This volume contains the Full Papers accepted for presentation at the V International Conference on Computational Methods for Coupled Problems in Science and Engineering, COUPLED PROBLEMS 2013, Santa Eulalia, Ibiza Island, Spain on June 17–19, 2013.

COUPLED PROBLEMS 2013 is the Fifth International Conference on this subject organized in the framework of Thematic Conferences of the European Community on Computational Methods in Applied Sciences (ECCOMAS) and is a Special Interest Conference of the International Association for Computational Mechanics (IACM).

The objective of COUPLED PROBLEMS 2013 is to become a forum for state of the art presentations and discussions of mathematical models, numerical methods and computational techniques for solving coupled problems of multidisciplinary character in science and engineering. The conference goal is to make a step forward in the formulation and solution of real life problems with a multidisciplinary nature and industrial interest, accounting for all the complex couplings involved in the physical description of the problem.

Previous editions of the conference were held on the Island of Santorini, Greece (2005); Ibiza, Spain (2007); Ischia, Italy (2009) and the Island of Kos, Greece (2011) with an increasing number of participants in each edition. COUPLED PROBLEMS 2013 has attracted over 450 participants, coming from all over the world. All together some 420 lectures will be presented, including 12 plenary lectures, which reflect the current state of the research and advances in engineering practice in the field. The CD Rom Proceedings contains contributions sent directly by the authors and the editors cannot accept responsibility for any inaccuracies and opinions expressed in the text.

The International Centre for Numerical Methods in Engineering (CIMNE) organizes this Conference jointly with the National Technical University of Athens (Greece) and the University of Padova (Italy). The organizers acknowledge the encouragement and support of ECCOMAS and IACM, under whose auspices this conference is held. The organizers would like to thank the authors for submitting their contributions and for their respect of the deadlines. Special thanks go to the colleagues who contributed to the organization of the 25 Invited Sessions.

In 2013, Eugenio Oñate turns 60 and for this reason we decided to dedicate, in his honor, the V COUPLED PROBLEMS 2013 Conference. We believe that Eugenio’s many contributions to the Numerical Methods in Engineering, in particular to the Coupled Problems, as well as his generous friendship with many people, are worth organizing this event for. We hope that you will enjoy the Conference, as well as the island of Ibiza and celebrating with us Eugenio’s 60th birthday.
ACKNOWLEDGEMENTS

The conference organizers acknowledge the support towards the organization of the COUPLED PROBLEMS 2013 Conference to the following organizations:

Center for Advanced Studies (CEAV), Ibiza, Spain

European Community on Computational Methods in Applied Sciences (ECCOMAS)

Ibiza Convention Bureau, Ibiza, Spain

International Association for Computational Mechanics (IACM)

International Center for Numerical Methods in Engineering (CIMNE), Spain

National Technical University of Athens (NTUA), Greece

Santa Eulalia City Hall, Ibiza, Spain

University of Padova, Italy

Universitat Politècnica de Catalunya (UPC), Spain

Plenary Speakers, European Commission Lecturer and Invited Session Organizers

We would also like to thank the Plenary Speakers and the Invited Session Organizers for their help in the setting up of a high standard Scientific Programme.

Plenary Speakers: Carlos Felippa, Thomas J. R. Hughes, Pierre Ladevèze, Wing Kam Liu, Rainald Löhner, Xavier Oliver, Eugenio Oñate, D.R.J. Owen, Tayfun Tezduyar, Peter Wriggers and Genki Yagawa.

European Commission Lecturer: Dietrich Knoerzer

SUMMARY

Preface ........................................................................................................................................... 7

Acknowledgements ......................................................................................................................... 9

Invited Sessions

Advances and Challenges in Computational Multiphysics and Multiscale .................................. 25
Organized by Carlos Felippa, Roger Ohayon and K.C. Park

Computational Models and Methods for Multiphysics Processes in Multiphase Porous Media ................................................................. 48
Organized by Lorenzo Sanavia and Claudio Tamagnini

Coupled Problems in Geotechnical Engineering ........................................................................ 80
Organized by Sebastià Olivella

Coupled Problems of the Dynamics of the Solid Systems with Friction ...................................... 92
Organized by Alexander V. Karapetyan and Alexey A. Kireenkov

Coupled Problems using Mesh-Reduction Methods .................................................................. 104
Organized by Alain Kassab, Eduardo Divo and Bozidar Sarler

Coupled Thermomechanical Modeling of Material Forming Processes .................................... 139
Organized by Carlos Agelet de Saracibar and Jean-Philippe Ponthot

Coupling Different Numerical Methods ..................................................................................... 193
Organized by Gernot Beer and Jerzy Rojek

Flows with Moving Boundaries and Interfaces ......................................................................... 217
Organized by Yuri Bazilevs, Kenji Takziawa and Tayfun Tezduyar

Fluid-Structure Interaction ......................................................................................................... 229
Organized by Tayfun Tezduyar, Kenji Takziawa and Yuri Bazilevs

Heterogeneous Materials & Computational Plasticity and Damage ........................................ 253
Organized by Francesco Marotti de Sciarra

Interaction with Time Resolved Turbulent Flow Fields ............................................................ 299
Organized by Dörte Carla Sternel

Inverse Problems in Mechanics .................................................................................................. 311
Organized by Assad Oberai

Mathematical Aspects of Interface Problems .......................................................................... 323
Organized by Daniele Boffi and Lucia Gastaldi

Modeling of Concrete at High Temperature .............................................................................. 335
Organized by Carmelo Majorana and Kaspar J. Willam

Multiscale-multiphysics Computational Techniques ............................................................. 355
Organized by Jacob Fish
Single and Multiobjective Optimization for Coupled Problems .................................................. 367
Organized by Massimo Guarnieri and Piergiorgio Alotto

Trends in the Theory and Numerics of Volume-Coupled Problems ........................................... 399
Organized by Bernd Markert and Stefan Turek

Unconventional Plasticity for Coupled Modelling ................................................................. 432
Organized by Koichi Hashiguchi and Valentina Salomoni

Contributed Sessions
Aeroelasticity ......................................................................................................................... 458
Coupled Problems in Bio-Medicine ..................................................................................... 479
Coupled Problems in Electro-Magnetics ............................................................................ 503
Coupled Problems in Fluid-Structure Interaction ............................................................ 550
Coupled Problems in Geomechanics .................................................................................. 744
Coupled Problems in Thermo-Mechanics .......................................................................... 828
Coupled Solution Strategies ............................................................................................... 895
High Performance Computing .......................................................................................... 966
Loose and Strong Coupling in FSI Problems ................................................................... 1000
Mathematical Formulations ............................................................................................... 1033
Multiphysics Problems .................................................................................................... 1062
Multiscale Problems ......................................................................................................... 1201
Numerical Methods in Coupled Problems Applied to Fluid Mechanics ....................... 1257
Numerical Methods in Coupled Problems, Finite Difference, Particle Methods, Meshless Methods, etc................................................................. 1302
Optimization Design ......................................................................................................... 1463
Thermo-Hydrodynamic Problems ..................................................................................... 1502
Authors Index ...................................................................................................................... 1515
CONTENTS

Invited Sessions

Advances and Challenges in Computational Multiphysics and Multiscale
Organized by Carlos Felippa, Roger Ohayon and K.C. Park

Keynote Lecturer: José Ángel González
Coupled analysis of nonlinear structural motion and fluid sloshing .................................................. 25
J.A. González* and K.C. Park

A coupled and multi-scale fluid-structure interaction and mass transfer model for biofilm growth simulations .......... 37
M. Coroneo*, L. Yoshihara and W.A. Wall

Computational Models and Methods for Multiphysics Processes in Multiphase Porous Media
Organized by Lorenzo Sanavia and Claudio Tamagnini

A model for the thermo-hydro-mechanical analysis of multiphase porous media in dynamics .................................. 48
M. Passarotto and L. Sanavia*

Anisotropic diffusion and propagation of sound waves in poroelastic media ...................................................... 58
B. Albers* and K. Wilmanski

Coupled thermo-hydro-mechanical modelling of crack development along fossil dinosaur’s footprints in soft cohesive sediments ........................................................................................................... 67
L. Nguyen-Tuan*, H. Viefhaus, M. Datcheva and T. Schanz

Coupled Problems in Geotechnical Engineering
Organized by Sebastià Olivella

Keynote Lecturer: Uwe-Jens Görke
On the necessity and a generalized conceptual model for the consideration of large strains in rock mechanics ........ 80
U-J. Görke*, T. Nagel and O. Kolditz

Coupled Problems of the Dynamics of the Solid Systems with Friction
Organized by Alexander V. Karapetyan and Alexey A. Kireenkov

A comparative study of computer models for friction and their influence on dynamics of the heavy rigid body on a horizontal surface ............................................................................................................. 92
I.I. Kosenko*

Coupled Problems using Mesh-Reduction Methods
Organized by Alain Kassab, Eduardo Divo and Bozidar Sarler

A high-resolution scalable meshless method for coupled steady poroelastic analysis, based on collocation with radial basis functions ........................................................................................................ 104
H. Power, D. Stevens* and A. Cliffe

Local radial basis function collocation method for solving thermo-mechanics of hot shape rolling of steel ............ 116
U. Hanoglu and B. Šarler*

Meshless 2D direct numerical simulation and heat transfer in a backward-facing step with heat conduction in the step .......... 127
A.G. Vidal*, A.J. Kassab and E. Divo

Coupled Thermomechanical Modeling of Material Forming Processes
Organized by Carlos Agelet de Saracibar and Jean-Philippe Ponthot

Keynote Lecturer: Stefanie Reese
Advanced material and structural behavior in innovative forming processes – Application to process-integrated ring rolling .... 139
R. Kebriaei and S. Reese*
A new frame for constitutive modeling of visco elastic polymers; accounting thermal effects and strain induced crystallization

N. Billon* and J. Bikard

Comparison of a fluid and a solid approach for the numerical simulation of friction stir welding with a non-cylindrical pin

P. Bussetta*, N. Dialami, R. Boman, M. Chiumenti, C. Agelet de Saracibar, M. Cervera and J-P. Ponthot

Numerical investigation of continuous fiber glass drawing

Q. Chouffart*, P. Simon and V.E. Terrapon

Tooling materials and solutions for the thixoforming steels

A. Rassili* and J. Lecomte-Beckers

**Coupling Different Numerical Methods**

*Organized by Gernot Beer and Jerzy Rojek*

Coupling level set methods with boundary integral methods for free surface potential flows

M. Garzon*, L.J. Gray and J.A. Sethian

Parallel application on high performance computing platforms of 3D BEM/FEM based coupling model for dynamic analysis of SSI problems

M.C. Genes*

**Flows with Moving Boundaries and Interfaces**

*Organized by Yuri Bazilevs, Kenji Takizawa and Tayfun Tezduyar*

Applications of an incompressible fluid-rigid body interaction on progressive moving-grid finite-volume method

S. Asao* and K. Matsuno

**Fluid-Structure Interaction**

*Organized by Tayfun Tezduyar, Kenji Takizawa and Yuri Bazilevs*

**Keynote Lecturer: Yuri Bazilevs**

Fluid-structure interaction simulations with novel structural mechanics discretizations

Y. Bazilevs*, A. Korobenko, X. Deng, J. Tippmann and M-C. Hsu

A parallel, black-box coupling algorithm for fluid-structure interaction

B. Uekermann*, H-J. Bungartz, B. Gatzhammer and M. Mehl

**Heterogeneous Materials & Computational Plasticity and Damage**

*Organized by Francesco Marotti de Sciarra*

**Keynote Lecturer: Francesco Marotti de Sciarra**

A constitutive model of coupled thermoelasticity with plasticity

R. Barretta, F. Marotti de Sciarra* and M. Salerno

Considerations on the plastic buckling of bi-layered cylindrical shells

F. Guarracino* and A.C. Walker

Non-proportionality of strain and coupling effects on dislocation distribution and ductility in ferritic steel polycrystals

T.O. Erinosho*, A.C.F. Cocks and F.P.E. Dunne

The geometric paradigm in computational elasto-plasticity

G. Romano, R. Barretta* and F. Marotti de Sciarra

**Interaction with Time Resolved Turbulent Flow Fields**

*Organized by Dörte Carla Sternel*

On the performance of detached-eddy simulation in fluid-structure interaction problems

S. Türk*, T. Reimann, D.C. Sternel and M. Schäfer
Inverse Problems in Mechanics
Organized by Assad Oberai

Shape and stress sensing of multilayered composite and sandwich structures using an inverse Finite Element Method ........ 311
P. Cerracchio*, M. Gherlone, M. Di Sciuva and A. Tessler

Mathematical Aspects of Interface Problems
Organized by Daniele Boffi and Lucia Gastaldi

Mass preserving, distributed lagrangian multiplier approach to immersed boundary method ........................................... 323
D. Boffi, N. Cavallini*, F. Gardini and L. Gastaldi

Modeling of Concrete at High Temperature
Organized by Carmelo Majorana and Kaspar J. Willam

Damage evaluation of RC members under high temperature by using RBSM-TRUSS Network Model ......................... 335
T. Unno*, H. Nakamura, Y. Yamamoto, M. Kunieda and N. Ueda

Moisture transport in concrete during fire: an NMR study .............................................................................................. 345
L. Pel*, G.H.A. van der Heijden and S.J. Jaspers

Multiscale-Multiphysics Computational Techniques
Organized by Jacob Fish

A multi-scale refined zigzag theory for multilayered composite and sandwich plates with improved transverse shear stresses ................................................................................................................................. 355
L. Iurlaro*, M. Gherlone, M. Di Sciuva and A. Tessler

Single and Multiobjective Optimization for Coupled Problems
Organized by Massimo Guarnieri and Piergiorgio Alotto

Keynote Lecturer: Eric J. Inclan
Effective modifications to differential evolution optimization algorithm ................................................................. 367
E.J. Inclan* and G.S. Dulikravich

Keynote Lecturer: Raffaelle Martone
A coupled thermo-electrical problem for the optimization of radio frequency ablation treatments ......................... 379
A. Formisano, F. Ferraioli and R. Martone*

Intelligent computing in optimization of coupled problems ........................................................................................... 387
T. Burczynski* and A. Dlugosz

Trends in the Theory and Numerics of Volume-Coupled Problems
Organized by Bernd Markert and Stefan Turek

Coupling generalised-α methods: Analysis, adaptivity, and numerics ................................................................. 399
J. Rang*

Fast solvers and efficient numerical CFD techniques for dynamic porous media problems ................................. 411
S. Turek, A. Obaid* and B. Markert

Numerical simulation of inductive heating processes .................................................................................................. 423
T. Gleim, D. Kuhl and B. Schröder*

Unconventional Plasticity for Coupled Modelling
Organized by Koichi Hashiguchi and Valentina Salomoni

Keynote Lecturer: David R. Hayhurst
Numerical analysis of the coupling between mechanical strain and thermal conductivity of ceramic matrix composites .......................................................... 432
D.R. Hayhurst*
### Keynote Lecturer: Vassilis Panoskaltsis
Decompositions of the stress and the rate of deformation tensors for materials undergoing phase transformations

V. Panoskaltsis*

### Contributed Sessions

#### Aeroelasticity

Determination of 3D wind induced vibration of cables for cable-stayed bridges

A. Hernández* and J.G. Valdés

Mesh deformation with exact surface reconstruction using a reduced radial basis function approach

D.S.C. Kowollik*, M.C. Haupt and P. Horst

#### Coupled Problems in Bio-Medicine

A finite element analysis of the performance of stents for angioplasty using the hydroforming process

R. de Araujo, S.A.G. Oliveira, T.A. Guimarães and F.R.F. Neto*

Femoral vectoring for hip dysplasia in neonates


#### Coupled Problems in Electro-Magnetics

**Keynote Lecturer: Javier Bonet**

A new variational framework for large strain hyperelastic piezoelectric materials

J. Bonet*, R. Ortigosa and A.J. Gil

Mathematical structure of the transport equations for coupled 2D-3D electron gasses in a MOSFET

V.D. Camiola and V. Romano*

The coupled problem among displacement, thermo and electromagnetic fields and applications in electronic equipments

B.Y. Duan* and P.Y. Lian

Thermoelectric simulation of electric machines with permanent magnets

A. Bermúdez, D. Gómez, F. Pena, M. Piñeiro and P. Salgado

#### Coupled Problems in Fluid-Structure Interaction

**Keynote Lecturer: Antonio J. Gil**

An Enhanced Immersed Structural Potential Method (ISPM) for the simulation of fluid-structure interaction problems

A.J. Gil*, A. Arranz Carreno, J. Bonet and C. Wood

**Keynote Lecturer: Paulo Lyra**

Minimization of the phenomenon of vibrations induced by vortices using surrogate based optimization

U.P. Farias Filho, A.R.E. Antunes, S.M.A. Bastos and P.M.R. Lyra*

3d conservative coupling method between a compressible fluid flow and a deformable structure

M.A. Puscas*, V. Daru, A. Ern, C. Mariotti, L. Monasse and C. Tenaud

A comparative study of immersed-boundary interpolation methods for a flow around a stationary cylinder at low Reynolds number

S.H. Madani*, J. Wissink and H. Bahai

A coupled mode – hpFEM for hydroelastic analysis of shear-deformable floating bodies of general thickness in variable bathymetry

K.A. Belibassakis, G.A. Athanassoulis, T.K. Papathanasiou* and S. Filopoulos

A numerical approach for static and dynamic analysis of deformable journal bearings

D. Benasciutti, M.Gh. Munteanu* and F. Flumian
Added resistance of ships in quartering seas .......................................................... 621
S. Liu and A.D. Papanikolaou*

An immersed boundary level-set based approach for fluid-shell interaction with impact .......................................................... 631
R.A.K. Sanches*, R. Carrazedo and H.B. Coda

Effects of slip condition on flow near the surface of hydro-gel ........................................... 643
M. Yokoyama*, Y. Kubota, K. Kikuchi, O. Mochizuki and G. Yagawa

FSI of high performance high-lift devices with circulation control via conditioned Coandă-Jets ........................................ 651
K. Sommerwerk*, M.C. Haupt and P. Horst

Mathematical modeling of hydrodynamic processes in geothermal plant ........................................ 661
M. Bogdevičius, J. Janutėnienė, S. Razmas and M. Drakšas

Modal characteristics of a flexible tube in turbulent axial flow:
A numerical approach and validation with experimental data .................................................. 668
J. De Ridder*, J. Degroote, K. Van Tichelen, P. Schuurmans and J. Vierendeels

Multidisciplinary and multiphysics computation of supersonic flow, via hybrid solutions
for compressible navier-stokes layer ........................................................................ 677
A. Nastase*

Numerical simulation of deformable cell in microchannels .................................................. 685
F. Nason*, L. Zhu, G. Dubini and L. Brandt

Numerical simulation of explosive fracturing with smoothed particle hydrodynamics ........................................ 696
Z. Guo, Y. Shen* and M. Liu

Numerical simulation of landmine explosions: comparison between different modelling approaches ........................................ 708
C. Fichera, L. Peroni* and M. Scapin

Prediction of sound fields in acoustic cavities coupled to absorptive structures due to vibrating surfaces ........................................ 720
M. Buchschmid*, G. Müller and M. Pospiech

Seizmic analysis of the Lesce dam including water-dam-soil dynamic interaction ........................................ 732
J. Radnić, A. Harapin and M. Sunara*

**Coupled Problems in Geomechanics**

**Keynote Lecturer: Scott Sloan**
Adaptive finite element limit analysis incorporating steady state pore pressures ........................................ 744
S.W. Sloan*, A.V. Lyamin and K. Krabbenhoft

**Keynote Lecturer: Umberto Perego**
A PFEM approach to the simulation of landslide generated water-waves ........................................ 756
M. Cremonesi, C. di Prisco and U. Perego*

Advances on mesh generation in numerical simulation of fractured network system ........................................ 763
S. Sun, J. Sui and M. Yuan*

Finite element modeling of thermo-hydro-mechanically (THM) coupled problems
in frozen ground engineering: State-of-the-Art ................................................................. 773
Y.W. Bekele*

Influence of cracks on the soil-atmosphere interaction: numerical coupled model of
thermo - atmosphere - porous media ........................................................................ 786
D. Tabbal*, I. Shahroul, A. Al-Qadad, F. Hage Chehade and M. Sadek

Influence of fluid-mechanical coupling in gas generation in undersaturated petroleum reservoirs ........................................ 794
Lattice Boltzmann based multicomponent reactive transport model coupled with geochemical solver for pore scale simulations


On numerical modeling of couple heat, air and moisture transfer through multilayered walls

M. Maliki*, N. Laredj, H. Naji, K. Bendani, H. Missoum and K. Menad

### Coupled Problems in Thermo-Mechanics

Coupled electro-elasticity at finite strains

C. Sansour, S. Skatulla and M. Hijiaj

Parametric identification of mathematical models of coupled conductive-radiative heat transfer

A.V. Nenarokomov*, O.M. Alifanov and D.M. Titov

Simulation of the thermal profile of a mushy metallic sample during tensile tests

F. Palama Bongo*, J. Favergeon, M. Rachik and D. Marceau

The DRBEM solution of the generalized magneto-thermo-viscoelastic problems in 3D anisotropic functionally graded solids

M.A. Fahmy*

Thermochemical heat storage for sustainable power generation – model development based on the theory of porous media and numerical implementation


Thermo-chemical modeling of fiber-polymer composites in fire for fluid/structure interaction

V. Biasi*, G. Leplat and F. Feyel

### Coupled Solution Strategies

Implementation of a Weak Coupling Approach between a CFD and an FE Software for Fires in Compartment

N. Tondini* and J-M. Franssen

Non-symmetrical BE-FE partitioned formulations for acoustic fluid-structure interaction problems

L. Rodríguez-Templeque*, J.A. González, A. Cerrato and R. Abascal

Numerical coupling procedure in steady conjugate heat transfer problems

M-P. Errera* and S. Chemin

Research on electromechanical field coupling model and solution strategy of large reflector antennas

M. Wang* and W. Wang

Rosenbrock time integration for unsteady flow simulations

D.S. Blom*, H. Bijl, P. Birken, A. Meister and A.H. van Zuijlen

Simultaneous flow of water and air across the land surface during runoff


Towards interaction of elastic structures with turbulence induced acoustics at low Mach numbers

F. Flitz* and M. Schäfer

### High Performance Computing

**Keynote Lecturer: Norberto Nigro**

Recent advances in the Particle Finite Element Method. Towards more complex fluid flow applications


Novel kinetic consistent 3D MHD algorithm for high performance parallel computing systems

B. Chetverushkin*, N. D’Ascenzo and V. Saveliev

On the difficulties of real-time co-simulation

G. Stettinger*, M. Benedikt, N. Thek and J. Zehetner
Loose and Strong Coupling in FSI Problems

Comparison between two different decompositions for the solution of fluid-structure interaction problems...........1000
J. Degroote*, J. Bols and L. Taelman

Macro-Step-Size selection and monitoring of the coupling error for weak-coupled subsystems in the
frequency-domain.................................................................1009
M. Benedikt, D. Watzenig, J. Zehetner* and A. Hofer

NEPCE - A nearly energy-preserving coupling element for weak-coupled problems and co-simulations...........1021
M. Benedikt*, D. Watzenig, J. Zehetner and A. Hofer

Mathematical Formulations

2D Velocity-vorticity viscous incompressible flows .......................................................................................1033
R. Téllez, H. Acevedo and A. Nicolás*

A functional for the momentum equations of incompressible viscous flow ...................................................1045
S. Li*, N. Li and W. Duan

Mathematical and numerical approach for a crashworthy problem ............................................................1051
S. Boria* and S. Pettinari

Multiphysics Problems

A coupled problem of heat and mass transfer applied to porous textile media surrounding the human foot........1062
J. Durany*, L. Poceiro and F. Varas

A coupled thermo-flexible multi-body approach for virtual rig testing of brake discs ....................................1074
N. Strömberg*

A large deformation and thermomechanically coupled interface approach .....................................................1082
M. Kaliske*, R. Fleischhauer, R. Behnke and I. Zreid

A Multiscale method to analyze the deterioration due to Alkali Silica Reaction considering the effects of
temperature and relative humidity .................................................................1094
T. Wu*, I. Temizer and P. Wriggers

Analytical solutions of nonstationary adiabatic processes for compression (extension) of viscoplastic
spherical and cylindrical shells, spherical and cylindrical layers from viscous liquid ....................................1101
A. Kiselev*

Coupled quantum-classical transport in Silicon nanowires ............................................................................1109
O. Muscato* and V. Di Stefano

Coupled problems in analysis of quantum dots with multiband models .........................................................1120
S. Prabhakar and R. Melnik*

Coupled swelling and large strain model for hydrogels: Application to the nucleus pulposus of the intervertebral
disc ................................................................................................1130
V. Brulliard*, S. Le Corre and E. Verron

Fundamental solutions for a coupled formulation of porous biphasic media with compressible
solid and fluid phases ..........................................................................................................................1142
R. Serpieri*, F. Travascio and S. Asfour

High performance conjugate heat transfer with the openPALM coupler ........................................................1154
F. Duchaine*, S. Jauré, D. Poitou, E. Quémerais, G. Staffelbach, T. Morel and L. Gicquell

Influence of the ionic strength on the deposit phenomenon and transport dynamic of microparticles
through saturated porous medium...........................................................................................................1168
Z. Mesticou*, M. Kacem and Ph. Dubujet
New coupled thermoelectric link finite element for FGM materials .......................................................... 1179
J. Paulech*, J. Murín, V. Kutíš and J. Hrabovský

Numerical simulation for evaporation of chemical agent droplet ............................................................. 1191
H. Hayakawa, M. Yamamoto, M. Kawai and K. Sakamoto

**Multiscale Problems**

A multiscale damage model for composite materials using a FFT-based method ........................................ 1201
J. Spahn*, H. Andrä, M. Kabel and R. Müller

Coupled thermoelastic simulation of nanovoid cavitation by dislocation emission at finite temperature .... 1213
M. Ponga, M. Ortiz and M.P. Ariza*

Multiscale analysis of piezoelectric material by using EBSD-measured realistic model ............................ 1226
H. Kuramae*, H. Sakamoto and Y. Uetsuji

Rule-based expert system application to optimizing of multiscale model of hot forging and heat treatment of Ti-6Al-4V ................................................................. 1237

Spatial and temporal instability of slightly curved particle-laden shallow mixing layers ............................ 1249
I. Eglite and A.A. Kolyshkin*

**Numerical Methods in Coupled Problems Applied to Fluid Mechanics**

Consolidation problem solution with a coupled hydro-mechanical formulation considering fluid compressibility ................................................................. 1257
S.R.C. Brant*, M.P. Cordão Neto and I.F. Gomes

Numerical Solution of P-SV waves with free surface boundary conditions by network method ............. 1269
J.L. Morales, I. Alhama and M. Alcaraz*

Schwarz Alternating domain decomposition approach for the solution of mixed heat convection flow problems based on the Method of Approximate Particular Solutions ........................................ 1280
C.A. Bustamante*, H. Power and W.F. Florez

Virtual reality 3D simulation of fluid-particle interaction ........................................................................... 1292
M. Chekired and V. Roubtsova*

**Numerical Methods in Coupled Problems Finite Difference, Particle Methods, Meshless Methods, etc.**

A LS-DYNA/FLUKA coupling for the numerical simulation of high energy particle beam interaction with matter ................................................................................................. 1302
M. Scapin*, L. Peroni, V. Boccone and F. Cerutti

A new functional for improving cell area distribution .................................................................................. 1310
P. Barrera, F.J. Domínguez-Mota and G. González*

An iterative coupling approach to the simulation of elastic wave .............................................................. 1322
J. António*, A. Tadeu and P. Ferreira

Automatic differentiation based formulation of coupled problems .......................................................... 1331
B. Hudobivnik* and J. Korelc

Computational optimisation strategies of tailored fiber placement in polymer matrix composites based on local shear stress minimisation ................................................................. 1343
T. Rettenwander*, M. Fischlschweiger and G. Steinbichler

Development of a maximum entropy approach for the thermomechanical modelling of the rotary friction welding process ...................................................................................... 1354
M. Foca*, G. Racineux and L. Stainier
Evaluation of automotive weatherstrip by coupled analysis of fluid-structure-noise interaction .............................................................. 1365
H.Y. Kim*, H.J. Kim and T.H. Kim

Low intrusive coupling of implicit and explicit integration schemes for structural dynamics: application to low energy impacts on composite structures ................................................................. 1373
T. Chantrait*, J. Rannou and A. Gravouil

Modal analysis of the FGM beam-like structures with effect of the thermal axial force ................................................................. 1383
J. Murín, M. Aminbaghai, J. Hrabovský, V. Kutiš* and J. Paulech

Model identification and on-line optimal control of food processes ................................................................................................. 1395

Network models for the numerical solution of coupled ordinary non-lineal differential equations .............................................. 1407
I. Alhama, J.L. Morales* and F. Alhama

Science & engineering software migration: Moving from desktop to mobile applications ............................................................ 1418
F. Améndola and L. Favre*

Semi-analytical approach for ratcheting strain and wear prediction in case of line rolling contact .................................................. 1430
M. Sofer* and R. Halama

Two dimensional solution of the advection-diffusion equation using two collocation methods with local upwinding RBF ............................................................. 1439
J.M. Granados*, C.A. Bustamante, H. Power and W.F. Florez

Use of a BEM model to simulate unsteady coupled heat and moisture flow through porous solid ................................................. 1451
L. Škerget* and A. Tadeu

Optimization Design

Optimal actuator and sensor placement with regard to coupled electro-mechanical behaviour of smart structures ................................................................. 1463
T. Nestorović* and M. Trajkov

Probability and variance-based stochastic design optimization of a radial compressor concerning fluid-structure interaction ........................................................................... 1475
D. Roos*, K. Cremanns and T. Jasper

Thermo-Hydrodynamic Problems

A fully coupled thermo-hydro-mechanical model for the analysis of the lining behavior of underground caverns in an AA-CAES system ........................................................................... 1502
A. El Murr*, F. Hadj-Hassen, A. Rouabhi and M. Tijani

Authors Index ........................................................................................................................................................................................................... 1515
INVITED SESSIONS
COUPLED ANALYSIS OF NONLINEAR STRUCTURAL MOTION AND FLUID SLOSHING

JOSÉ A. GONZÁLEZ* AND K. C. PARK†

*Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Camino de los Descubrimientos s/n, 41092 Sevilla, Spain
e-mail: japerez@us.es

†Division of Ocean Systems Engineering
Korea Advanced Institute of Science and Technology (KAIST)
335 Science Rd, Yuseong-gu, Daejeon 305-70, Republic of Korea
e-mail: kcpkaist@kaist.edu

Key words: Fluid-structure interaction, Partitioned formulations, Localized Lagrange multipliers

Abstract. Fluid sloshing in containers is modeled using a finite element formulation previously proposed by the authors for problems with moderate motions [1], extending in this work its application to arbitrarily large rotations and small deformations relative to a floating frame of reference moving with the fluid. This novel approach is used to investigate the coupling effects originated by the incidence of environmental sea waves on rigid floating vessels with internal flexible structural parts and fluids oscillating inside rigid or flexible tanks.

1 INTRODUCTION

This work presents a partitioned finite element formulation for the solution of structure-structure and fluid-structure coupled problems, under the assumption of relatively small deformations but arbitrarily large rotations, based on the floating frame of reference approach. The main difference with the classical approach is in the algebraic separation of pure-deformational from pure-rigid modes, defining the position of the body-frame at any point of the undeformed state of the body. Pure-deformational modes are then measured with respect to this corotated configuration. The same concept was introduced by Fraeijs de Veubeke in [1], treating the case of a complete structure presenting coupled rigid-body and deformational motions. An important consequence of this definition of the reference frame is the uncoupling of rigid-body from deformational motions in the inertia mass matrix.
The mechanical response of the multibody system can also be coupled with a fluid domain. Fluid and structure systems are treated separately and connected using localized Lagrange multipliers to a common frame tracking the interface motion. This coupling strategy permits easier parallelization and facilitates the enforcement of slip condition between the fluid and the structure walls. The fluid is modeled using a finite element formulation previously developed by the authors for sloshing problems with moderate motions, extending its application here to arbitrarily large rotations and small deformations relative to a floating frame of reference moving with the fluid. This novel approach is used to investigate the coupling effects of a rigid floating structure with flexible structural parts and a fluid oscillating inside rigid or flexible tanks due to global motion.

2 KINEMATICS

Two reference systems are introduced. The first one is a fixed inertial-frame and vectors expressed in this system are represented using capital letters. The second system is a floating frame of reference that is fixed to the body and moves with it, vectors defined in this system are written using lowercase. The position of an arbitrary point of the body is expressed in the inertial frame as:

\[ \mathbf{X} = \mathbf{X}_0 + \mathbf{A}(\mathbf{r} + \mathbf{d}) \]  

where \( \mathbf{X}_0 \) is the vector defining the position of a fixed point 0 of the body in the inertial frame, \( \mathbf{A} \) is the rotation matrix of the body-frame, \( \mathbf{r} \) is the vector defining the position of the point in its undeformed position expressed in the body-fixed frame and \( \mathbf{d} \) is the deformational displacement vector of the same point expressed in the same frame.

Using the small deformational-displacements approximation, the position, velocity and acceleration of a particle expressed in the inertial frame can be simplified:

\[ \dot{\mathbf{X}} = \dot{\mathbf{X}}_0 + \dot{\mathbf{A}}\dot{\mathbf{d}} - \dot{\mathbf{A}}\tilde{\mathbf{r}}\omega \]  
\[ \ddot{\mathbf{X}} = \ddot{\mathbf{X}}_0 + \ddot{\mathbf{A}}\ddot{\mathbf{d}} - \ddot{\mathbf{A}}\tilde{\mathbf{r}}\alpha - \dddot{\mathbf{A}}\tilde{\mathbf{r}} + 2\dddot{\mathbf{A}}\tilde{\mathbf{r}}\]  

We can see in equation (4) that acceleration of a particle is obtained by composition of linear, deformational, angular, centrifugal and Coriolis acceleration terms.

Finally, given a set of virtual displacements \( (\delta\mathbf{d}, \delta\mathbf{X}_0, \delta\theta) \), the virtual displacement \( \delta\mathbf{X} \), expressed in the body-frame, is approximated:

\[ \delta\mathbf{X} = \mathbf{A}\phi \begin{bmatrix} \mathbf{d} \\ \delta\mathbf{X}_0 \\ \delta\theta \end{bmatrix} \]  

where \( \phi \) represents the displacement interpolation matrix.
2.1 Description of the deformation

As explained earlier, the deformation of each floating substructure is described using a local frame of reference that translates and rotates following the undeformed picture of the solid. Merely condition to define this system is that the substructure displacement field does not present rigid body components or free modes.

Total displacements of a discretized free floating substructure can be separated into a pure deformational component plus a rigid-body part, as illustrated in Figure 1:

\[ \mathbf{u} = \mathbf{d} + \mathbf{R}\alpha \]  

(6)

where \( \mathbf{u} \) are the nodal displacements, \( \mathbf{d} \) represents the vector of pure-deformational displacements, \( \mathbf{R} \) is a basis of the rigid-body modes and vector \( \alpha \) collects the amplitudes of these rigid-body motions. Rigid-body matrix \( \mathbf{R} \) is then a block-matrix composed of nodal contributions:

\[ \mathbf{R}^T = \begin{bmatrix} \mathbf{R}_1^T & \ldots & \mathbf{R}_n^T \end{bmatrix} \]  

(7)

with sub-blocks that can be formed directly for each node of the mesh as:

\[ \mathbf{R}_i = \begin{bmatrix} \mathbf{I} & -\mathbf{r}_i \end{bmatrix} \quad (i = 1 \ldots n) \]  

(8)

where \( \mathbf{r}_i \) is the position vector of node \( i \) relative to a rotation center 0 and \( n \) is the total number of nodes of the free floating subdomain.

Once the rigid-body modes are obtained, separation of the total displacements into deformational and a rigid-body contributions can be done as described by Felippa and Park [2]. This is accomplished by using the projector:

\[ \mathbf{P} = \mathbf{I} - \mathbf{MRM}^{-1}_\alpha \mathbf{R}^T \]  

(9)

where \( \mathbf{M} \) is a symmetric definite positive mass matrix and \( \mathbf{M}_\alpha = \mathbf{R}^T\mathbf{MR} \) is the principal mass matrix introduced by Park et al. in [3] a \((6 \times 6)\) matrix for a three-dimensional floating substructure. This operator presents the filtering properties \( \mathbf{P}^T\mathbf{R} = \mathbf{0} \) and \( \mathbf{P}\mathbf{MR} = \mathbf{0} \),
allowing to separate pure deformational modes from rigid-body motions using the following expressions:

\[ d = \mathcal{P}^T \mathbf{u}, \quad \mathbf{R} \mathbf{\alpha} = (\mathbf{I} - \mathcal{P}^T) \mathbf{u} \] (10)

where, by definition, the projector \( \mathcal{P}^T \) performs an orthogonal projection in the subspace defined by the rigid-body modes and therefore is acting as a filter for the deformational component of displacements.

For our derivation of the equations of motion, it will be useful to separate the rigid-body modes \( \mathbf{R} \) into its translational and rotational components:

\[ \mathbf{R} = \begin{bmatrix} \mathbf{R}_t & \mathbf{R}_r \end{bmatrix} \] (11)

using subscripts \((t)\) and \((r)\) to refer to the translational or the rotational component.

### 3 Variational formulation

The total virtual work for a group of deformable substructures undergoing arbitrary large-rotations is composed by the following terms:

\[ \delta W_T = \delta W_i + \delta W_d + \delta W_f + \delta W_c \] (12)

corresponding to the virtual work of inertia forces, internal forces, external forces and constraints.

Virtual work of inertia forces \( \delta W_i \) for a free-floating substructure is obtained integrating in the volume \( V \) the product of particle acceleration (4) times virtual displacement (5):

\[ \delta W_i = \int_V \rho \dddot{\mathbf{X}} \cdot \delta \mathbf{X} \, dV \] (13)

Introducing a FEM discretization, and considering equations (4) and (5), evaluation of the virtual work of inertia forces yields:

\[ \delta W_i = \begin{bmatrix} \frac{\delta \mathbf{d}}{\delta \mathbf{\theta}} \end{bmatrix}^T \begin{bmatrix} \mathbf{M} & \mathbf{S}_t & \mathbf{S}_r \\ \mathbf{S}_t^T & \mathbf{M}_t & -m \mathbf{\tilde{r}}_G \\ \mathbf{S}_r^T & -m \mathbf{\tilde{r}}_G^T & \mathbf{M}_r \end{bmatrix} \begin{bmatrix} \mathbf{A}^T \dddot{\mathbf{X}}_0 \\ \mathbf{A}^T \dddot{\mathbf{X}}_0 \\ \mathbf{A}^T \dddot{\mathbf{X}}_0 \end{bmatrix} + \begin{bmatrix} \mathbf{g}_{\text{cor}}^{\text{d}}(\mathbf{\omega}, \dot{\mathbf{d}}) \\ m \dot{\mathbf{\omega}}^2 \mathbf{r}_G \\ \mathbf{\tilde{\omega}}^T \mathbf{M}_r \mathbf{\omega} \end{bmatrix} \] (14)

where \( m \) is the total mass of the body, \( \mathbf{r}_G \) is the position vector of the body COG in local coordinates, \( \mathbf{M}_t = m \mathbf{I}_3 \) is the (3x3) translational mass matrix, \( \mathbf{M}_r \) is the (3x3) inertia tensor and \( \mathbf{M} = \int_V \rho \mathbf{\phi}^T \mathbf{\phi} \, dV \) is the finite-element mass matrix.
Coupling inertia terms in (14) take the form:

\[ S_t = \int_V \rho \phi^T dV, \quad S_r = \int_V \rho \phi^T \tilde{r}^T dV \]  \hspace{1cm} (15)

and the quadratic velocity terms due to centrifugal and Coriolis accelerations:

\[ g_{\text{cen}}(\omega) = \int_V \rho \phi^T \omega^2 r dV, \quad g_{\text{cor}}(\omega, \dot{d}) = 2 \int_V \rho \phi^T \tilde{\omega} \phi \dot{d} dV \]  \hspace{1cm} (16)

with contributions of the Coriolis acceleration to the translational and rotational rigid-body equations given by:

\[ g_{\text{cor}}(\omega, \dot{d}) = 2 \tilde{\omega} S_t \dot{d}, \quad g_{\text{cor}}(\omega, \dot{d}) = 2 \int_V \rho \phi^T \tilde{r} \phi \dot{d} dV \]  \hspace{1cm} (17)

It is observed that the rigid-deformational inertia coupling terms can be expressed as the product of the deformational mass matrix and the rigid-body modes as \( S_t = MR_t \) and \( S_r = MR_r \). Similarly, for the rigid-body mass matrix terms of (14), we have:

\[ M_t = R_t^T MR_t, \quad M_r = R_r^T MR_r, \quad -m \tilde{r} G = R_t^T MR_r \]  \hspace{1cm} (18)

relations previously derived by Park et al. in [3].

Assuming that the reference point of the body is located at the COG, i.e. \( r_G = 0 \), the discrete approximation of the variational is expressed:

\[ \delta W_i = \delta X_0^T \{ M_t \ddot{X}_0 \} + \delta q^T \{ M_G \ddot{q} - 2M_G \dot{q} \} + \delta d^T \{ M_p \ddot{d} + g_d \} \]  \hspace{1cm} (20)

where deformational and rigid-body inertia terms are uncoupled.

Assuming that the reference point of the body is located at the COG, i.e. \( r_G = 0 \), the discrete approximation of the variational is expressed:

\[ \delta W_i = \delta X_0^T \{ M_t \ddot{X}_0 \} + \delta q^T \{ M_G \ddot{q} - 2M_G \dot{q} \} + \delta d^T \{ M_p \ddot{d} + g_d \} \]  \hspace{1cm} (20)

where deformational and rigid-body inertia terms are uncoupled.

Assuming that the reference point of the body is located at the COG, i.e. \( r_G = 0 \), the discrete approximation of the variational is expressed:

\[ \delta W_i = \delta X_0^T \{ M_t \ddot{X}_0 \} + \delta q^T \{ M_G \ddot{q} - 2M_G \dot{q} \} + \delta d^T \{ M_p \ddot{d} + g_d \} \]  \hspace{1cm} (20)

where deformational and rigid-body inertia terms are uncoupled.
with rotations parametrized using *Euler-parameters* denoted by the quaternion \( q \).

To compute the virtual work due to deformations, the substructure is discretized using the classical linear FEM approximation, where the assembly of element contributions by the direct stiffness method leads to the semi-discrete equations of motion:

\[
\delta W_d = \delta d^T \{ K_P d \} \tag{21}
\]

\[
K_P = \mathcal{P} K \mathcal{P}^T
\]

where \( K \) is the small-displacements finite element stiffness matrix, presenting a null-space given by the rigid-body modes, that is, \( K_R = 0 \) and where the projector operator has been introduced to eliminate the rigid-body component.

The virtual work produced by body loads and boundary tractions acting on the body is approximated by the discrete equation:

\[
\delta W_f = -\delta d^T \{ Pf \} - \delta X_b^T \{ AR_1^T f \} - \delta q^T \{ G^T R f \} \tag{22}
\]

provided that the total external force vector is expressed as \( f \).

Flexible substructures are then connected using classical Lagrange multipliers, see Figure 2, that appear in the system as internal forces to satisfy compatibility conditions expressed in the body frame of reference.

The virtual work done by the constraints \( \delta W_c \) can be expressed:

\[
\delta W_c = \int_{\Gamma_c} \delta \{(A \lambda) \cdot (X - X_f)\} \, dV \tag{23}
\]
where $\lambda$ represents the localized Lagrange multipliers attached to each substructure and expressed in the body frame of reference, used to enforce the kinematic compatibility condition between the body and a frame with position $X_f$.

After discretization, equation (23) transforms into:

$$
\delta W_c = \delta \lambda^T \{ B^T (R_t A^T X_0 + r + P^T d) - L_f A^T X_f \} + \delta X_0^T \{ A R_f^T B \lambda \} + \delta q^T \{ G^T R_f^T B \lambda \} + \delta d^T \{ B P \lambda \} - \delta X_f^T \{ A L_f^T \lambda \}
$$

with $B_P = PB$, providing independent Lagrange multipliers for each substructure. This approach allows to express substructural constraint equations in the particular local system of each body involving only the unknowns of one body and its frame.

Finally, an additional constraint enforcing the unity norm of the quaternion should be added. This is done introducing a new Lagrangian multiplier $\mu$ to enforce this condition and adding to (24) a new term:

$$
\delta W_q = \delta \{ \mu (q^T q - 1) \}
$$

4 PARTITIONED EQUATIONS OF MOTION

The total virtual-work of a FEM partitioned system undergoing arbitrarily large rotations with small deformations is derived from (12), (20), (21), (22), (24) and (25):

$$
\delta W_T = \delta d^T \{ M_P \ddot{d} + K_P d + B_P \lambda + P M \Omega R_r G \dot{q} + 2 P \Omega M \dot{d} - P f \} + \delta \lambda^T \{ B^T (R_t A^T X_0 + r + P d) - L_f A^T X_f \} + \delta X_0^T \{ M \dddot{X}_0 + A R_f^T B \lambda - A R_f^T f \} + \delta \mu \{ 2 q^T \dot{q} \} - \delta X_f^T \{ A L_f^T \lambda \}
$$

and from the stationary-point condition of total virtual-work, equations of motion are obtained:

$$
\begin{bmatrix}
(M_P \frac{d^2}{dt^2} + K_P) & B_P & 0 & 0 & 0 & 0 \\
B_P^T & 0 & B^T R_t A^T & 0 & 0 & -L_f A^T \\
0 & A R_f^T B & M \frac{d^2}{dt^2} & 0 & 0 & 0 \\
0 & G^T R_f^T B & 0 & (M_G \frac{d^2}{dt^2} - 2 M_G) & 2q & 0 \\
0 & 0 & 0 & 2q^T & 0 & 0 \\
0 & -A L_f^T & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
d \\
\lambda \\
X_0 \\
q \\
\mu \\
X_f
\end{bmatrix}
= \begin{bmatrix}
P (f - M \Omega R_r G \dot{q} - 2 \Omega \dot{M} \dot{d}) \\
- B_t^T r \\
A R_f^T f \\
G^T R_f^T f - 2 R_f^T \Omega \dot{M} \dot{d} \\
0 \\
0
\end{bmatrix}
$$

(27)
where all terms can be clearly identified; first equation is the deformational part of the FEM elastic equations, second equation imposes the interface compatibility condition between the substructure boundary and the frame, third equation represents the translational global equilibrium condition expressed in the inertial frame of reference, fourth equation represents the rotational equilibrium condition expressed in the floating-frame of reference, fifth equation enforces unity of the quaternions and last equation represents the frame equilibrium condition expressed in the inertial frame.

5 FLUID FORMULATION

The fluid is considered to be inviscid and incompressible, confined in a carrier structure under the action of gravity field \( g \), initially at rest with density \( \rho_0 \) and with initial hydrostatic pressure due to gravity \( p^0 = p_{\text{ext}} + \rho_0 g (H - z) \). Next, we define in the body-frame a Lagrangian deformational displacement field \( d \) following the fluid particle and assume small deviations from this equilibrium position, as represented in Figure 3.

\[
\delta W^{(e)}_d = \rho^0 c^2 V_f (\mathbf{\nabla} \cdot \mathbf{d})^{(e)} (\mathbf{\nabla} \cdot \mathbf{\delta d})^{(e)} + \int_{V^{(e)}_f} \rho^0 \mathbf{I} : (\mathbf{\nabla} \mathbf{d} \mathbf{\nabla} \mathbf{\delta d}) dV - \int_{V^{(e)}_f} \rho (\mathbf{A} - \mathbf{I}) g dV
\]

a linearized approximation around the initial hydrostatic equilibrium state, assuming that the fluid deviations from equilibrium are small and where the last term accounts for the rotation of the gravity field.

Element displacements are discretized as \( d = \phi d^{(e)} \), where \( \phi \) collects the element shape functions while \( d^{(e)} \) gathers nodal values of the element. The element stiffness matrix is
then composed of two terms:

\[ K^{(e)} = K^{(e)}_{ac} + K^{(e)}_{geo}, \]  

\[ K^{(e)}_{ac} = \frac{\rho c^2}{V_f} \int_{V_f}^{} (\nabla \cdot \phi)^T dV \int_{V_f}^{} (\nabla \cdot \phi) dV \]  

\[ K^{(e)}_{geo} = \int_{V_f}^{} p_f^T (\nabla \phi)^T (\nabla \phi) dV. \]

where \( K^{(e)}_{ac} \) is the acoustic stiffness matrix and \( K^{(e)}_{geo} \) is the geometrical stiffness matrix.

Upon assembling the element matrices, we arrive to a discrete variational for the complete fluid mesh analog to (21):

\[ \delta W_d = \delta d^T \{ K_P d - P(A - A^0)^T F_g \} = \delta d^T \{ K_P d - f \} \]  

in which \( K_P \) is the projected stiffness matrix of the fluid, \( d \) the vector of fluid deformational displacements and \( f \) the nodal forces increment due to gravity expressed in the body-frame.

6 FLOATING STRUCTURES

We study the particular case of a fluid contained in a flexible tank that is transported by a marine vessel. The hull of the ship is treated as a rigid body connected to the internal flexible structure and the influence of the external fluid into the system is introduced as an input. It is important to mention that the only external effect considered in the simulations are the buoyancy restoring forces and moments. Other hydrodynamic effects due to the interaction with the external fluid like, diffraction forces, added-mass, or viscous damping effects, are not considered and could be important in some applications.

The static stability condition of the ship under the effect of restoring forces, which are buoyancy and weight, is called the metacentric stability. Considering the roll motion of the ship, see Figure 4, there is a restoring moment in the form:

\[ m_\phi = -\rho g \nabla |G M_T| \sin(\phi) \]

where \( \rho \) is the density of water, \( \nabla \) is the displaced volume of water, \( G M_T \) is the transverse metacentric height, and \( \phi \) is the roll angle. Similarly, there is a longitudinal metacenter, \( G M_L \), which acts as center of rotation in case of disturbance in the pitch degree of freedom.

Heave motion of the ship is dominated by the effect of restoring forces. If the hull is assumed to have a shape of a rectangular prism, and a constant waterline area \( A_w \), which is the area enclosed by the curve at the intersection of the body and the water surface, then the vertical force that forms as a result of the deviation in the vertical position of the ship with respect to the equilibrium position, can be given by

\[ F_z = -\rho g A_w \delta Z_0 \]
Figure 4: Transversal restoring force on a floating structure due to roll rotation. Restoring moment is proportional to the distance between the center of gravity (G) and the transversal metacentre (MT) in the y-z plane.

Finally, the restoring forces vector may be obtained by the collection of these force components as follows:

$$f_h(X_0, q) = -\rho g A_w A^T A_Z (X_0 - X_0^b), \quad A_Z = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (35)$$

The buoyancy moments can be expressed:

$$m_h(q) = -\rho g \bigg\{ \begin{array}{c} |GM_T| q^T A_\phi q \\ |GM_L| q^T A_\theta q \\ 0 \end{array} \bigg\} \quad (36)$$

with constant matrices

$$A_\phi = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} ; \quad A_\theta = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$

Variation of the buoyancy forces and moments is obtained by differentiation from (35) and (36):

$$\Delta f_h(X_0, q) = -\rho g A_w A^T A_Z \Delta X_0 - \rho g A_w \tilde{f}_h G \Delta q \quad (37)$$

and the moments increment are obtained from:

$$\Delta m_h(q) = -2\rho g \bigg\{ \begin{array}{c} |GM_T| \\ 0 \\ 0 \end{array} \bigg\} q^T A_\phi + \bigg\{ \begin{array}{c} 0 \\ |GM_L| \\ 0 \end{array} \bigg\} q^T A_\theta \bigg\} \Delta q \quad (38)$$
7 SIMULATION OF SLOSHING IN FLOATING STRUCTURES

In this Section, a two-body problem is used to validate the proposed computational technique and demonstrate its potential. We consider the case of a rigid floating structure connected to a rigid tank transporting 75 m$^3$ of water. The geometrical properties and dimensions of the ship, including the exact position of the tank inside the ship, are presented in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>Mass</td>
</tr>
<tr>
<td>$I_\phi$</td>
<td>Rolling inertia</td>
</tr>
<tr>
<td>$I_\theta$</td>
<td>Pitching inertia</td>
</tr>
<tr>
<td>$I_z$</td>
<td>Pitching inertia</td>
</tr>
<tr>
<td>$\nabla$</td>
<td>Displaced volume of water</td>
</tr>
<tr>
<td>$A_w$</td>
<td>Water plane area</td>
</tr>
<tr>
<td>$GM_t$</td>
<td>Transverse metacentric height</td>
</tr>
<tr>
<td>$GM_l$</td>
<td>Longitudinal metacentric height</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Fluid density</td>
</tr>
<tr>
<td>$T_z$</td>
<td>Heave period</td>
</tr>
<tr>
<td>$T_\phi$</td>
<td>Roll period</td>
</tr>
<tr>
<td>$T_\theta$</td>
<td>Pitch period</td>
</tr>
</tbody>
</table>

Table 1: Ship design parameters. 3D cubical tank configuration. Mesh, dimensions and relative position of the tank inside the marine vessel.

Ship motion is induced by imposing an initial rigid-body rotational velocity to the ship $\omega^0 = \{0.1, 0.05, 0\}^T \text{rad/s}$. The transported fluid, with properties $\rho = 1000 \text{kg/m}^3$ and $c = 1500 \text{m/s}$, is modeled using 125 hexahedral finite elements with three degrees of freedom per node.

Implicit time integration of the equations of motion is performed by using a fixed time step $\Delta t = 0.01 \text{s}$, with Newmark parameters $(\gamma = \frac{1}{2}, \beta = \frac{1}{4})$ combined with a dissipation parameter $\alpha = 0.01$. Figure 5 shows different positions in time of the hull and the free-surface of the fluid due to the prescribed rolling-pitching initial velocity. Time history of the tank displacement and free-surface elevation is given in Figure 5 for two points A and B aligned in the vertical direction near the wall.

8 CONCLUSIONS

- A fully implicit partitioned finite-element formulation for the analysis of structure-structure and fluid-structure systems presenting large rotations and small deformations has been presented. The proposed computational framework is based on the floating frame of reference approach separating rigid-body motions from de-
formational displacements and uses localized Lagrange multipliers to satisfy the constraints.

- It has been demonstrated that this new procedure is very well suited for modeling moderate sloshing phenomena in coupled carrier-internal fluid problems where the carrier presents large translations and rotations while deformations in the fluid can be considered small. A numerical example is used to demonstrate the accuracy, robustness, and efficiency of the proposed solution algorithm.

REFERENCES


A COUPLED AND MULTI-SCALE FLUID-STRUCTURE INTERACTION AND MASS TRANSFER MODEL FOR BIOFILM GROWTH SIMULATIONS

MIRELLA CORONEO, LENA YOSHIHARA AND WOLFGANG A. WALL

Institute for Computational Mechanics (LNM)
Technische Universität München
Boltzmannstr. 15, 85747 Garching b. München, Germany
e-mail: {coroneo, yoshihara, wall}@lnm.mw.tum.de, web page: http://www.lnm.mw.tum.de

Key words: biofilm growth, mass transport, fluid-structure interaction, multi-scale model, finite element method

Abstract. The formation and development of different biofilm structures is known to be influenced by nutrients availability and flow conditions. For this reason an approach which takes into account the effect of local structure deformation and fluid flow on mass transfer is essential for the understanding of biofilm macro-scale dynamic. The objective of the present work is therefore to study the fluid-structure interaction (FSI) and the substrate transport and reaction of big growing biofilm aggregates, for which continuum models can be applied. For this purpose we propose a novel growth model for the simulation of biofilm structures development. It is based on a finite element approach, developed in our in-house research code, for the numerical simulation of a sequential one-way coupling of the FSI and the scalar transport models [1]. The biofilm growth is coupled to the other processes through a multi-scale approach and takes into account the effects of mass transfer and shear stresses. First numerical examples are run at the purpose to demonstrate the suitability of the growth model to catch the main features of growing biofilm structures. This type of approach can give an important contribution to the understanding of biofilm architectures living in a complex environment. It allows to study the development of complex and real-life biofilm structure shapes often seen in nature and industry, to understand the influence of operating conditions, and therefore can enable the control of biofilm behaviour.

1 INTRODUCTION

Biofilm architecture, mechanics and interaction with surrounding fluid have a profound influence on their behaviour and potential treatments [2]. For the purpose of being able to control their formation and development, it results then extremely important to understand the mutual interaction of all the involved mechanisms. There are, however,
still a lot of points which need to be properly investigated, mainly because of the many difficulties in carrying out experiments which could isolate each single mechanism.

In this context, biofilm modelling is arising as a mean of producing quantitative tools [3] and a way to study biofilm behaviour. From the mathematical point of view, biofilm models can be divided in three main categories: cellular automata, individual-based models and continuum models [4]. Among them the last one is suitable to study bigger biofilm aggregates. The assumptions of continuum models that the biomass concentration can be adequately described by one or more density fields and that each density field obeys to some conservation law [5] can be considered satisfied for big aggregates. Indeed, while constituted of micro-scale objects, biofilm structures interact with the surrounding fluid as macro-scale materials and can be studied as flexible structures located in a moving liquid flow. Modelling approaches available at present however result to be limited to fixed biofilm geometries, neglecting biofilm deformation and growth. Only few works account for shear-induced deformation of the biofilm structure [1, 6, 7, 8, 9], but none of them takes in account the effect of local structure deformation and fluid flow on both mass transfer and growth.

For the purpose of understanding the influence of operating conditions on biofilm growth and erosion and of being able to control their behaviour, we propose a novel multi-scale computational approach, developed in our in-house research code BACI, taking in account both the fluid-structure interaction and the transport processes [1], and the effect they have on biofilm growth. Biofilm interaction with the surrounding fluid is modelled through an Arbitrary Lagrangian-Eulerian (ALE) approach, while the mass transfer is calculated through the solution of the dynamic convection-diffusion-reaction equation, assuming a non-linear Monod kinetic. The growth model is coupled to the fluid-structure interaction and substrate transport taking in account the different phenomena time-scales and calculates the growth based on the local characteristics of the biofilm structures. In this way it is possible to properly predict the effect of local biofilm deformation and of fluid flow on the transport and reaction of nutrients, and the influence that transport processes and shear stresses have on biofilm growth.

The governing equations of the fluid-structure interaction (FSI) and mass transport subproblems are summarized in sections 2 and 3, respectively. Afterwards, the growth model is presented in section 4, while in section 5 a description of the multi-scale algorithm is reported. Some selected numerical examples are then presented in section 6 before conclusions and outlook, provided in section 7.

2 FLUID-STRUCTURE INTERACTION

Biofilm interaction with the surrounding fluid is modelled through an arbitrary Lagrangian-Eulerian (ALE) approach. As a consequence, the FSI problem domain consists of three fields: the fluid $\Omega^F$ and the structure $\Omega^S$ domains, sharing a common interface $\Gamma$, and a third, non-physical mesh field produced by the ALE approach $\Omega^{G,F}$, on which a different equation has to be solved.
Biofilms are well-known to be viscoelastic materials, i.e. they behave elastically, when subjected to external forces over short periods of time, and in a viscoelastic way, when external forces are applied at longer time periods. Since the fluid-structure interaction is considered only at very small time scales, the structural motion of the biofilm is safely described by an elastic material model. In the biofilm domain the following non-linear structural elastodynamic equation is solved at the purpose to obtain the displacement field \( \mathbf{d}^S \)

\[
\rho^S \frac{d^2 \mathbf{d}^S}{dt^2} = \nabla \cdot (\mathbf{F} \cdot \mathbf{S}) + \rho^S \mathbf{b}^S \quad \text{in} \quad \Omega^S \times (0,T),
\]

where \( \rho^S \) denotes the structural density, \( \Omega^S \) is the undeformed domain while \( \mathbf{b}^S \) represent the external body forces. The internal forces are expressed in terms of the second Piola-Kirchhoff stress tensor \( \mathbf{S} \) and the deformation gradient \( \mathbf{F} \). At boundaries Dirichelet or Neumann conditions are imposed and the initial boundary value problem is completed by imposing appropriate initial conditions. In case of displacement-based finite element (FE) formulation, equation (1) is multiplied by the virtual displacements and then integrated by parts. The resulting weak form is the starting point for the spatial discretization.

In the deforming fluid domain, the fluid flow is considered to be laminar, since biofilm structures usually grow attached to surfaces and are considered to be in the hydrodynamic boundary layer. As a result, in the fluid domain the following incompressible time-dependent ALE version of Navier-Stokes equations are solved for both pressure \( \rho^F \) and velocity \( \mathbf{u}^F \) fields

\[
\rho^F \frac{\partial \mathbf{u}^F}{\partial t} + \rho^F (\mathbf{c}^F \cdot \nabla) \mathbf{u}^F - 2\mu \nabla \cdot \varepsilon(\mathbf{u}^F) + \nabla p^F = \rho^F \mathbf{b}^F \quad \text{in} \quad \Omega^F \times (0,T),
\]

\[
\nabla \cdot \mathbf{u}^F = 0 \quad \text{in} \quad \Omega^F \times (0,T).
\]

In the momentum equation (2), \( \varepsilon(\mathbf{u}^F) \) denotes the strain rate tensor of the fluid, \( \mu \) is its dynamic viscosity, \( \mathbf{b}^F \) a prescribed body force, and \( \mathbf{c}^F \) the fluid ALE convective velocity, representing the fluid velocity \( \mathbf{u}^F \) relative to the arbitrarily moving fluid domain

\[
\mathbf{c}^F = \mathbf{u}^F - \mathbf{u}^{G,F}.
\]

The fluid grid velocity \( \mathbf{u}^{G,F} \) is defined by

\[
\mathbf{u}^{G,F} = \frac{\partial \varphi}{\partial t} \quad \text{in} \quad \Omega^F \times (0,T),
\]

where \( \varphi \) represents a unique, arbitrary mapping for the deformation of the fluid domain \( \mathbf{d}^{G,F} \)

\[
\mathbf{d}^{G,F}(\mathbf{x},t) = \varphi \left( \mathbf{d}^{G,F}_t, \mathbf{x},t \right) \quad \text{for} \quad (\mathbf{x},t) \in \Omega^F \times (0,T),
\]
while \( d_{G,F} \) represents the mesh interface displacement, later related to the structure interface displacement \( d_{S} \). Also for the fluid domain Dirichlet and Neumann boundary conditions are imposed together with a divergence-free initial velocity field. For the fluid field the weak form is obtained by multiplying equations (2) and (3) with test functions for the velocity and pressure and then integrating by parts.

In order to calculate the fluid ALE convective velocity \( c^F \) appearing in equation (2), it is necessary to define the mapping \( \varphi \) introduced in equation (6). For this purpose, the boundary of the fluid ALE mesh is coupled to the Lagrangian mesh of the structures and to an Eulerian mesh at the in- and outflow portions. Within the domain, it is allowed to deform arbitrarily and, in the present study, it was chosen to treat the fluid ALE field as a quasi-elastostatic pseudo-structure [10]. The ALE equation of motion results then to be

\[
\nabla \cdot \sigma_{G,F} = 0 \quad \text{in} \quad \Omega_{G,F} \times (0, T)
\]

with \( \sigma_{G,F} \) defined as in [1]. Due to the continuous position change of the fluid-structure interaction surface, \( \Gamma \), and to the ALE formulation, the mesh will be continuously deformed and this deformation is controlled by kinematic and dynamic constraints imposed at the interface \( \Gamma \). First of all, an equilibrium of forces has to be fulfilled, resulting in equal surface tractions at the fluid and structure surface; second, the fluid grid velocity \( u_{G,F}^\Gamma \) and the fluid velocity \( u_F^\Gamma \) have to match at the interface. In addition, the fluid velocity, \( u_F^\Gamma \), is imposed to be equal to the structure deformation rate, since a mass flow across \( \Gamma \) as well as a relative tangential movement of fluid and structure at \( \Gamma \) are prohibited. For insights in the formulation of the weak forms and of their discretizations refer to Yoshihara et al. [1].

3 SCALAR TRANSPORT

The scalar transport has to be solved on the fluid and on the solid domain, which also in this case share a common mass transfer interface \( \Gamma \). For the calculation of the scalar field, the water-substrate solution is considered diluted and in the fluid domain \( \Omega^F \) the following convection-diffusion equation is solved for \( \Phi^F \)

\[
\frac{\partial \Phi^F}{\partial t} + c^F \cdot \nabla \Phi^F - \nabla \cdot (D^F \nabla \Phi^F) = 0 \quad \text{in} \quad \Omega^F \times (0, T).
\]

In the previous equation \( D^F \) is the fluid diffusion coefficient and \( c^F \) is the ALE convective velocity, which has already been introduced in equation (4).

On the other hand, in the solid domain \( \Omega^S \) the following diffusion-reaction equation in the conservative form is solved

\[
\frac{\partial \Phi^S}{\partial t} + \Phi^S (\nabla \cdot u^S) - \nabla \cdot (D^S \nabla \Phi^S) + R^S = 0 \quad \text{in} \quad \Omega^S \times (0, T).
\]

Here, \( D^S \) is the solid diffusion coefficient, \( u^S \) denotes the solid velocity, while \( R^S \) represents the reaction rate term. The microbial growth rates in an aqueous environment is usually
related to the concentration of a limiting nutrient through the Monod kinetic and the substrate consumption $R^S$ reads

$$R^S = k \frac{\Phi^S}{K + \Phi^S},$$

(10)

where $k$ represents the reaction rate constant and $K$ the half-saturation concentration of substrate.

The convection-diffusion transport of solute in the liquid domain, solved on spatial coordinates, is coupled to the diffusion-reaction transport equation in the biofilm domain, solved on material coordinates, with the constraint of equal concentrations and fluxes at the fluid-structure interface. The initial boundary value problem is completed through the imposition of Dirichlet and Neumann conditions at the boundaries of the domain and with appropriate initial concentration fields. Also in this case the weak form is obtained multiplying equations (8) and (9) with the virtual concentrations and integrating by parts. For insights in the formulation of the weak forms and of their discretizations refer also in this case to Yoshihara et al. [1].

4 THE GROWTH MODEL

The formation and development of different biofilm structures are known to be influenced by nutrients availability and flow conditions [11]. For this reason, a reliable modelling of biofilm growth has to take in account both variables connected to mass transfer as well as to flow conditions. The present novel biofilm model uses stresses and mass fluxes at the interface resulting from the FSI and mass transfer step to calculate the local amount of growth and erosion. The growth is calculated in term of displacement perpendicular to the interface in the following way

$$\tilde{d}^S = K_1 J^S - K_2 \sigma^S \text{ on } \Gamma \times (0, T).$$

(11)

Here $K_1$ and $K_2$ are constants, while $J^S$ and $\sigma^S$ are the mass flux through the interface and the shear stress at the interface, respectively. The latter two variables can be evaluated at the end of the calculation of the coupled FSI and mass transfer model, if it reaches a stationary solution, or as an average over a fixed period of time, if the problem reaches a periodic steady state solution, as it is the case for flapping streamers [9].

In order to permit biofilm to grow and the structure mesh to be appropriately deformed, in this step an ALE approach is applied also to the structure. For this purpose, the interface displacement due to growth calculated from equation (11) is applied as a Dirichlet boundary condition on the grid structure interface

$$d^{G,S} = \tilde{d}^S \text{ on } \Gamma \times (0, T),$$

(12)

while all the other boundaries of the structure ALE mesh are considered fixed. Within the domain, the ALE field is treated also in this case as a quasi-elastostatic pseudo-structure.
and an equation of motion similar to equation (7) is solved on the structure ALE mesh

\[ \nabla \cdot \sigma^{G,S} = 0 \quad \text{in} \quad \Omega^{G,S} \times (0, T). \]  

(13)

5 MULTI-SCALE COUPLED ALGORITHM

Monolithic schemes were found to be the most stable and efficient approach to model complex biological problems involving the coupling of incompressible flows to soft structures [12, 13]. For this reason the fully coupled non-linear FSI problem is solved monolithically. Moreover, since in biofilm applications the fluid flow and the structural deformation are not influenced by mass transport processes, a one-way coupling of fluid-structure interaction and transport models has been applied, as reported in [1]. For what concern the growth model, since the different involved phenomena happen in different time-scales, a multi-scale algorithm was applied, consisting of (i) an inner timeloop solving FSI and scalar transport at fluid-dynamic time-scale, (ii) an outer timeloop solving only the biofilm growth at biological time-scale.

Hence, in each fluid dynamic time step, at first the non-linear equations governing the fluid flow and the structure displacement, with the mentioned boundary conditions, are solved monolithically till when residuals meet a problem-specific tolerance. Subsequently, the local deformations and velocities obtained from the FSI calculation are transferred to the mass transport subproblem and the dynamic convection-diffusion-reaction equations for the mass transport are solved on the deformed fluid and solid domains till when also in this case residuals meet a problem-specific tolerance. These steps are repeated till when a stationary or a periodic steady state condition for the flow and concentration fields is reached. Afterwards, information regarding the mass flux through the interface and the shear stresses on it are transferred to the growth and erosion subproblem. At this point, the correct amount of growth is calculated for a longer time step and applied to the structure domain.

6 NUMERICAL EXAMPLES

The proposed simulations are based on the implementation of the algorithm discussed above in our in-house research code BACI [14]. The numerical solution of the field equations is obtained discretizing in space through a finite element method and in time through implicit time integration schemes. The resulting set of non-linear algebraic equations is then solved using a Newton-type method. A one-step-\( \theta \) time integration scheme with \( \theta = 0.66 \) is used for time discretization of fluid, structure and transport equations, while for space discretization trilinear, hexahedral finite elements are used. The presented examples are based on three-dimensional models and discretizations, although a pseudo two-dimensional deformation and flow state is enforced by specific boundary conditions.

The presented approach for coupling FSI and mass transport has been already successfully applied to the simulation of flapping biofilm streamers [9, 1] for flow conditions.
similar to those experimentally investigated by Stoodley et al. [15]. Those simulations have demonstrated the importance and the suitability of the proposed coupling of FSI and mass transport for simulating convective and diffusive mass transport on coupled and deformable fluid and solid domains under real operating and material conditions. For these reasons, the numerical examples proposed in the present study focus only on the suitability of the multi-scale growth model to simulate simple growing biofilm structures, and in particular on the possibility of the proposed model to catch the effect of operating conditions on biofilm growth, giving minor importance to real material parameters.

The presented examples reproduce a section of a fluid channel, where a uniform biofilm layer is present on the wall. The fluid domain $\Omega^F$ (dimensions $10\text{mm} \times 8\text{mm} \times 0.2\text{mm}$) is bound by a biofilm structure (dimensions $10\text{mm} \times 2\text{mm} \times 0.2\text{mm}$) at the bottom, as reported in Figure 1. The interface between the biofilm and the fluid is denoted by $\Gamma$ and represents the interface for FSI, mass transfer and biofilm growth. For the purpose of highlighting the effect of operating conditions on biofilm growth, different boundary conditions are applied to the same domains, while material parameters are taken constant. In all the presented cases, only the interface $\Gamma$ is allowed to move in any direction, while all the other boundaries are kept fixed. When not specified also a zero velocity field is applied at the fluid boundaries and a zero-flux conditions at the solid and fluid boundaries. As initial condition always a zero velocity field is prescribed. Differences between the three presented cases are reported in the following.

(a) In the first simulation a constant unitary concentration is applied at the top of the fluid domain and as initial condition a linear concentration gradient in the vertical direction is prescribed on all the domain.

(b) In the second simulation, a cosine profile with a unitary maximum value at its center is applied at the top boundary of the fluid domain

$$\Phi^F(x, y, z) = 0.5 + 0.5 \cos (0.2\pi x)$$
with \(-5\text{mm} \leq x \leq 5\text{mm}\) and \(-0.1\text{mm} \leq z \leq 0.1\text{mm}\). While as initial condition the following function is prescribed
\[
\Phi^F(x, y, z) = [0.5 + 0.5 \cos (0.2\pi x)] [1 + \cos (0.05\pi y + 1.25\pi)]
\]
with \(-5\text{mm} \leq x, y \leq 5\text{mm}\) and \(-0.1\text{mm} \leq z \leq 0.1\text{mm}\).

(c) Finally, the third simulation is similar to the second one except for the fact that in this case also a non-zero velocity boundary was applied at the left fluid boundary. A linear velocity gradient in the vertical direction with maximum value of \(0.1\frac{\text{mm}}{\text{ms}}\) is imposed to the velocity \(x\)-component
\[
\nu^F_x(x, y, z) = 0.0375 + 0.0125y
\]
with \(-5\text{mm} \leq y \leq 5\text{mm}\) and \(-0.1\text{mm} \leq z \leq 0.1\text{mm}\). Consequently also flow through the right fluid boundary is allowed and a slip boundary condition is prescribed at the top boundary.

For running these simulations a multi-scale approach is applied and 100 steps with a time step of 1 ms are used in the inner timeloop and were found to be sufficient to reach a steady-state solution. After that a growth step is calculated based on the last calculated values and with a biological time step of 1000 ms. In Figure 2 the substrate distribution in the fluid domain for the three simulations and at different biological time steps is reported together with the biofilm structure grid. Results highlight the influence of the boundary conditions on the final biofilm shape. As a matter of fact in simulation (a) the structure growth is uniform, while in simulation (b) the presence of a non-uniform nutrients source produces higher growth of the initial flat biofilm shape in the center of the biofilm structure. Finally, in case (c) the presence of the velocity field and also its effect on the substrate distribution produces a non-symmetric final biofilm shape.

7 CONCLUSION AND OUTLOOK

To enable the investigation of the influence of operating conditions on biofilm growth, we have developed an advanced computational model for the multi-scale and coupled numerical simulation of fluid-structure interaction and mass transfer of moving and growing biofilm structures. For this purpose, the one-way coupling of the fully coupled non-linear FSI problem and nonlinear multi-field mass transport model is encapsulated in a multi-scale approach. In this way it is possible to calculate FSI and mass transfer at hydrodynamic time scale and biofilm growth at biological time scales. The new methodology presented in this paper enables the mutual coupling of (i) flow and solid deformation, (ii) transport processes on deformed fluid and solid domains, (iii) as well as growth, based on mass transport and interfacial stresses. Hence, our approach can provide insights into how local deformations influence transport processes in both fluid and solid fields and how both fluid dynamic and mass transfer influence biofilm growth.
Selected examples demonstrated the general suitability of the proposed model for modeling the fluid-structure interaction and the mass transfer of growing biofilm structures. Ongoing work is concerned with an appropriate calibration of the proposed growth model based on experimental data, while outlook will concern the application of the presented coupled and multi-scale approach to reproduce experimental biofilm development at different operating conditions and the utilization of the developed approach as a prediction tool also in conditions difficult to investigate experimentally.

REFERENCES


A MODEL FOR THERMO-HYDRO-MECHANICAL ANALYSIS OF
MULTIPHASE POROUS MEDIA IN DYNAMICS

MAREVA PASSAROTTO AND LORENZO SANAVIA

Dipartimento di Ingegneria Civile, Edile ed Ambientale (DICEA)
Università degli Studi di Padova
Via F. Marzolo 9, 35131 Padova, Italy
e-mail: mareva.passarotto@unipd.it, lorenzo.sanavia@unipd.it, www.dicea.unipd.it

Key words: Dynamics analysis, Variably saturated porous media, Non-isothermal problems, Thermo-hydro-mechanical coupling.

Abstract. This work presents the development of a mathematical and numerical model for the analysis of the thermo-hydro-mechanical behaviour of multiphase porous materials in dynamics. The fully coupled multiphase model for non-isothermal deformable porous media is developed within the Hybrid Mixture Theory. In order to analyse the thermo-hydro-mechanical behaviour of soil structures in the low frequency domain, e.g. under earthquake excitation, the u-p-T formulation is advocated neglecting the relative fluids acceleration and their convective terms. Moreover, the dynamic seepage forcing terms and the compressibility of the solid grain at the microscopic level are neglected. The standard Bubnov-Galerkin method is applied to the governing equations for the spatial discretization, whereas the generalized Newmark scheme is used for the time discretization. The final algebraic, non-linear and coupled system of equations is solved by the Newton method within the monolithic approach. The formulation and the implemented solution procedure are validated through the comparison with other finite element solutions or analytical solutions when available.

1 INTRODUCTION

The analysis of the dynamic response of multiphase porous media has many applications in civil engineering. Onset of landslide due to earthquake or rainfall and the seismic behaviour of earth dams are just few examples where inertial forces cannot be neglected because of the mass involved. Moreover, there are situations where it is important to consider the effect of temperature variation that causes a decrease of the solid skeleton strength and an increase of the pore water pressure. We could observe these phenomena for example during the onset of catastrophic landslide, as described in Vardoulakis [1], where the mechanical energy, dissipated in heat inside the slip zone, may also lead to the vaporization of pore water, creating a cushion of zero friction. Another example is the seismic behaviour of deep nuclear waste disposal, because an increment of temperature due to the failure of the canisters could create localized failure zones, which could act as preferential escape zones for the fluids containing radionuclides.

Many authors have developed models for the analysis of the dynamic behaviour of multiphase porous media in isothermal conditions. A state of art can be found in Zienkiewicz et al. [2] and Schanz [3]. Recently, Nenning and Schanz [4] presented an infinite element for

This work presents, as a novel contribution, a formulation of a fully coupled model for deformable multiphase geomaterials in dynamics including thermal effects.

The multiphase model is developed following Lewis and Schrefler [7]. The u-p-T formulation is obtained by neglecting the relative fluids acceleration and their convective terms, which is valid for low frequency problems as in earthquake engineering [2]. In the model development, the dynamic seepage forcing terms in the mass balance equations and in the enthalpy balance equation and the compressibility of the grain at the microscopic level are neglected. The implemented model is validated through the comparison with analytical or finite element quasi-static and dynamic solutions.

2 MACROSCOPIC BALANCE EQUATIONS

The full mathematical model necessary to simulate the thermo-hydro-mechanical behaviour of partially saturated porous media was developed within the Hybrid Mixture Theory by Lewis and Schrefler [7], Gawin and Schrefler [8] using averaging theories according to Hassanizadeh and Gray [9],[10]. The partially saturated porous medium is treated as multiphase system composed of the solid skeleton (s) and voids filled with liquid water (w) and gas (g); that the latter is assumed to behave as an ideal mixture of dry air (ga) and water vapour (gw). At the macroscopic level the porous material is modelled by a substitute continuum of volume Β with boundary ∂Β that simultaneously fills the entire domain, instead of the real fluids and the solid which fill only a part of it. In this substitute continuum each constituent π has a reduced density which is obtained through the volume fraction \( η_π(x,t) = dv_π(x,t)/dv(x,t) \), where \( dv \) is the volume of the average volume element (representative elementary volume, REV) of the porous medium and \( dv_π \) is the volume occupied by the constituent \( π \) in \( dv \). \( x \) is the vector of the spatial coordinates and \( t \) the current time.

The solid is deformable and non-polar, and the fluids, solid and thermal fields are coupled. The constituents are assumed to be isotropic, homogeneous, immiscible except for dry air and vapour, and chemically non-reacting. At micro level, solid is incompressible, while liquid water and gas are considered compressible. Local thermal equilibrium between solid matrix, gas and liquid phases is assumed. Heat conduction and convection, vapour diffusion, water flow due to pressure gradients or capillary effects and water phase change (evaporation and condensation) inside the pores are taken into account. In the partially saturated zones the liquid water is separated from its vapour by a meniscus concave toward gas. Due to the curvature of this meniscus the sorption equilibrium equation is assumed valid (for simplicity) and gives the relationship \( p_c = p_g - p_w \) between the capillary pressure, gas pressure and liquid water pressure.

The general field equations of the model are now written at macroscopic level in the geometrical linear setting.

The primary variables are chosen to be the displacements of the solid matrix, \( u(x,t) \), the capillary and gas pressure, \( p_c(x,t) \) and \( p_g(x,t) \), and the absolute temperature, \( T(x,t) \).

The small terms related to relative accelerations of the fluids and their convective terms are neglected following [2]. This approximation is valid for dynamics at lower frequencies, as in earthquake engineering [2], [7], [11]. Dynamic seepage forcing terms connected with the solid acceleration are also neglected because their contribution is very small compared with other
terms [2] (the effect of dynamic seepage can be of importance in the high frequency range where the \( u-p \) formulation is no longer valid [2]).

With the assumptions described above we obtain the \( u^c-p^g-T \) formulation of the balance equations to be implemented [12]:

**Linear momentum balance equation of the mixture**

\[
\text{div} \, \sigma + \rho g = \rho a^g
\]  

(1)

**Water species (liquid and vapour) mass balance equation**

\[
\begin{align*}
\rho^w \frac{n S_w}{K_w} \frac{\partial p^w}{\partial t} + \left[ \rho^w S_w + \rho^w S_g \alpha \text{div} \, v^s - \beta_{\text{reg}} \frac{\partial T}{\partial t} + n[\rho^w - \rho^v] \frac{\partial S^w}{\partial t} + n S_e \frac{\partial \rho^v}{\partial t} \right] \\
+ \text{div} \, J^g + \text{div} \left( \rho^w \frac{k^v}{\mu^w} \left[ - \text{grad} \, p^w + \rho^v \mathbf{g} \right] \right) + \text{div} \left( \rho^v \frac{k^v}{\mu^v} \left[ - \text{grad} \, p^v + \rho^v \mathbf{g} \right] \right) = 0
\end{align*}
\]  

(2)

**Dry air mass balance equation**

\[
S_g \text{div} \, v + n S_e \frac{\partial \rho^e}{\partial t} + \frac{1}{\rho^e} \text{div} \, J^e - n \frac{\partial S^w}{\partial t} + \frac{1}{\rho^e} \text{div} \left( \rho^e \frac{k^v}{\mu^e} \left[ - \text{grad} \, p^g + \rho^g \mathbf{g} \right] \right) + n \beta_g \frac{\partial T}{\partial t} = 0
\]  

(3)

**Enthalpy balance equation for the multiphase medium**

\[
\begin{align*}
\left( \rho C_p \right)_{\text{eff}} \frac{\partial T}{\partial t} - \text{div} \left( \chi_{\text{eff}} \text{grad} \, T \right) - \Delta H_{\text{vap}} \rho^w S_w \alpha \text{div} \, v^s - \Delta H_{\text{vap}} \rho^w \frac{n S_w}{K_w} \frac{\partial p^w}{\partial t} \\
+ \left[ \rho^w C_p n S_w \frac{k^v}{\mu^w} \left[ - \text{grad} \, p^w + \rho^w \mathbf{g} \right] \right] \cdot \text{grad} \, T + \Delta H_{\text{vap}} \beta_{\text{sw}} \frac{\partial T}{\partial t} \\
+ \left[ \rho^e C_p n S_e \frac{k^v}{\mu^e} \left[ - \text{grad} \, p^g + \rho^g \mathbf{g} \right] \right] \Delta H_{\text{vap}} = 0
\end{align*}
\]  

(4)

The meaning of each variable of equations 1-4 are described in [7], [12] or [13].

**3 CONSTITUTIVE RELATIONSHIP**

For the gaseous mixture of dry air and water vapor, the ideal gas law is introduced. The equation of state of perfect gas (Clapeyron’s equation) and Dalton's law are applied to dry air (\( \text{ga} \)), water vapor (\( \text{gw} \)) and moist air (\( g \)).
In the partially saturated zones, the equilibrium water vapor pressure \( p_{gw}(x,t) \) can be obtained from the Kelvin-Laplace equation, where the water vapor saturation pressure, \( p_{gw,s} \), depending only upon the temperature, can be calculated from the Clausius-Clapeyron equation or from an empirical correlation. The saturation degree \( S_s(x,t) \) and the relative permeability \( k_{r,π}(x,t) \) are experimentally determined functions.

The solid skeleton is assumed elastic, homogeneous and isotropic in the numerical simulations described in Section 5.

4 SPATIAL AND TIMES DISCRETIZATION

The finite element model is derived by applying the Galerkin procedure for the spatial integration and the Generalized Newmark Method for the time integration of the weak form of the balance equations of the previous section [2], [7], [14]. In particular, after spatial discretization within the isoparametric formulation, the following non-symmetric, non-linear and coupled system of equations is obtained [12]:

\[
\begin{align*}
\int_{Ω} B^T \sigma' \, dΩ - Q \bar{p}^c + R \bar{p}^c + M \ddot{u} &= f_u \\
U \dddot{p}^c + S \dddot{p}^c + T \dddot{T} + Q' \dddot{u} - I \dddot{p}^c + H \dddot{p}^c + E \dddot{T} &= f_v \\
S' \dddot{p}^c - T' \dddot{T} + R' \dddot{u} + I' \dddot{p}^c - H' \dddot{p}^c - E' \dddot{T} &= f_g \\
T'' \dddot{T} - U' \dddot{p}^c + S'' \dddot{p}^c - Q'' \dddot{u} + E'' \dddot{T} - L'' \dddot{p}^c + H'' \dddot{p}^c &= f \end{align*}
\]

(5)

where the displacements of the solid skeleton \( u(x,t) \), the capillary pressure \( p_c(x,t) \) and the temperature \( T(x,t) \) are expressed in the whole domain by global shape function matrices \( N_u(x), N_v(x), N_c(x), N_p(x) \) and the nodal value vectors \( \bar{u}(t), \bar{p}(t), \bar{p}(t), T(t) \). Following the Generalized Newmark Method, equations (5) are rewritten at time \( t_{n+1} \).

After time integration, the non-linear system of equation is linearized, thus obtaining the equation system that can be solved numerically:

\[
G(x_{n+1}) \equiv G(x_{n+1}) + \left. \frac{\partial G}{\partial x} \right|_{x=x_{n+1}} \, dx_{n+1} = 0
\]

(6)

where \( x = [\Delta \dddot{u} \quad \Delta \dddot{p}^c \quad \Delta \dddot{p}^c \quad \Delta \dddot{T}]_{n+1} \) is the vector of the unknown, \( G = [G_u \quad G_v \quad G_c \quad G_T]_{n+1} \) and \( \partial G/\partial X \) is the Jacobian matrix. Finally, the solution vector is updated by the incremental relationship \( X_{n+1} = X_{n+1} + \Delta X_{n+1} \).

5 VALIDATION AND APPLICATION EXAMPLES

In the following the numerical validation of the finite element model is presented by solving two numerical tests [12].
5.1 Numerical validation of the non-isothermal water saturated model

This problem deals with a fully saturated thermo-elastic consolidation problem [15], simulating a column, 7 m high and 2 m wide, of a linear elastic material subjected to an external surface load of 1000 Pa and to a surface temperature jump of 50 K above the initial temperature of 293.15 K (Figure 1). The material parameters used in the computation are summarised in [16]. The liquid water and the solid grain are assumed incompressible for the static analysis, whereas the compressibility of the liquid water is taken into account in the dynamic analysis. For the numerical calculation, the problem is solved as a two-dimensional problem in plane strain condition. The column is discretized with eight-node isoparametric elements (4 elements/meter); nine Gauss points are used.

![Figure 1: Geometry, loading conditions and finite element discretization of the saturated soil column](image)

The initial and boundary conditions are described in Table 1.

<table>
<thead>
<tr>
<th>Initial condition</th>
<th>Boundary condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P^g = P_{atm}$ fixed</td>
<td>$P^e = P_{atm}$ fixed</td>
</tr>
<tr>
<td>$P^c = $ idrostatic fixed</td>
<td>$P^c = 0.0$ at the top</td>
</tr>
<tr>
<td>$T = 293.15$ K fixed</td>
<td>$T$ not fixed</td>
</tr>
<tr>
<td>$u_x = 0.0$ on the lateral nodes</td>
<td>$u_x = 0.0$ on the lateral nodes</td>
</tr>
<tr>
<td>$u_y = 0.0$ on the bottom</td>
<td>$u_y = 0.0$ on the bottom</td>
</tr>
</tbody>
</table>

The solution of the finite element model presented in this work is compared with the quasi-static solution [16] and is plotted in Figures 2-4.
Mareva Passarotto and Lorenzo Sanavia

Figure 2: Temperature time history for node 319 up to the steady state solution (a) and in the first period (b) highlighted in a)

Figure 3: Capillary pressure time history for node 319 up to the steady state solution (a) and (b) during the time highlighted in a)

Figure 4: Vertical displacement time history for node 319 up to the steady state solution (a) and in the first period (b) highlighted in a)

It can be observed that the dynamic solution is faster than the quasi-static one at the beginning of the analysis, and that, at the end of the analysis, the dynamic solution reaches the quasi-static one.
5.2 Drainage of liquid water from initially water saturated soil column

The proposed benchmark is based on an experiment performed by Liakopoulos [17] on a column, 1 meter high, of Del Monte sand and instrumented to measure the moisture tension at several points along the column during its desaturation due to gravitational effects. Before the start of the experiment, water was continuously added from the top and was allowed to drain freely at the bottom through a filter, until uniform flow conditions were established. Then the water supply was ceased and the tensiometer readings were recorded. The finite element simulation is performed with the two-phase flow model in isothermal conditions, with switching between saturated and unsaturated solution performed at \( p^c = 2000 \) Pa \((S_w = 0.998)\), which corresponds to bubbling pressure of the analysed sand, and an additional lower limit for the gas relative permeability of 0.0001 [8]. For the numerical calculation, a two-dimensional problem in plane strain conditions is solved; the spatial domain of the column is divided into 20 eight-node isoparametric finite elements of equal size. Furthermore, nine Gauss integration points were used. The material parameters are listed in [8] or [20].

This problem has been solved considering single or two-phase flow mainly in quasi-static condition; a finite element solution in dynamics was presented in [18]. The initial hydro-mechanical equilibrium state is obtained via a preliminary quasi-static solution. The initial and boundary conditions for the dynamic analysis are summarized in Table 2.

<table>
<thead>
<tr>
<th>Initial condition</th>
<th>Boundary condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P^g = P_{atm} ) on the top</td>
<td>( P^g = P_{atm} ) on the top, on the bottom</td>
</tr>
<tr>
<td>( P^c = ) idrostatic</td>
<td>( P^c = 0.0 ) on the bottom</td>
</tr>
<tr>
<td>( T = 293.15 ) K fixed</td>
<td>( T = 293.15 ) K fixed</td>
</tr>
<tr>
<td>( u_x = 0.0 ) on the lateral nodes</td>
<td>( u_x = 0.0 ) on the lateral nodes</td>
</tr>
<tr>
<td>( u_y = 0.0 ) on the bottom</td>
<td>( u_y = 0.0 ) on the bottom</td>
</tr>
</tbody>
</table>

The comparison between the dynamic and the quasi-static solution is plotted in Figures 5-7, where the profiles for water pressure, water saturation and vertical displacement along the column are plotted. Since the inertial loads are negligible in the experiment, the finite element solution in dynamics gives almost the same results of the quasi-static solution (Figures 5b-7b).

![Figure 5](image_url)

**Figure 5**: Profiles of water pressure versus height (a – dynamic solution) and comparison between the quasi-static (S) and the dynamic solution (D) at 5, 10, 20 and 30 minutes (b)
7 CONCLUSIONS

A model for the analysis of the thermo-hydro-mechanical behaviour of porous media in dynamics was developed. Starting from the generalized mathematical model developed by Lewis and Schrefler [7] for deforming porous media in non-isothermal conditions, the u-p-T formulation was derived following [2]. The validity of such an approximation is limited to low frequencies problems [2], as in earthquake engineering. In this formulation the relative accelerations of the fluids and the convective terms related to these accelerations are neglected. Moreover, in the model development, the compressibility of the solid grain at microscopic level and the dynamic seepage forcing terms were neglected.

The numerical model was derived within the finite element method: the standard Bubnov-Galerkin procedure [14] was adopted for the discretization in space, while the implicit and unconditionally stable Newmark procedure was applied for the discretization in time [14]. The independent variables chosen are: the displacement of the solid skeleton $u$, the capillary pressure $p^c$, the gas pressure $p^g$ and the temperature $T$.

The model was implemented in the finite element code Comes-Geo [7], [8], [13], [16], [19], [20], [21], [22]. The formulation and the implemented solution procedure were validated through the comparison with literature benchmarks, finite element solutions or analytical solutions [12]. In this paper, comparison between the finite element solution in dynamics and
the corresponding quasi-static solution is presented by studying the non-isothermal consolidation in a water saturated column and the drainage of liquid water in an initially water saturated soil column.

This work extends the model developed in [18] to non-isothermal conditions and removes the passive air phase assumption of the multiphase porous media model in dynamics developed in [2], [23], [24], [25] and used in [26] to study the seismic behaviour of an earth dam.

ACKNOWLEDGEMENTS

The authors would like to thank the University of Padua for financial support (projects: CPDA097373 and STPD08JA32).

REFERENCES


ANISOTROPIC DIFFUSION AND PROPAGATION OF SOUND WAVES IN POROELASTIC MEDIA

BETTINA ALBERS* AND KRZYSZTOF WILMANSKI†

* Department for Soil Mechanics and Geotechnical Engineering
Technische Universität Berlin (TU Berlin)
Sekr. TIB1-B7, Gustav-Meyer-Allee 25, 13355 Berlin, Germany
e-mail: albers@grundbau.tu-berlin.de, www.mech-albers.de

† TU Berlin, Germany and ROSE School Pavia, Italy
http://en.wikipedia.org/wiki/Krzysztof_Wilmanski

Key words: Anisotropy of tortuosity, Acoustic Waves, Poroelastic Media, Shear Polarization.

Abstract. In an earlier paper it was already suggested that the anisotropy of the tortuosity yields essential changes of the attenuation of the waves in poroelastic media depending on the propagation direction in relation to the principal directions of tortuosity and on the mode of the wave. In the region of low frequencies appearing in geotechnical applications the orientation of the principal directions of tortuosity plays a secondary role in measured speeds and attenuations and most likely cannot be used for practical purposes. However, due to the appearance of two shear waves anyway the construction of a device for measuring the anisotropy of the permeability may be possible. It would have to induce shear waves of different polarization and different propagation directions. Then, also for the range of low frequencies, one could measure the principal values and directions of the tortuosity by comparing the amplitudes of arrivals for different polarization of the signals.

1 INTRODUCTION

In elastic solids the anisotropy of a material (e.g. of wood or fibre-reinforced synthetics) enters a model through the constitutive law. The relation between stresses and deformations is described by a generalized Hooke law in which the elasticity tensor is of fourth order. The number of elastic constants (the coefficients of the elasticity tensor) can be reduced from 81 by considering special cases of anisotropy (e.g. for monoclinic symmetry to 13, for orthotropic symmetry to 9 or for transversal isotropy to 5). For isotropic solid materials only two independent elastic constants, e.g. the Lamé parameters \( \lambda \) and \( \mu \), remain.

In poroelastic materials the relative motion of fluid components with respect to the skeleton, i.e. the diffusion, is of high importance. This characteristic feature of a permeable porous material distinguishes porous materials from other multicomponent systems such as composites. The macroscopic permeability of a porous medium is influenced by the microstructure of the solid material, in particular, by the shape of the channels and their volume contribution to the total volume of the material. The permeability characterizes the intensity of diffusion. In the description of rocks and porous materials deformations of the
anisotropic skeleton are less important than anisotropic diffusion properties. Therefore, here, the anisotropy of the material is described not by anisotropic stress-strain relations but by anisotropic permeability. It is induced by a symmetric tensor of permeability defined by a permeability coefficient and an inverse of the tortuosity tensor introduced by Bear and Bachmat [4].

Four modes of propagation arise by use of this model. For the special choice of orientation of the propagation direction these are two pseudo longitudinal modes $P_1$ and $P_2$, one pseudo transversal mode $S_2$ and one transversal mode $S_1$. Speeds of propagation and the attenuation of these waves as well as the polarization properties in dependence on the orientation of the principal directions of the tortuosity are presented.

### 2 GOVERNING EQUATIONS INCLUDING ANISOTROPIC PERMEABILITY

The two-component model of a poroelastic material including anisotropic permeability has been already presented in [1] and [6]. Its linearized partial momentum balances have the form

$$\rho^S \frac{\partial v^S}{\partial t} = \text{div} \ T^S + \hat{p}, \quad \rho^F \frac{\partial v^F}{\partial t} = -\text{grad} \ p^F - \hat{p}, \quad (1)$$

where $\rho^S$ and $\rho^F$ are the initial constant partial mass densities of solid and fluid, $v^S$ and $v^F$ are the partial velocities of these two components. The partial stress tensor $T^S$, the partial pressure $p^F$ and the momentum source $\hat{p}$ are given by the constitutive relations

$$T^S = T_0^S + \lambda^S e_1 + 2\mu^S e_\perp + Q \varepsilon, \quad p^F = p_0^F - Qe - \rho^F \kappa, \quad \hat{p} = \pi_y (v_y^F - v_y^S) e_y,$$  

where $\lambda^S$, $\mu^S$, $\kappa$ and $Q$ are material parameters describing an isotropic poroelastic skeleton, an ideal fluid (with compressibility $\kappa$) and the coupling of both. Quantities with index zero are initial values and $e_i$ are Cartesian base vectors, i.e. $e_i \cdot e_j = \delta_{ij}$. With $e_i$ denoting the Almansi-Hamel tensor for small deformations of the solid, the volume changes of skeleton and fluid are $e = \text{tr} e^S$ and $e$, respectively. The following compatibility conditions are satisfied

$$\frac{\partial e^S}{\partial t} = \text{grad} v^S, \quad \frac{\partial e^F}{\partial t} = \text{div} v^F \quad (3)$$

The matrix $\pi_{ij}$ describes Cartesian components of the symmetric and positive definite permeability tensor. The influence of inertial forces attributed to added mass effects (i.e. the influence of relative accelerations) is neglected in this model. As pointed out in [7], in our opinion, it is erroneous to relate these forces to tortuosity effects as is often done in the literature. Moreover, such effects seem to be of higher order of magnitude.

The permeability tensor $\pi_{ij}$ is the product of a positive scalar which reflects the physical conditions of diffusion $\pi_0 = \frac{\nu^F}{D_0}$ and of the inverse of the tortuosity matrix $T_y^{-1}$ which is reflecting geometrical properties of the curvy channels. The former depends on the hydraulic diameter $D_0$, the initial porosity $n_0$, the true dynamic viscosity of the fluid in the pores $\mu^F$, the capillary shape factor $b$ and the earth gravity $g$. The latter has been introduced and discussed by Bear and Bachmat in [4]. If the direction $e_i$ is chosen as the principal direction $t_i$ of the symmetric tortuosity tensor the latter can be written in the following spectral form
The tortuosity tensor possesses three real eigenvalues \( \{ \tau^{(1)}, \tau^{(2)}, \tau^{(3)} \} \) and three corresponding eigenvectors \( \{ t_1, t_2, t_3 \} \). The so-called principle tortuosities \( \{ \tau^{(1)}, \tau^{(2)}, \tau^{(3)} \} \), according to Bear and Bachmat [4], measure an average inverse of the cosines of the angles between a short straight interval in a chosen principal direction and a streamline between the endpoints of this interval.

3 MONOCHROMATIC WAVES

In this paper, the Cartesian reference system is chosen in which the relations \( e_1 = t_1, \ e_2 \cdot t_1 = \sin \alpha, \ e_1 \cdot t_2 = \cos \alpha, \) and \( e_1 \cdot t_2 = -\sin \alpha \) between the base vectors and the principal directions of the tortuosity are satisfied. The propagation of monochromatic waves of a given frequency \( \omega \) in a two-component poroelastic material is investigated by use of the above presented model. The waves are solutions of the governing equations and result from the following presumptions

\[
\begin{align*}
\textbf{v}^S &= \textbf{V}^S \textbf{E}, & \textbf{v}^F &= \textbf{V}^F \textbf{E}, & \textbf{e} &= \textbf{E}^F \textbf{E}, & \textbf{E} &= e^{i(k \cdot x - \omega t)} = e^{i[(\text{Im} \textbf{n}) \cdot x] \text{Re} \textbf{e}^{i(k \cdot x - \omega t)}}, \\
\textbf{k} &= k \textbf{n}, & \textbf{n} &= \textbf{e}_1, & \textbf{n} \cdot \textbf{n} &= 1, & c_{ph} &= \frac{\omega}{\text{Re} k},
\end{align*}
\]

where \( \textbf{V}^S, \textbf{V}^F, \textbf{E}^S \) and \( \textbf{E}^F \) are constant amplitudes, \( \textbf{k} \) is the wave vector, \( k \) the wave number, \( \textbf{n} \) denotes the propagation direction and \( c_{ph} \) is the propagation speed of the monochromatic wave of frequency \( \omega \).

Applying (5) to Eq. (2) yields

\[
\begin{align*}
\textbf{E}^S &= -\frac{1}{2\omega} \left( \textbf{V}^S \otimes \textbf{k} + \textbf{k} \otimes \textbf{V}^S \right), & \textbf{E}^F &= -\frac{1}{\omega} \textbf{v}^F \cdot \textbf{k}.
\end{align*}
\]

From the momentum balances together with the constitutive relations the following eigenvalue problem arises

\[
\begin{align*}
&\left( -\rho^S \omega^2 \textbf{I} + \kappa^S \textbf{I} \otimes \textbf{k} + \mu^S \left( k^2 \textbf{I} + \textbf{k} \otimes \textbf{k} \right) - i \pi_\omega \textbf{T}^{-1} \right) \textbf{V}^S + \\
&\left( Q \textbf{k} \otimes \textbf{k} + i \pi_\omega \textbf{T}^{-1} \right) \textbf{V}^F = 0,
\end{align*}
\]

where it is assumed that \( \textbf{n} = \textbf{e}_1 \), Eq. (7) can be simplified:

\[
\begin{align*}
&\left( -\rho^S \omega^2 \textbf{V}^S + \kappa^S k^2 \textbf{e}_1 \textbf{V}^S + \mu^S \left( k^2 \textbf{V}^S + k^2 \textbf{V}^F \right) - i \pi_\omega \textbf{T}^{-1} \textbf{V}^S + \\
&\left( Q k^2 \textbf{e}_1 \textbf{V}^F + i \pi_\omega \textbf{T}^{-1} \textbf{V}^F \right) = 0,
\end{align*}
\]

It is assumed that one of the principal directions of \( \textbf{T} \) is perpendicular to \( \textbf{k} \), i.e. \( \textbf{k} = kn = ke_1 \) and \( t_1 = e_1 \). This simplifying assumption, which was also proposed in [1], causes that one of the propagation modes is purely transversal. In [6] the even simpler case was considered that the propagation direction \( \textbf{k} \) coincides with the direction \( \textbf{e}_1 \) and simultaneously with one of the
principal directions of the tortuosity tensor, i.e. $\alpha = 0$.

### 3.1 Decoupled transversal wave

Multiplication of Eqs. (3) with $e_3$ yields two equations for the two components $V_3^S$ and $V_3^F$. Combination of these equations yields the dispersion relation

$$\left( \omega^2 - \frac{k^2}{\rho^2} + \frac{i\pi_\alpha}{\rho^2} \tau(\omega) \right) \left( \omega^2 - \frac{k^2}{\rho^2} + \frac{i\pi_\alpha}{\rho^2} \tau(\omega) \right) + \frac{\rho^2}{\rho^2} \left( \frac{\pi_\alpha}{\rho^2} \tau(\omega) \right) = 0,$$

where (4) has been used and $V_3^S \neq 0$, $V_3^F \neq 0$. Since the remaining components of the vectors $V^S$ and $V^F$ are zero, the result is clearly a transversal wave, i.e. its amplitude is perpendicular to the propagation direction. The existence of such a purely transversal $S1$-wave follows from the assumption that the eigenvector of the tortuosity tensor $t_3$ is orthogonal to the propagation direction $n$. Relation (9) yields the phase speeds and the attenuations of the $S1$-wave which are illustrated in Figure 1 for different values of the principal tortuosity $\tau^{(3)}$.

![Figure 1: Speeds (left) and dimensionless attenuations (right) of the purely transversal wave appearing in Alermoehe sandstone for different values of the principal tortuosity $\tau^{(3)}$](image)

For the illustration of the wave properties in this paper roughly data of Arnold [3] for Alermoehe sandstone saturated by water (Table 1) are used. She reported on tortuosities in the interval between 1.06 and 6.50. Similar results are obtained by Wang et al. [5] who measured values between 3.45 and 7.69 for the tortuosity of various rocks. Combining these results, here, $\tau^{(3)}$ is chosen between 1 and 8.

The value of the shear modulus $\mu^s = \frac{4}{3}K_s(1 - 2\nu)/(1 + \nu)$ follows from Gassmann and Geertsma relations (for details see e.g. [2]).

### 3.2 Coupled waves

In order to achieve the phase speeds and attenuations of the coupled waves, Eqs. (3) have to be multiplied with $e_1$ and $e_2$. This yields four equations for the four unknown components $V_1^S, V_2^S, V_1^F$ and $V_2^F$. Thus, the set of equations $AX = 0$ has to be solved. The components of matrix $A$ depend on the frequency, on the angle $\alpha$ reflecting the anisotropy of the tortuosity
Table 1: Example of the construction of one table

<table>
<thead>
<tr>
<th>$K_s$</th>
<th>$K_f$</th>
<th>$\nu$</th>
<th>$n_0$</th>
<th>$\rho_s^{\text{true}}$</th>
<th>$\rho_f^{\text{true}}$</th>
<th>$\rho_f^{\text{part}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 GPa</td>
<td>2.25 GPa</td>
<td>0.2</td>
<td>0.3</td>
<td>2500 $\frac{\text{kg}}{\text{m}^3}$</td>
<td>1750 $\frac{\text{kg}}{\text{m}^3}$</td>
<td>300 $\frac{\text{kg}}{\text{m}^3}$</td>
</tr>
</tbody>
</table>

$K_s, K_f$ – true compressibility moduli of solid and fluid, $\nu$ – Poisson number, $n_0$ – initial porosity, $K_s = K_f/(1 + 50n_0)$, $\rho_s^{\text{true}}, \rho_f^{\text{true}}$ – true and partial mass densities of the solid, $\rho_f^{\text{part}}$ – partial mass density of the fluid, $\lambda^s, \mu^s$ – Lamé parameters, $\kappa$ – compressibility, $Q$ – coupling parameter, $K_c = [(1-n_0)/K_s + n_0/K_f]^{-1}$, $\pi_0$ – permeability coefficient.

(introduced in (4)) and on the material parameters \{$\rho^s, \rho^f, \lambda^s, \mu^s, \kappa, Q, n_0, \tau^{(1)}, \tau^{(2)}, \tau^{(3)}$\}. A complete presentation of Matrix A can be found in [1]. Vector $X$ is defined by $X = (V_s^1, V_s^2, V_f^1, V_f^2)^T$. The dispersion relation $\det (A) = 0$ determines the relation between $\omega$ and $k$. For a given real frequency $\omega$ three complex solutions for the wave number $k$ are obtained. From these follow the phase speeds $\omega/(\text{Re } k)$ and the attenuations $\text{Im } k$. The solutions, illustrated in Figure 2, besides on frequency and the values of principle tortuosities, depend on the angle $\alpha$. For values of the angle different from the limit values $\alpha = 0$ and $\alpha = \pi/2$ the modes of propagation are neither longitudinal nor transversal.

The phase speeds and dimensionless attenuations shown in Figures 1 and 2 are illustrated for a very large range of frequencies even if it is known that such high frequencies do not appear in geophysics. However, this is done in order to indicate the asymptotic properties of the waves. In view of a better comparability the attenuation is normalized by $\text{Im } k \to c_\infty \sqrt{2}/a_0 \text{Im } k$ where $c_\infty$ is the high frequency limit of the phase speed. In Figures 2-4 three different choices of pairs of principal tortuosities are chosen, i.e. $\tau^{(1)} = 1.5$ and $\tau^{(2)} = 6$, $\tau^{(1)} = 2$ and $\tau^{(2)} = 4.5$ as well as $\tau^{(1)} = 2.8$ and $\tau^{(2)} = 3.2$.

The wave with the highest speed and the lowest attenuation is the fast longitudinal wave pseudo $P_1$. The medium speed and attenuation belong to the pseudo $S_2$ wave. The smallest

![Figure 2](image_url): Speeds (left) and dimensionless attenuations (right) of the three pseudo waves appearing in Alermoeh Sandstein for three pairs of principal tortuosities and two values of angle $\alpha$. 
speed and the highest attenuation are those of the slow longitudinal wave pseudo $P_2$. It is obvious that for both waves which exhibit high- and low-frequency limits different from zero, the pseudo $P_1$ and the pseudo $S_2$-wave, only at very high frequencies the angle $\alpha$ plays any role for the wave speed. The attenuations exhibit a strong dependence on the angle $\alpha$ predominantly for high values of the frequency. As for the classical $P_1$- and $S$-waves of isotropic porous media the attenuation is rather small for the pseudo $P_1$- and pseudo $S_2$-wave. The maximum value for pseudo $P_1$ is 1.25, for pseudo $S_2$ it is around 4.5. The $P_2$-wave in isotropic porous media is strongly damped. This is also the case for the pseudo $P_2$-wave. The corresponding maximal dimensionless attenuation is 28. The values of the dimensionless attenuations indicate a strong influence of the difference in the principal tortuosities on the attenuation. For the three pseudo waves the maximal value of the dimensionless attenuation is for a big difference in the principal tortuosities around the double of this of a medium difference and for a small difference a half.

These results indicate that in the low frequency range – in contrast to the high frequency range – the orientation of the principal directions of tortuosity plays a secondary role in measured speeds and attenuation. Thus, most likely it cannot be used for geotechnical applications.

3.3 Shear polarization

Shear waves play a particular role in the analysis of anisotropy. This is due to the fact that in isotropic materials the properties of shear waves are independent of the polarization. All waves with an amplitude perpendicular to the direction of propagation have the same speed and the same attenuation. For this reason they are called $S$-waves without indicating the direction of the amplitude on the plane perpendicular to the propagation direction. This is not the case for anisotropic media and, in particular, for porous media with an anisotropic permeability.

In order to investigate the deviation from the plane perpendicular to the propagation direction of the pseudo transversal wave $S_2$ the shear polarization factor is built. It is defined by

$$d = \left| \text{Re} \frac{V_s^2}{V_s^2} \right| = \left| \text{Re} \frac{N}{D} \right|,$$

$$N = \frac{\pi \rho_0 \omega}{\rho^5} \left( \tau^{(1)} - \tau^{(2)} \right) \sin \alpha \cos \alpha \left\{ -\omega^2 + \frac{\mu_s^2}{\rho^2} k^2 \right\},$$

$$D = \frac{\pi \rho_0 \omega}{\rho^5} \left( \tau^{(1)} \cos^2 \alpha + \tau^{(2)} \sin^2 \alpha \right) \left\{ -\omega^2 + \frac{\lambda_s^2 + 2\mu_s^2}{\rho^2} k^2 + \frac{\sigma k}{\rho^2} \right\} +$$

$$+ \frac{\sigma}{\rho^2} k^2 - r \left( -\omega^2 + \kappa k^2 \right) \frac{\sigma}{\rho^2} k^2 + r \left( -\omega^2 + \kappa k^2 \right),$$

$$r = \frac{\rho^r}{\rho^2}.$$
Details on the derivation of $N$ and $D$ can be found in [1], $k(\omega)$ is the solution of the dispersion relation corresponding to the pseudo shear wave.

In Figure 3 several dependencies of the shear polarization factor, i.e. of the deviation of the direction of the pseudo shear wave from the plane perpendicular to the propagation direction, are shown. From the left column it gets obvious, that $d$ both in the limits of the angle $\alpha$ ($\alpha = 0$ and $\alpha = \pi/2$) and of the frequency ($\omega = 0$, $\omega \to \infty$) is zero. Hence, the coincidence of the propagation direction with a principal direction of tortuosity yields pure transversal waves with an amplitude perpendicular to the propagation direction. Simultaneously, in both limits of the frequency waves do not feel the anisotropy of the tortuosity and become pure transversal waves as well.

In both the left and the middle columns of Figure 3 the dependence of the polarization factor $d$ on the frequency and on the angle $\alpha$ is shown for the above mentioned pairs of principle tortuosities – large difference at the top, medium difference in the middle and small difference at the bottom. While in the left column the original values are presented, in the

![Figure 3: Different views of the shear polarization factor $d$. Left: as function of frequency and angle $\alpha$, middle: logarithmic presentation of $d$, right: range of $d$ in a certain frequency interval in dependence on $\alpha$. Pairs of tortuosities – top row: $\tau^{(1)} = 1.5$, $\tau^{(2)} = 6.0$, middle row: $\tau^{(1)} = 2.0$, $\tau^{(2)} = 4.5$, bottom row: $\tau^{(1)} = 2.8$, $\tau^{(2)} = 3.2$](image-url)
middle column the factor $d$ is illustrated logarithmically within the limits $10^{-5} \leq d \leq 0.185$. On the right-hand side for the range of frequencies $10^0 \text{ Hz} \leq \omega \leq 3 \cdot 10^9 \text{ Hz}$ the polarization factor – also logarithmically – is shown in dependence on $\alpha$. The maximum of $d$ for different angles $\alpha$ varies between around 0.18 for a big difference of the principal tortuosities, 0.13 for a medium difference and 0.02 for a small difference.

Inspection of the left-hand side of the figure immediately points up the limit values of $d$. From middle and right columns – due to the logarithmic presentation – these limits are not evident. However, these figures show that the difference in the tortuosities is not that important for the frequency dependence than for the $\alpha$-dependence.

Additional to the presentations of $d$ for high frequencies, in Figure 4 for a value of the frequency appearing in geotechnics ($\omega = 1000 \text{ Hz}$) the dependence on $\alpha$ is presented in normal scale. In this way, again, the limits get obvious and the differences for the three choices of principal tortuosities become evident as well even if the values are very small (of the order of $10^{-11}$).

The above numerical results show, similarly to the dependencies of speeds and attenuations on the angle $\alpha$, that a particular orientation of the principal directions of the tortuosity tensor with respect to the propagation direction is of bigger influence for higher frequencies than for lower ones. The anisotropy is primarily reflected by the influence of the principal tortuosities on the attenuation of monochromatic waves and this gives rise to a possibility of a new method of nondestructive acoustic testing of the permeability of geomaterials. For such a method the polarization of the shear waves would be of high importance.

4 CONCLUSIONS

- In a two-component poroelastic material with anisotropic permeability for a special choice of orientation of the propagation direction four waves occur: a transversal mode $S1$, a pseudo transversal mode $S2$ and two pseudo longitudinal waves $P1$ and $P2$.
- In contrast to the high frequency region, in the range of low frequencies – appearing in geotechnical applications – the orientation of principal directions of tortuosity is of
marginal importance for speeds and attenuations.

Even though, the different permeabilities in different directions may be measured: Due to the appearance of two shear modes a device may be constructed which induces shear waves of different polarization and different propagation directions. A comparison of the amplitudes of arrivals for different polarization of signals would give rise to the principal values and directions of the tortuosity. This would be a nondestructive test of geophysical materials and could be used in applications as seepage processes in road and dam constructions or tunneling in rocks.

REFERENCES


COUPLED THERMO-HYDRO-MECHANICAL MODELLING OF CRACK DEVELOPMENT ALONG FOSSIL DINOSAUR’S FOOTPRINTS IN SOFT COHESIVE SEDIMENTS

LONG NGUYEN-TUAN*, HANNA VIEFHAUS*, MARIA DATCHEVA† AND TOM SCHANZ*

*Chair for Foundation Engineering, Soil and Rock Mechanics
Ruhr-Universität-Bochum, Germany
Universitätsstr. 150, IA 4, 44780 Bochum
e-mail: [long.nguyentuan; hanna.viefhaus; tom.schanz]@rub.de, web page: http://www.gbf.rub.de/

†Institute of Mechanics - Bulgarian Academy of Sciences
Acad. G. Bonchev, block 4, 1113 Sofia Bulgaria
e-mail: datcheva@imbm.bas.bg - web page: http://www.imbm.bas.bg/112

Key words: coupled Thermo-Hydro-Mechanical (THM) modelling, dinosaur’s footprints, 3D finite element analysis, crack analysis.

Abstract. The fossil footprints have been used to back calculate the properties of the soil in the Age of Dinosaurs. The interpretation of fossil footprints requires the simulation of the processes during as well as after the footprint was generated. Some radial and circumferential cracks were observed occurring on track walls of the footprints. It is supposed that the origin of these cracks can be elucidated by means of footprint’s drying after the footprint is generated.

In order to verify this hypothesis and to allow for a precise interpretation of the dinosaur tracks, a series of laboratory and numerical simulation tests was carried out. The tests were designed to mimic the shape evolution of the footprint of the dinosaur during drying. Within the experiments the change of environmental humidity and temperature was monitored and recorded. The laboratory experiment showed that both the radial and circumferential cracks appear during the drying process.

The numerical simulation has been performed to better understand and to account for the cracking mechanism in dinosaur’s tracks. In this study the behaviour of a silty soil during drying is numerically simulated by means of a 3D model and performing coupled thermo-hydro-mechanical analysis utilizing the finite element program CODE_BRIGHT. Based on the analysis of the tensile stresses along the sample, it was found that the highest tensile stress is on the track wall and it is due to soil shrinkage. It can be concluded that the high tensile stress induced during drying is the most possible reason for cracks to appear in radial and circumferential direction along the foot’s imprint.
1 INTRODUCTION

The study of fossil footprints plays an important role for investigation of the animal characteristics, behaviour and its natural environment during the Dinosaur Age. Many paleontologists contributed studies on the morphology of the fossil footprints, the motion of the vertebrates and the sediment structure based on analysis of the imprinting process [7, 8, 9].

This paper contributes to the analysis of the change of the mechanical properties of soil composing the footprint starting from the initial saturated sediment and posterior imprinting process. If the sediment is soft enough to allow the foot sinking to a certain depth, there will be formed a vertical wall due to the pail up from the true footprint to the soil surface. This vertical wall is termed a track wall [9]. Palaeontologists have discovered many track walls on which fractures were observed at radial direction and along the contour around the imprint. Figure 1 presents the footprints of dinosaurs from Colorado, USA, showing the existence of many fractures on the track wall of the footprints. A question now arises what physical conditions and mechanisms caused these fractures. It is to be mentioned that a crack may be formed at different periods – during the imprinting period (peak weight-bearing [7]) or during the period of track drying, and because of the tectonic movement and weathering. Cracks during the peak weight-bearing period were observed in several virtual experiments [7, 8], where fractures have circumferential shape surrounding the footprint. This fact was explained by the shearing in the thrust shear zone induced by the dynamic movement of the dinosaur’s foot. Another hypothesis is that the tectonic movement induced shear and tension forces in the rock mass that initiated the fractures. Theories explaining the fracture process were analysed using stress approach and energy approach [2] and modelled accounting for the pre-existing flaw [6]. However, the tectonic cracks have different aspects compared to the cracks found at the track wall of the dinosaur footprints. It is supposed that the origin of track wall fractures/cracks can be understood by means of soil drying after the footprint is generated. Therefore, in this study an experiment was designed to simulate the footprint alteration with the progress of the drying process.

The paper is focused on the application of a thermo-hydro-mechanical model to qualitatively interpret the cracking initiation phenomenon. The analysis of the tension stress at the soil surface made it possible to explain the crack initialisation. Therefore a 3-dimensional analysis based on THM numerical simulation was performed. The footprint behaviour during drying has been numerically simulated in 3D by means of coupled THM analysis available in CODE_BRIGHT program [3]. The stress-strain behaviour is reproduced by a thermo-elasto-plastic model being a modification of the Barcelona Basic Model [1] accounting soil swelling and shrinking with the change of water content and temperature [5].
2 EXPERIMENTAL MOTIVATION

A soil ring sample having approximately the same dimensions as the dinosaur’s footprint was prepared to mimic qualitatively the formation and the alteration of the footprint after the weight-bearing period. A heaved ring aimed to simulate the track wall was formed reproducing the track wall geometry. The material of the soil ring was a silty clay that was initially water saturated. The soil ring and the base are shown in Fig. 2(a). The inner diameter is 40 cm, the outer diameter is 80 cm. The height of the ring is 6 cm, the thickness of the square base is 12 cm. The environmental humidity and temperature were varied during the experiment and two stages of the experiment can be distinguished. In the first stage, the specimen was situated for 20 days in laboratory conditions at temperature 20°C and relative humidity of 53%. In the second stage, the specimen was exposed to the sun for 20 days. Under the change of environment the soil specimen was drying by soil-water evaporation.

The laboratory experiment showed that during the drying process both the radial and circumferential cracks appeared. The cracks that were formed after 40 days are presented in Fig. 2(b). It can be seen that radial cracks appeared on the ring, and the circumferential cracks appeared at the outside and inside edges of the ring. In the drying process the soil water was released by evaporation. The reduction of soil pore water induced shrinkage of the specimen. The magnitude of this shrinkage is different along the specimen, e.g. the shrinkage at the surface of the specimen and inside the specimen is not the same. This difference induces tension stress on the surface of the specimen. If the tension stress is higher than the tensile strength of the soil, cracks will occur.
3 COUPLED THERMO-HYDRO-MECHANICAL ANALYSIS

A thermo-hydro-mechanical analysis of expansive soil, which is available in the finite element code in CODE_BRIGHT [3], was used for numerical simulation of the performed experiment. The equations that govern the THM response can be categorised into four main groups, namely, balance equations, constitutive equations, equilibrium restrictions and definition constraints. The balance equations can be seen in [12, 11]. The constitutive relations used for the THM numerical simulation are presented in the following part.

3.1 Constitutive equations of the coupled thermo-elasto-plastic model

For mechanical stress-strain behaviour thermo-elasto-plastic (TEP) model is adopted. It is presented in an incremental form and the increment of the strain as a sum of the increments of the elastic, \( \varepsilon^{e} \), and plastic, \( \varepsilon^{p} \), strains:

\[
\varepsilon = \varepsilon^{e} + \varepsilon^{p}
\]

Following the two stress variable concept in unsaturated soil mechanics, the elastic part of the strain increment is taken to be the sum of the increments of suction induced \( \varepsilon^{s-e} \), net stress induced \( \varepsilon^{\sigma-e} \) and the strain increment due to temperature change \( \varepsilon^{T-e} \). The final relation for the elastic strain increment reads:

\[
\varepsilon^{e} = \varepsilon^{s-e} + \varepsilon^{\sigma-e} + \varepsilon^{T-e}
\]

The nonlinear elastic law for the volumetric strain induced by the net stress is expressed in Eq. 3.

\[
d\varepsilon^{s-e} = \frac{K_i(s)}{1 + e} \frac{dp'}{p'} \quad \text{and} \quad p' = p - \max(P_g, P_l)
\]
\( \kappa_i(s) = \begin{cases} \kappa_{io}(1 + \alpha_i s) & \text{if } 1 + \alpha_i s \geq 0.001 \\ 0.001 k_{io} & \text{if } 1 + \alpha_i s < 0.001 \end{cases} \) \hspace{1cm} (4)

where \( p \) is the mean total stress, \( p' \) is the mean net stress in unsaturated state or effective stress in saturated state, \( P_g \) and \( P_l \) are gas pressure and liquid pressure, \( e \) is the void ratio, \( \kappa_{io} \) and \( \alpha_i \) are model parameters.

Suction and temperature induce only volumetric strains with constitutive equations given as follows:

\[
d\varepsilon^s_v = \kappa_s(p', s) \frac{ds}{1 + e} s + p_{at} \quad ; \quad d\varepsilon^T_v = \alpha_o dT
\]

with

\[
\kappa_s(p', s) = \kappa_{so} \kappa_{sp} \exp (\alpha_{ss} s)
\]

and

\[
\kappa_{sp} = \begin{cases} 1 + \alpha_{sp} \ln \left( \frac{10^{-20}}{p_{ref}} \right) & \text{if } p' \leq 10^{-20} \\ 0 & \text{if } p' \geq p_{ref} \exp \left( \frac{-1}{\alpha_{sp}} \right) \\ 1 + \alpha_{sp} \ln \left( \frac{p'}{p_{ref}} \right) & \text{elsewhere} \end{cases}
\]

The parameters involved are: \( \alpha_o \) is for the elastic thermal strain; \( \kappa_{so} \) is the elastic stiffness parameter in changing of suction at zero net stress; \( p_{at} \) is the atmospheric pressure; \( \alpha_{ss} \) and \( \alpha_{sp} \) are model parameters. The elastic slope \( \kappa_i \) and \( \kappa_s \) may be considered not dependent on temperature in case of moderate temperature gradients.

The yield surface in TEP model is given in the deviatoric plane \( p - q \) via the following equation:

\[
F = q^2 - M^2 (p' + p_s)(p_o - p') = 0
\]

where \( q = \sqrt{\frac{3}{2}} \sigma^D : \sigma^D \), with deviatoric stress defined as \( \sigma^D = \sigma' - \frac{1}{3} \sigma' : I \). The pre-consolidation pressure \( p_o \) depends on suction and according to [1] it is defined as:

\[
p_o = p^c \left( \frac{p^o}{p^c} \right)^{\lambda(0) - \kappa_{io}} \lambda(s) - \kappa_{io}
\]

where \( p^c \) is a reference pressure, \( p^o \) is the pre-consolidation pressure at the saturated state, \( \lambda(0) \) is the plastic stiffness parameters for the change in effective stress at saturated state.
The stiffness parameter for the change in the mean net stress at given suction is defined by:

$$\lambda(s) = \lambda(0) \left[ (1 - r) \exp(-\beta s) + r \right]$$  \hspace{1cm} (10)

where $r$ and $\beta$ are model parameters.

The tensile strength $p_s$ is proportional to suction and is a given function of temperature:

$$p_s = p_{s0} + k s \exp(-\rho \Delta T) \quad \text{and} \quad \Delta T = T - T_{ref}$$  \hspace{1cm} (11)

where $k$ is the parameter that takes into account the increase of tensile strength due to suction, $p_{s0}$ is tensile strength in saturated state, $\rho$ is a parameter that takes into account the decrease of the tensile strength due to temperature increase, $T_{ref}$ is reference temperature.

The isotropic hardening depends on the plastic volumetric strain according to:

$$\Delta p_o^* = \frac{1 + e}{\lambda(0) - \kappa_{io}} p_o^* \varepsilon_v^p$$  \hspace{1cm} (12)

### 3.2 Constitutive equations of the hydraulic model

Mass or liquid transfer is realised via advective flow and diffusion. Both components are considered in this analysis.

The **advective flow of the water phase** is described by the generalized Darcy’s law:

$$q_l = \frac{k_{vl}}{\mu_l} (\nabla P_l - \rho_l g)$$  \hspace{1cm} (13)

where $\mu_l$ is the dynamic viscosity of the pore liquid, $g$ is the gravity acceleration, $\rho_l$ is the liquid density. The tensor of intrinsic permeability $k$ is supposed to depend on porosity according to the Kozeny’s model:

$$k = k_o \frac{\phi^3}{(1 - \phi)^2} \frac{(1 - \phi_o)^2}{\phi_o^3}$$  \hspace{1cm} (14)

where $\phi$ is the porosity, $\phi_o$ is a reference porosity, $k_o$ is the intrinsic permeability for matrix with porosity $\phi_o$. The relative permeability $k_{rl}$, is derived using the Mualem-van Genuchten closed form model, [15]:

$$k_{rl} = \sqrt{S_e} \left( 1 - \left( 1 - S_c^{1/\lambda} \right)^{\lambda} \right)^2$$  \hspace{1cm} (15)

where $\lambda$ is a shape parameter for the retention curve and $S_e$ is defined as:

$$S_e = \frac{S_l - S_{rl}}{S_{ls} - S_{rl}} = \left( 1 + \left( \frac{P_g - P_l}{P} \right) \frac{1}{1 - \lambda} \right)^{-\lambda}$$  \hspace{1cm} and \hspace{1cm} $P = P_0 \left( \frac{\sigma_T}{\sigma_{T0}} \right)$  \hspace{1cm} (16)
where $S_t$, $S_{ls}$ and $S_{rl}$ are the current, the maximum and the residual liquid degree of saturation, $P_0$ is a model parameter, $\sigma_T$ is surface tension of liquid and $\sigma_{T0}$ is surface tension of liquid at which $P_0$ was measured. According to [13] $\sigma_T$ is calculated by the empirical relation:

$$\sigma_T = 0.03059 \exp \left( \frac{252.93}{273.15 + T} \right)$$ (17)

The molecular diffusion of the vapour in air is governed by the Fick’s law. Fick’s law is adopted to define the equation for the diffusive flux of the water vapour $i^v$ reads:

$$i^v = - (\phi \rho_v S_t D_m I) \nabla \omega^v$$ (18)

where $\rho_v$ is the vapour density, $\omega^v$ is the mass fraction of the vapour, $I$ is the identity matrix and $D_m$ is the diffusion coefficient of vapour in $m^2/s$ and is defined by:

$$D_m = \tau D \frac{(273.15 + T)^n}{P_g}$$ (19)

where $\tau$ is the tortuosity, $D$ is the molecular diffusion coefficient at temperature 273.15K, and $n$ is a parameter.

3.3 Constitutive equations of the thermal conductivity model

Fourier’s law is adopted for the heat conduction flux $i_c$:

$$i_c = -\lambda_T \nabla T \quad \text{where} \quad \lambda_T = \lambda_{sat} S_i^{(1-S_i)}$$ (20)

where $\lambda_T$ is the soil thermal conductivity, $\lambda_{sat}$ and $\lambda_{dry}$ are soil thermal conductivity at the saturated and dry states, respectively.

4 MODEL PARAMETERS, INITIAL AND BOUNDARY CONDITIONS

4.1 Model parameters for the THM analysis

The material used to prepare the soil sample in the experiment is a silty clay, whose the material parameters are collected from reported in the literature experimental data, where parameters for similar soils are given, [4, 1]. Parameters for the coupled thermo-elasto-plastic (TEP) model for typical silty soil are presented in Table 1 and Table 2. The tension stress may induce the tension failure that conducts the FE program to non-converged situation. Therefore the parameters regarding to tensile strength are given higher than the real soil properties. The parameters values are as follows: $\rho = 0.2 ~ (^\circ C^{-1})$, $k = 0.01 ~ (-)$, $p_{s0} = 0.5 ~ (MPa)$. 

73
Table 1: TEP elastic parameters

<table>
<thead>
<tr>
<th>$\kappa_{io}$</th>
<th>$\kappa_{so}$</th>
<th>$K_{\text{min}}$</th>
<th>$\nu$</th>
<th>$\alpha_{ss}$</th>
<th>$\alpha_i$</th>
<th>$\alpha_{sp}$</th>
<th>$p^{ref}$</th>
<th>$\alpha_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>(-)</td>
<td>(MPa)</td>
<td>(-)</td>
<td>(-)</td>
<td>(-)</td>
<td>(-)</td>
<td>(MPa)</td>
<td>($^\circ$C$^{-1}$)</td>
</tr>
<tr>
<td>0.01</td>
<td>0.0581</td>
<td>10</td>
<td>0.27</td>
<td>0</td>
<td>-0.01</td>
<td>0.3</td>
<td>0.126</td>
<td>1.0E-05</td>
</tr>
</tbody>
</table>

Table 2: TEP plastic parameters

<table>
<thead>
<tr>
<th>$\lambda(0)$</th>
<th>$r$</th>
<th>$\beta$</th>
<th>$\rho$</th>
<th>$k_o$</th>
<th>$p^c$</th>
<th>$M$</th>
<th>$\alpha$</th>
<th>$p_0^*$</th>
<th>$\varepsilon_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>(-)</td>
<td>(-)</td>
<td>(MPa$^{-1}$)</td>
<td>(-)</td>
<td>(MPa)</td>
<td>(-)</td>
<td>(MPa)</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>0.115</td>
<td>0.98</td>
<td>0.01</td>
<td>0.2</td>
<td>7.32E-03</td>
<td>0.01</td>
<td>1.12</td>
<td>0.426</td>
<td>0.28</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The basic parameter for temperature evolution and distribution is the thermal conductivity $\lambda$, which depends on liquid saturation. Following [14] and [10], the parameters are used as $\lambda_{\text{dry}} = 0.7$ and $\lambda_{\text{sat}} = 1.3$.

Intrinsic permeability is calculated according to the Kozeny’s model based on experimental result from saturated hydraulic conductivity test in THM column at $20^\circ$C. Parameters for this model are shown in Table 3.

4.2 Initial and boundary conditions

From the performed in this study experiment it can be seen that cracks occur on the surface of the soil sample. These cracks are radial and circumferential cracks and therefore, in order to assess the expectation of the cracks to appear on the surface of the soil sample during drying, the finite element analysis has to be performed in its 3D formulation. With the 3D model the stress-strain behaviour in condition of tension stress on the surface of the sample can be clearly seen. Because of the existing symmetry in the test sample and for reducing the number of element, thus reducing the cost of the numerical analysis, the geometry of the numerical model was created as one fourth of the whole real soil samples as depicted in Fig. 3. The observation points are selected on the top of the ring (point 2) and on the inside and outside edges of the ring (points 1 and 3). The observation points correspond with the place where cracks occurred in the experiment. The finite element discretization is also illustrated in Fig. 3.

Table 3: Hydraulic parameters

<table>
<thead>
<tr>
<th>$P_0$ (MPa)</th>
<th>$\lambda$</th>
<th>$k_o$ (m$^2$)</th>
<th>$\sigma_{T_0}$ (N/m)</th>
<th>$\phi_o$</th>
<th>$S_{rl}$</th>
<th>$S_{ls}$</th>
<th>$D$ (m$^2$s$^{-1}$K$^{-n}$Pa)</th>
<th>$\tau$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.61</td>
<td>0.446</td>
<td>8.78E-16</td>
<td>0.072</td>
<td>0.481</td>
<td>0.11</td>
<td>1.0</td>
<td>5.6E-7</td>
<td>0.8</td>
<td>2.3</td>
</tr>
</tbody>
</table>
Table 4: Initial condition for THM test

<table>
<thead>
<tr>
<th>Porosity Unit</th>
<th>Suction (MPa)</th>
<th>Temperature (°C)</th>
<th>Stress (MPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.48</td>
<td>0.0</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 3: Discretization: (a) View from top. (b) finite element mesh

The simulation is divided into two stages corresponding to the change in the sample environment during the experiment. For the first stage, the applied suction at the upper face is increased gradually from initial 15 MPa to 30 MPa corresponding to 75% ambient room relative humidity. The temperature on the top surface of the sample is kept constant \( (T=20^\circ C) \) and equal to the initial temperature of the sample. For the second stage, the applied suction on the upper face is increased gradually from 30 MPa to 45 MPa corresponding to 62% ambient relative humidity after exposing the soil sample to the sun. The temperature on the upper face of the soil sample is increased from 20°C to 37°C in this stage. Fig 4 illustrates the thermal and hydraulic (suction) boundary conditions applied during the simulation. The initial condition is described in Table 4.

The considered time intervals corresponding to the two stages of the experiment are: first stage of 500 hours, corresponding to 20 days keeping the sample in room condition; second stage of 500 hours, corresponding to 20 days keeping the sample outdoor.

5 RESULTS

Figure 5 presents the visualization of degree of saturation after 1000 hours. It is clear that the water saturation on the surface is much smaller than water saturation inside the sample. The difference is even higher at the beginning of the simulation. The reduction of the water saturation by evaporation on the upper surface induces the
shrinkage phenomenon in this area.

Employing the TEP model, it is possible to simulate the soil shrinkage induced by decreasing suction. The relation between suction and strain is given by equation 5. At the beginning of the simulation, the sample is fully water saturated and no initial mechanical stress is applied. It is coming from the solution that after several days the water saturation at the sample surface is becoming lower than the saturation inside the sample body. Therefore, the shrinkage strain near the sample surface is higher than inside the sample body. This results in a tension stress on the surface of the sample. The obtained tension stress is shown in Fig. 6. It can be seen that the tension stress is more pronounced at the inside and outside edges of the soil ring (track wall). The stress contour analysis reveals that there are lines from the inner edge to the outer edge of the track wall along which the tension stress is higher than at the neighbourhood. This indicates that most probably radial crack will appear along the black lines shown Fig. 6. Thus, the results from the numerical analysis agree with the observation in the experiment.

For analysing the difference between the stress values on the sample surface and the stress inside the sample, we have selected points 1', 2' and 3'. These points are locate at the same place but 2 cm below the respective points 1,2,3. Figure 7a presents the tension stress at the points 1, 2, and 3 with time evolution and corresponding points 1', 2', and 3'. It can be seen that the water saturation decreases rapidly inducing a rapid increase of the surface tension stress. The tension stress at the sample surface later decreases because of the reduction of the difference between the water saturation at the surface and in the sample body, see Fig. 7b. The same process is repeated in the very beginning of the second stage, when the applied suction and temperature increase to 45 MPa and 37°C,

Figure 4: Suction and temperature boundary condition on the upper face of the soil sample
respectively. The net tensile mean stress increases for a short time at the beginning of the second stage, then slowly decreases.

Figure 7b presents the evolution of the water saturation during the test at point 2 and point 2’, and at point 2” located at depth of 7 cm below the point 2. The degree of saturation on the surface (point 2) decreases rapidly to 0.2, whereas at the points 2’ and 2” in the body of the soil sample the degree of saturation decreases slowly. In point 2 at the beginning of the stage 2 there is a small decrease of water saturation, but the change of the water saturation in points 2’ and 2” is undistinguished. The variation of the degree of saturation suggests an increase of the tension stress on the surface of the sample. The degree of saturation of the three observation points is expected to be the same at infinite time.

6 CONCLUSIONS

An experimental setup and numerical simulation using coupled THM analysis are presented and discussed. The experiment and the numerical modelling of the experimental
setup are done in order to analyse the initiation of cracks on the surface of dinosaur’s footprints. The stress analysis shows that the potential for crack appearance exists on inside and outside edges of the soil ring (track wall), and along the radial direction of the ring. The results show an agreement between observed cracks and simulated tensile stress concentrators. It can be concluded that the high tensile stress on the footprint’s surface that appears due to soil shrinkage phenomenon is most probably the reason for the appearance of cracks in radial and circumferential directions.

References


ON THE NECESSITY AND A GENERALIZED CONCEPTUAL MODEL FOR THE CONSIDERATION OF LARGE STRAINS IN ROCK MECHANICS

Uwe-Jens Görke*, Thomas Nagel* and Olaf Kolditz*†

* Department of Environmental Informatics
Helmholtz Centre for Environmental Research - UFZ
Permoserstr. 15, D-04318 Leipzig, Germany
e-mail: uwe-jens.goerke@ufz.de - Web page: http://www.ufz.de

† University of Technology Dresden, Germany

Key words: Porous Media Theory, Coupled Problems, Large Deformations, Finite Element Method, Rock Mechanics, Thermodynamical Consistency

Abstract. This contribution presents a generalized conceptual model for the finite element solution of quasi-static isothermal hydro-mechanical processes in (fractured) porous media at large strains. A frequently used averaging procedure, known as Theory of Porous Media, serves as background for the complex multifield approach presented here. Within this context, a consistent representation of the weak formulation of the governing equations (i.e., overall balance equations for mass and momentum) in the reference configuration of the solid skeleton is preferred. The time discretization and the linearization are performed for the individual variables and nonlinear functions representing the integrands of the weak formulation instead of applying these conceptual steps to the overall nonlinear system of weighted residuals. Constitutive equations for the solid phase deformation are based on the multiplicative split of the deformation gradient allowing the adaptation of existing approaches for technical materials and biological tissues to rock materials in order to describe various inelastic effects, growth and remodeling in a thermodynamically consistent manner. The presented models will be a feature of the next version of the scientific open-source finite element code OpenGeoSys developed by an international developer and user group, and coordinated by the authors.

1 INTRODUCTION

Certain rock materials, like rock salt and claystone, play an important role in geotechnical applications (e.g., energy storage), and are characterized by a highly complex material behavior. Irreversible deformations, rate dependent stress-strain effects as well as creep, swelling and shrinking are observed for such materials under realistic load regimes.
Constitutive relations for the mathematical modeling of the mechanical deformation behavior of rock materials or their substitutes (e.g., bentonite) have been developed for decades. Usually, they are phenomenologically, i.e., macroscopically based. Nevertheless, corresponding constitutive models that are known from literature have certain conceptual limitations. Rarely, they consider couplings to thermal and/or hydraulic processes. Additionally, in general such models are formulated as scalar equations to consider specific deformation processes, for instance uniaxial or principal axes states, which makes their transformation to the tensorial description of generalized, three-dimensional stress-strain states difficult. Usually, specific functions for the modeling of particular constitutive effects are defined in a more heuristic manner based on the results of simplified lab experiments. Thus, the compliance with the requirements of the axioms of rational thermodynamics (e.g., thermodynamic consistency, i.e., the a priori compliance with the fundamental theorems of thermodynamics for arbitrary stress-strain states) that are well-approved in the continuum mechanics of materials is not verified. In addition, only a few authors discuss large strain models for rock materials. However, in particular within the context of the above mentioned swelling effects in reality strains of more than 40% are measured for claystone [1], which cannot be numerically simulated in a physically useful manner based on small strain approaches. Consequently, the aforementioned limitations in constitutive modeling of rock materials result in uncertainties of the transformation of most of the existing models from simplified to more complex stress-strain states.

The theoretical framework of large strain mechanics is very well established in the literature (cf. [2, 3] and references given therein). Starting in the 50s of the previous century the development of large strain models for technical materials such as elastomers improved the accuracy of numerical analyses of the mechanical behavior of corresponding components substantially [4]. Based on experimentally observed similarities in the stress response of mechanically loaded biological soft tissues and certain technical materials, in the 80s and 90s large strain models have increasingly been discussed in biomechanics and subsequently extended to biphasic materials such as articular cartilage (cf. [5] and other authors). Currently, large strain models are state-of-the-art in biomechanics. For more realistic results in the numerical prediction of consolidation processes, large strain models have been introduced in soil mechanics starting from the late 90s of the previous century [6]. Currently, large strain models are discussed in a huge amount of publications addressing deformation processes as well as their couplings with other physical effects (e.g., thermal, hydraulic) for technical materials and biological tissues.

Although various experiments for rock materials show comparable large strain effects, hardly any corresponding model is known from the rock mechanics literature. However, well-discussed phenomena such as creep of rock salt and swelling of claystone indicate the necessity of considering large strain approaches in order to avoid physically inappropriate results of numerical simulations. In this study, we present a generalized numerical model for the finite element solution of quasi-static coupled processes in (fractured) porous media at large strains. Without loss of generality, for simplicity of the representation the
presented model is restricted to isothermal hydro-mechanical (HM) processes in fully saturated biphasic materials neglecting mass production of the constituents as well as mass transfer between them. The consideration of non-isothermal, multiphase-multicomponent and/or partly saturated effects follows straightforwardly from the presented procedure.

In the following, tensors will be denoted by bold-faced characters in direct notation. Their juxtaposition implies the scalar product of two vectors (e.g., \( \mathbf{a} \mathbf{b} = a_i b^i \)), or a single contraction of adjacent indices of two tensors, while double dots indicate a double contraction of adjacent indices of tensors of rank two and higher (e.g., \( \mathbf{a} \cdot \cdot \mathbf{b} = a_{ij} b^i \)). A superposed dot indicates the material rate of a tensor, a superscript \( .^T \) the transposed tensor. Tensors belonging to the reference configuration of the solid skeleton are denoted by capital letters (additionally labeled by the subscript \( ._S \)), tensors in the current configuration by small letters. The subscripts \( ._S \) and \( ._F \) indicate variables corresponding to the solid skeleton and the pore fluid, respectively.

2 CONCEPTUAL MODELING

The model, which is discussed in this paper, is mainly based on the so-called Theory of Porous Media (TPM). In brief, the TPM is a combination of the physically based mixture theory (see [7]) with the concept of volume fractions (cf. [8–10] and others). Within the context of this enhanced approach of the mixture theory all kinematical and physical quantities can be interpreted on the macro scale as local statistical averages of the corresponding values of the underlying microstructure.

A comprehensive overview of the history and the current state of the TPM is given, for instance, by [11]. The development of material-independent basic principles (kinematics of transport and deformation, balance relations) to model the behavior of fully and partially saturated porous continua within the context of the TPM, and the formulation of appropriate numerical schemes based on standard Galerkin procedures are discussed in detail by [11–13] and [14] (see also the huge number of references therein).

For the first time, [15] presented a model for large elastoplastic solid skeleton deformations within the context of hydro-mechanical porous media behavior based on the multiplicative split of the deformation gradient. More recently mixed large strain formulations for porous media mechanics are discussed by [6] (elastic solid skeleton), [16] and [17] (elastoplasticity in case of partially saturated models), [18] (dynamic hyperelastic model) and [19] (dynamic elastoplastic approach). While these papers are mainly dedicated to applications in soil mechanics, [5, 20–22] and many other authors present various large strain porous media models adopted to biomechanical problems.

The conceptual basics, numerical aspects and examples of application of the TPM under large strain conditions have been studied by many authors (only a very short overview could be given here), but few of the previous works analyzed a consistent representation of the weak formulation of the governing equations in the reference configuration of the solid skeleton. This description is preferred here, and serves as the foundation of a generalized material approach, the details of which are discussed in [23].
2.1 Preliminary Remarks

Within the context of the TPM, all constituents of the porous medium are understood as smeared substitute continua with reduced mass density. Consequently, the porous medium is considered as a substitute continuum model, which is constituted by overlapping homogenized partial continua, and which is able to characterize physical processes in heterogeneously structured materials using the well-known assumptions and thermodynamically based approaches of continuum mechanics.

As usual in the context of the concept of volume fractions, the pore structure as well as the pore distribution are described in a statistically averaged sense using scalar variables representing the fraction of the partial volume of the constituent with respect to the overall volume $d\Omega_0$ in the reference state of a representative elementary volume of the control space. In case of biphasic porous media the volume fractions $\phi_{S_0}$ for the solid skeleton and $\phi_{F_0}$ for the pore fluid (i.e., the porosity) at time $t = t_0$ are defined as follows:

$$\phi_{S_0} = \frac{d\Omega_{S_0}}{d\Omega_0}, \quad \phi_{F_0} = \frac{d\Omega_{F_0}}{d\Omega_0}$$

with the partial volume $d\Omega_{S_0}$ of the solid skeleton and the partial volume $d\Omega_{F_0}$ of the pore fluid. The saturation condition

$$d\Omega_0 = d\Omega_{S_0} + d\Omega_{F_0} \Rightarrow \phi_{S_0} + \phi_{F_0} = 1$$

which is assumed to be fulfilled at each time $t$, represents a constraint condition.

Considering porous media constituents, two different definitions of their mass density are given. For the effective (aka realistic) density the differential elements of mass $dm_{S_0}$ and $dm_{F_0}$ of the constituents are related to the partial elementary volumes.

$$\varrho_{SR_0} = \frac{dm_{S_0}}{d\Omega_{S_0}}, \quad \varrho_{FR_0} = \frac{dm_{F_0}}{d\Omega_{F_0}}$$

In contrast, the partial (aka global) mass density of the constituents is related to the elementary volume of the overall continuum.

$$\varrho_{S_0} = \frac{dm_{S_0}}{d\Omega_0} = \phi_{S_0} \varrho_{SR_0}, \quad \varrho_{F_0} = \frac{dm_{F_0}}{d\Omega_0} = \phi_{F_0} \varrho_{FR_0}$$

The averaged density of the (homogenized) overall porous structure is defined as

$$\varrho_0 = \varrho_{S_0} + \varrho_{F_0}$$

2.2 Kinematics of Transport and Deformation

Below, the description of the kinematics of a multiphase medium is based on two fundamental assumptions

1. At the current time $t$, each particle located at the position $x$ of the mapping of the real body into the physical space simultaneously consists of material points of all of the partial constituents, and
2. all constituents are characterized by an individual, independent motion process (i.e., transport, deformation) of their material points.
The reference configuration of the porous body is identical to the reference configuration of the solid skeleton, and represents a set \( \Omega_0 \subset \mathbb{R}^3 \) of material points with the boundary \( \Gamma_0 \) (i.e., an area within the three-dimensional Euclidian space \( \mathbb{E}^3 \)). The material points of the solid skeleton are uniquely defined by their position vectors \( \mathbf{X}_S \in \Omega_0 \). Regarding the individual motion of the considered constituents, material points of the solid skeleton and the pore fluid, both belonging to \( \mathbf{x} \) at the current time \( t \), were located at different positions in the reference configuration. With the individual laws of motion

\[
\mathbf{x} = \varphi_S(\mathbf{X}_S, t), \quad \mathbf{x} = \varphi_F(\mathbf{X}_F, t)
\]

unique relations between the current position of material points of the constituents in \( \mathbb{E}^3 \) at any time \( t \), and their assignment to the reference state are given.

Within the context of TPM applications, usually a Lagrangian description is used for the kinematics of the solid skeleton. For physical correctness, the fluid flow as motion relative to the motion of the solid skeleton is originally referred to the current configuration, which is actually a description of Eulerian nature. In order to smooth out some shortcomings of an inconsistent formulation of the motion of individual constituents, [24–26] and others proposed a so-called generalized material description of the balance relations of the TPM considering the reference configuration of the solid skeleton as reference configuration of the overall continuum.

The displacement vector for material points of the solid skeleton is the primary kinematical variable of the TPM. Using the motion law (6.1), the displacement vector can be represented as a function of the coordinates of the reference configuration and the time.

\[
\mathbf{u}_S = \mathbf{u}_S(\mathbf{x}, t) = \mathbf{u}_S(\varphi_S(\mathbf{X}_S, t), t) = \mathbf{U}_S(\mathbf{X}_S, t) = \mathbf{x}(\mathbf{X}_S, t) - \mathbf{X}_S
\]

Kinematical reflections regarding the balance relations and constitutive models at large strains are usually based on the deformation gradient

\[
\mathbf{F}_S = (\text{Grad}_S \mathbf{x})^T = (\text{Grad}_S \mathbf{U}_S)^T + \mathbf{I}
\]

providing the mapping of material line elements of the solid skeleton from the reference into the current configurations. The determinant \( J_S \) of the deformation gradient

\[
J_S = \det \mathbf{F}_S = d\Omega / d\Omega_0
\]

represents the volume ratio of the solid skeleton smeared over the current configuration with respect to the reference configuration. Based on the deformation gradient, different strain measures can be defined. Within the context of the generalized material description the right Cauchy-Green tensor \( \mathbf{C}_S = \mathbf{F}_S \mathbf{F}_S^T \) and Green’s strain tensor \( 2\mathbf{E}_S = \mathbf{C}_S - \mathbf{I} \) are of particular interest.

2.3 Effective Stress Concept

Based on experimental observations on saturated soils [27] introduced the concept of effective stresses, in order to calculate the overall stress state in porous media. This
heuristic principle implies the decomposition of the stress state at any spatial point of
the current configuration into partial stresses. Characterizing the interaction between
the moving pore fluid and the deforming solid skeleton, one of the partial stresses acts
equally in material points of the pore fluid and the solid skeleton that occupy the same
location in the current configuration. This partial stress can be represented using a second
order isotropic tensor (hydrostatic state), whose coefficient is known as pore pressure $p$.
Consequently, the overall stress state has to be determined by additional partial stresses,
which are caused by the history of the fluid transport as well as the deformation of the
solid skeleton themselves. These partial stresses are called effective stresses.

The generalized material description of the partial stresses for the solid skeleton and the
pore fluid of a saturated biphasic porous medium in terms of 2nd Piola-Kirchhoff partial
stress tensors $T_S$ and $T_F$ follows from the corresponding spatial description performing
usual pull-back operations.

\begin{align}
T_S &= -J_S \phi_S p C_S^{-1} + T^E_S \tag{10a} \\
T_F &= -J_S \phi_F p C_S^{-1} + T^E_F \tag{10b}
\end{align}

Here, $T^E_S$ and $T^E_F$ denote the effective stress tensors for the solid skeleton and the pore
fluid, respectively. Neglecting internal friction forces of the pore fluid, and considering
the saturation condition $S_v$, the 2nd Piola-Kirchhoff overall stress tensor $T$ is defined as
follows:

\begin{equation}
T = T^E_S - p S_v = T^E_S - T_v \quad \text{with} \quad S_v \overset{\text{def}}{=} J_S C_S^{-1}
\end{equation}

### 2.4 Constitutive Models for the Solid Skeleton

The effective stress tensor $T^E_S$ is characterized by the deformation of the solid skele-
ton as well as non-mechanical processes (e.g., thermal and/or (electro-)chemical effects
in geo- and biomechanics). Preferring phenomenological, macroscopic constitutive con-
cepts, this causes certain additive decompositions of the effective stress tensor. As known
from respective approaches in solid mechanics, appropriate evolutional relations for par-
tial stresses can be thermodynamically consistently formulated considering the classical
axioms of material theory (cf. [2, 28]).

In particular, the formulation of thermodynamically consistent constitutive relations is
based on the conceptual analysis of the combination of the first and second laws of ther-
modynamics, which is frequently called the Clausius-Duhem inequality. Based on corre-
sponding definitions of the energy and entropy balances for the individual constituents, the
Clausius-Duhem inequality for the considered porous media problem can be represented
in generalized material description as follows:

\begin{equation}
\frac{1}{2} T^E_S \cdot \dot{C}_S - g_{S0} \ddot{\psi}_S \geq 0 \tag{12}
\end{equation}

Defining relation (12) certain assumptions on the relationships between partial balance
laws of the constituents and their counterparts for the overall continuum, and on the vari-
able characterizing the momentum exchange due to the interaction between the inviscid
pore fluid and the solid skeleton have been taken into account. Furthermore, the volume balance of the overall continuum, the decomposition of the partial stress tensors (10a), (10b), and the saturation condition (2) have been used (for details see [10, 23]).

Dependent on the material properties of the solid skeleton, its partial free energy \( \tilde{\psi}_S \) represents a function of an elastic strain measure and, occasionally, on different internal variables. Consequently, analyzing relation (12) a thermodynamically consistent constitutive relation for \( T^E_S \) can be defined as well as potential evolutional equations for stress-type quantities, which are work-conjugated to the internal variables. Within this context, the definition of specific constitutive relations is based on the multiplicative split of the deformation gradient into an elastic and various inelastic parts \( F^e_S = F_S^e F^i_S \), dependent on the relevant physical effects. Obviously, the following approaches that are well-approved for technical materials and/or biological tissues are of interest for applications in rock mechanics (in brackets only one early relevant reference is given, respectively, although, currently a huge number of publications exist on this topic): elastoplastic models \( (F^i_S = F^p_S; [29]) \), viscoelastic models \( (F^i_S = F^v_S; [30]) \), growth models \( (F^i_S = F^g_S; [31]) \).

3 GOVERNING EQUATIONS

In this section, the balance laws for mass and momentum related to the overall continuum are presented to provide a complete set of governing equations, which enables the solution of coupled quasi-static initial-boundary value HM problems for saturated biphasic porous media in terms of a generalized total Lagrangian finite element approach. The proposed conceptual steps can be summarized as follows:

1. Development of the weak formulation from the strong form of the problem
2. Time discretization of rate terms of the integrands of the weak formulation
3. Consistent linearization of individual nonlinear functions of the integrands of the weak formulation using Taylor series representations
4. Spatial discretization of the linearized weak formulations using standard Galerkin procedures

3.1 Governing Balance Relations of the Overall Continuum

For the important case of quasi-static loading, inertial forces are neglected, and the local balance of (linear) momentum of the overall continuum reduces to the classical form of the equilibrium conditions given in generalized material description (cf. [14] and others)

\[
\text{Div}_S (TF_S^e) + \varrho_0 B = 0
\]

where \( \varrho_0 B \) represents the barycentric overall volume force.

Starting from the material time derivative of the saturation condition (2) with respect to the motion of a material point of the solid skeleton using the relations between the effective and partial density (4) of the constituents as well as the individual mass balances neglecting mass exchange between the constituents, and assuming intrinsic incompressibility of the constituents, the volume balance of the overall biphasic saturated porous
medium is originally obtained in spatial description (cf. [14] and others). Its generalized material formulation is given as:

\[ J_S C_S^{-1} \cdot \dot{E}_S + \text{Div}_S \mathbf{W}_F = 0 \]  \hspace{1cm} (14)

Frequently, the term \( \mathbf{\tilde{W}}_F \) is called filter velocity representing the relative velocity of the pore fluid against the deforming solid skeleton. In the mathematical sense, \( \phi_{F0} \mathbf{W}_F \) can be understood as the mapping of the spatial mass flux vector per unit time and surface area into the reference configuration of the solid skeleton.

### 3.2 Weak Formulation

Multiplying Eq. (13) with an arbitrary test function \( V_S = V_S(X_S) \) \((V_S = 0\) at the Dirichlet part \( \Gamma_{0DU} \) of the boundary of the domain under consideration\) and integrating the result over the volume of the undeformed domain \( \Omega_0 \), the weighted form of the equilibrium conditions for the overall continuum, i.e., the weak formulation follows. With

\[ \text{Div}_S(T_S^T V_S) = \text{Div}_S(T_S^T V_S) - T_S^T \cdot (\text{Grad}_S V_S)^T \]  \hspace{1cm} (15)

neglecting volume forces, using Eq. (8) as well as the decomposition (11) of the overall stress tensor, based on Gauss-Ostrogradski’s integral theorem and considering the symmetry of the 2nd Piola-Kirchhoff stress tensor, the weak formulation of the equilibrium conditions of the overall continuum becomes:

\[ \int_{\Omega_0} T_S^E (U_S; V_S) d\Omega_0 - \int_{\Omega_0} p (S_v \cdot E_S (U_S; V_S)) d\Omega_0 = \int_{\Gamma_{0NU}} \mathbf{R}_U V_S d\Gamma_0 \]  \hspace{1cm} (16)

Here, \( \mathbf{R}_U \) is the given external loading related to the corresponding Neumann surface \( \Gamma_{0NU} \), and the kinematic variable

\[ 2E_S(U_S; V_S) \equiv (\text{Grad}_S V_S)^T + \text{Grad}_S U_S + \text{Grad}_S(U_S \cdot \text{Grad}_S V_S)^T + \text{Grad}_S V_S \cdot (\text{Grad}_S U_S)^T \]  \hspace{1cm} (17)

has been defined in order to simplify further representations.

Considering the stress decomposition (11), multiplying Eq. (14) with an arbitrary test function \( q = q(X_S) \) \((q = 0\) at the Dirichlet part \( \Gamma_{0DP} \) of the boundary of the domain under consideration\) and integrating the result over \( \Omega_0 \), the weighted form of the volume balance relation for the overall continuum, i.e., the weak formulation follows. Adopting Gauss-Ostrogradski’s integral theorem to the corresponding flux term in an analogous manner as it was done for the equilibrium conditions, a simplified representation can be derived

\[ \int_{\Omega_0} \left( S_v \cdot E_S \left( U_S; \dot{U}_S \right) \right) q d\Omega_0 - \int_{\Omega_0} (\text{Grad}_S q) \mathbf{W}_F d\Omega_0 = - \int_{\Gamma_{0NP}} \bar{R}_p q d\Gamma_0 \]  \hspace{1cm} (18)

where \( \bar{R}_p \) defines the prescribed fluid flux related to the part \( \Gamma_{0NP} \) of the surface affected by an external impact. In order to integrate the pore pressure as primary variable into
Eq. (18), the filter velocity is eliminated using Darcy’s (filter) law in generalized material description neglecting volume forces.

\[ \mathbf{\tilde{W}}_F = -K \text{Grad}_S p \]

### 3.3 Numerical Scheme

The coupled problem (16), (18) shows a nonlinear dependency on the primary variable \( U_S \), and linear dependency on the primary variable \( p \), on the rate variable \( \dot{U}_S \) and on the test functions. In order to eliminate rate-dependent terms of primary variables by discretizing them in time, a generalized trapezoidal single-step scheme is applied. Within this context, regarding the material time derivative of the solid skeleton displacements

\[ (U_S)_{n+1} = (U_S)_n + \left[ \alpha (\dot{U}_S)_{n+1} + (1 - \alpha) (\dot{U}_S)_n \right] \Delta t \]

follows, where the time increment \( \Delta t \) is given by \( t_{n+1} - t_n \), the subscript \((\cdot)_n\) denotes variables at time \( t_n \), which are known from the solution of the previous time step within the context of an incremental numerical scheme, and the subscript \((\cdot)_{n+1}\) denotes variables, which belong to the unknown current solution at time \( t_{n+1} \).

Linearization of the system (16), (18) is required in order to solve it numerically within the context of a Newton-Raphson scheme. It is performed based on Taylor series representations applied to functions of the primary variables \( U_S \) and \( p \) instead of adopting Gateaux derivatives of functionals. The idea is to find the solution

\[ (U_S + \Delta U_S, p + \Delta p) := (U^{i+1}_S(t + \Delta t_i), p^{i+1}_j(t + \Delta t_i)) \]

of the coupled two-field problem for the current \((i + 1)st\) Newton iteration at time \( t_{n+1} \) based on the given solution \((U_S, p) := (U^i_S(t + \Delta t_i), p^i_j(t + \Delta t_i))\) for the \(i\)th Newton iteration at time \( t_{n+1} \), where \((U^0_S(t + \Delta t_i), p^0_j(t + \Delta t_i)) = ((U_S)_n, p_n)\) serves as initial solution known from the previous time step at time \( t_n \).

Considering the backward Euler case \((\alpha = 1)\) in terms of the generalized time discretization scheme (20), based on the proposed linearization procedure, and neglecting terms that are at least quadratic with respect to the increments of the primary variables \( \Delta U_S \) and \( \Delta p \) as well as with respect to products of these increments, the solution of the mixed initial-boundary value problem for saturated biphasic media within the TPM at large strains results in the incremental-iterative solution of the linear system

\[
\begin{align*}
\int_{\Gamma_0} \mathbf{E}_S(U_S; V_S) \cdot \nabla \mathbf{T}(E_S(U_S), p) \cdot \mathbf{E}_S(U_S; U_S) d\Gamma_0 & \quad + \int_{\Omega_0} T(E_S(U_S), p) \cdot \text{Grad}_S \Delta U_S (\text{Grad}_S V_S)^T d\Omega_0 - \int_{\Omega_0} \Delta p (S_y(U_S) \cdot \mathbf{E}_S(U_S; V_S)) d\Omega_0 \\
= & \int_{\Gamma_0 \text{NU}} \mathbf{R}_U V_S d\Gamma_0 - \int_{\Omega_0} T(U_S) \cdot \mathbf{E}_S(U_S; V_S) d\Omega_0
\end{align*}
\]

(22a)
\[
\int_{\Omega_0} \left[ S_v(U_S) \cdot E_S(U_S; \Delta U_S) \right] q \, d\Omega_0 + \Delta t \int_{\Omega_0} (\text{Grad}_S q) \cdot K (\text{Grad}_S \Delta p) \, d\Omega_0 \\
= -\Delta t \int_{\Gamma_{0NP}} \bar{R}_p q \, d\Gamma_0 - \Delta t \int_{\Omega_0} (\text{Grad}_S q) \cdot K (\text{Grad}_S p) \, d\Omega_0 \\
- \int_{\Omega_0} \left[ S_v(U_S) \cdot E_S(U_S; U_S - (U_S)_n) \right] q \, d\Omega_0
\] (22b)

in terms of the increments $\Delta U_S$ and $\Delta p$ of the primary variables, valid for all functions $V_S \in (H^1_0(\Omega_0))^3$ as well as $q \in L^2(\Omega_0)$.

For the numerical solution of the system (22a), (22b) in appropriate partial spaces $\Delta U_S$, $V_S \in \mathcal{V}_h \subset \mathcal{V}$ and $\Delta p$, $q \in \mathcal{X}_h \subset \mathcal{X}$ of ansatz functions within the context of a mixed finite element formulation, the usual spatial discretization procedures are applied.

4 CONCLUSIONS

This paper has been contributed to the conceptual modeling and the numerical realization of a mixed finite element formulation for hydro-mechanical processes in saturated biphasic porous media at large strains based on the Theory of Porous Media. Within this context, the work was focused on isothermal, quasi-static models neglecting mass exchange between the phases. Defining all field variables consistently on the reference configuration of the solid skeleton, a generalized material approach has been presented resulting in the realization of a generalized total Lagrangian finite element approach.

The presented finite element procedure has been successfully applied to study several examples from biomechanics and soil mechanics illustrating the capabilities of the algorithms [23]. In comparison to small strain solutions, the presented results show the importance to consider large strain models developed in a consistent way. Consequently, the discussed numerical concept represents a kind of blueprint for the numerical modeling of porous media mechanics considering multiphase-multicomponent flow, reactive transport aspects and nonisothermal effects as typical phenomena observed for geotechnical applications in rock mechanics.

The discussed numerical scheme is prepared to be realized in an object oriented scientific software tool, OpenGeoSys, developed under the coordination of the authors, and applied to various hydrological and geotechnical applications (for details see [32]).

5 ACKNOWLEDGEMENTS

The development of the presented model was performed within the context of the package grant “Numerical simulation and analysis of coupled problems in continuum mechanics” funded by the German Research Foundation (DFG), grant Kr 1419/9.

REFERENCES


A COMPARATIVE STUDY OF COMPUTER MODELS FOR FRICTION AND THEIR INFLUENCE ON DYNAMICS OF THE HEAVY RIGID BODY ON A HORIZONTAL SURFACE

IVAN I. KOSENKO*

*Department of Technological Support of an Aircraft Reliability
Tsiolkovsky Russian State University of Technology
Orshanskaya st. 3, Moscow, 121552, Russia
e-mail: kosenko@ccas.ru, web page: http://www.mati.ru/

Key words: object-oriented modeling, multibond graph, canonical junction structure, rattleback, friction model, tippe-top

Abstract. Using an example of a heavy rigid body moving on a horizontal surface and having with it a permanent contact the process of construction and verification for spatial dynamical models of the multibody systems is analyzed. Two approaches to formal representation of the models: object-oriented, and bond graph based are applied. Energy based similarities between these approaches are analyzed.

A detailed description of the bond graph representation for the most general type of constraint is presented. It turned out the resulting total bond graph model of the multibody system dynamics always has exactly a so-called canonical junction structure. This representation has a tight correspondence with our object-oriented implementation of the mechanical constraint architecture. As an example Modelica implementation of several classes in the row for mechanical contact is investigated.

Computer implementations for three examples of the heavy rigid body dynamics are under investigation: (a) the rattleback, (b) example of A. P. Markeev, (c) the Tippe-Top. Among all of three examples each one demonstrates in its own manner a peculiar dynamical behaviour.

1 INTRODUCTION

Development of a computer model for multibody system (MBS) dynamics is a process needing in a reliable unified technology to construct the models in an efficient way. It turns out Modelica language provides a tools to resolve such a problem successively step by step using its natural approaches. One of them is connected tightly with the so-called multiport representation of the models initially based on the bond graph use. These latter based in turn on the idea of energy interaction, and substantially on energy conservation for physically interconnected subsystems of any engineering type.
Verification process is critically important for the reliable MBS dynamical model design. One has to remark that mechanical examples with friction are exclusively useful for testing process while verifying the model. In this way we apply several cases of MBS with contacts having well-known dynamical behavior to verify our library of classes for composing the MBS dynamical models.

2 BOND GRAPHS AS A BACKGROUND

Geometric formalisms to represent the multibody system (MBS) dynamics are well known [1]. These formalisms operate with known mechanical objects, twists and wrenches, having known information description concerning a causality in the MBS models. The model representation under consideration is tightly connected with the power based approach to modeling, so-called bond graphs [2].

Indeed, let the rigid body kinematics be defined by the twist \((v, \omega)\), where \(v\) is the mass center velocity, and \(\omega\) is the body angular velocity. All velocities for definiteness assumed to be defined with respect to (w. r. t.) any base body rigidly connected with an inertial frame of reference. Further let all the forces acting upon the body be reduced to the wrench \((F, M)\) with the total force \(F\) acting at the mass center and the total torque \(M\). Thus the total power of all the forces acting on the body is computed by the known formula

\[
W = (v, F) + (\omega, M)
\]

being used for representing a multibond in the bond graph simulating the MBS dynamics. We have in such the case an evident canonical duality between twists and wrenches.

In our approach which is more natural in traditional classical mechanics we assume twist for the flow and wrench for the effort variable in the multibond. In the further course we present an illustration for this approach and demonstrate its convenience to construct the mechanical constraints of contact type in a relatively simple way. One can find also a description of other cases of constraints in [3].

Doing so we can associate dynamical object of the rigid body with 1-junction. The reason is that 1-junction has zero sum of efforts for all incident multibonds. We can simulate the d’Alembert principle using this latter property of the junction in the following way: one multibond connects this junction with its only element of inertance, while other multibonds deliver forces of different nature to the junction. Similarly, one can associate 0-junction with the (dynamical) object of mechanical interaction, constraint in particular. This time incident multibonds have zero sum of velocities (flow variables). 0-junctions of this type always have three multibonds each. Then one of incident multibond has a flow variable being a subtraction of two other multibond flows. This property exactly is one of relative velocity at contact.

Finally one can represent a general bond graph structure of the constraint in any MBS in a way as it is depicted in Figure 1. Triangles in the Figure: \(A, B\) correspond to interacting bodies, \(C\) denotes constraint or contact force element.
All multibonds here consist of the twist \((v, \omega)\) signals representing the flow component, and the wrench \((F, M)\) signals as an effort. Causality of an inertance elements arranges according to the Newton–Euler system of ODEs. Left and right transformers are to shift the twist from the mass center to the contact point \(P\) (being computed inside block \(C\)) according to the known Euler formula: \((v, \omega) \mapsto (v + [\omega, r], \omega)\), where the vector \(r\) begins at the corresponding center of mass and ends at the contact point. Reciprocally the wrenches shift to the body mass center from point of the contact in a following way: \((F, M) \mapsto (F, M + [r, F])\). As one can see easily the transformers conserve the power. Central transformer is responsible for the transfer to orthonormal base at the contact point with the common normal unit vector and two others being tangent ones to both contacting bodies’ surfaces supposed regular enough.

Note that it is a usual practice to attach the inertance element to 1-junction, in particular because of its causality nature, see for example \([4, 5]\). Figure 1 can remind us in some degree a bond graph element of the lumped model for the flexible beam dynamics.

Causality for some multibonds inside the constraint object is defined individually for each particular scalar bond \([6]\) depending on the type of the constraint and is assigned finally after the whole MBS model compilation. If we will continue to build the bond graph model for the whole MBS in a proposed way then finally we can arrive exactly at the so-called canonical junction structure \([6]\) useful for the formal procedures of the bond graph optimal causality assignment. For this we have to add an intermediate 0-junctions for elements attached to 1-junction in the constraint component \(C\), see Figure 1.

Leaving some multibonds without the causality assignment and trusting this work to compiler we apply a so-called acausal modeling \([7]\). On the other hand if we will act in a manner close to the real cases of constraints with elasticity then instead of the constraint elements \(FC/EC\), we have to use an element of the compliance with the causality uniquely determined.
3 OBJECT-ORIENTED IMPLEMENTATION

When using Modelica language [8] we construct a unified computer model of the constraint, or, in a more general way, of any physical interaction between two rigid/deformable bodies we define [9, 10, 3] two classes of communication ports: (a) class of the kinematic port and (b) one of the effort port. These ones are twist and wrench connectors. It turned out the connections of such types make it possible to construct a model of the bodies interactions based on the causality physically motivated.

Briefly speaking we consider a constraint/contact between two rigid bodies as a communication network working in a following way. Namely, the constraint object imports the kinematic information, twists plus additional data, accepting it from the objects of interacting bodies and reciprocally exports wrenches (plus additional data) as a response in the opposite direction. Thus the constraint “computes” an efforts the bodies interact by, see Figure 2. And now we are going to point out the similarity between the bond graph description above and our MBS model. Indeed, the pair of twist/wrench ports plays a role of the multiport notion, and corresponding pairs of connections in Figure 2 stand for the notion of a bond with a causality identically defined.

Furthermore, in this way we can associate objects ob bodies with 1-junction each, while 0-junction is associated with the object of the constraint. A superclass of our package for objects of bodies encapsulates dynamics of rigid body and is described by means of Newton’s differential equations for the body mass center, and by Euler’s differential equations for the body rotation about it. The Euler equations are constructed using quaternion algebra [11].

If we use multiport concept for developing model of the MBS dynamics then the most complicated task is a design of the class hierarchy for the constraint/contact objects. A base constraint superclass is a root of such class hierarchy tree. According to Newton’s third law this superclass must contain equations of the form

\[ F_A + F_B = 0, \quad M_A + M_B = 0. \]  

(1)

in its behavioral section. Here arrays \( F_A, M_A \) and \( F_B, M_B \) represent constraint forces.
and torques “acting in directions” of bodies A and B respectively. Additional equations for different types of constraints are to be added to equations (1) in different classes—inheritors corresponding to these particular types of constraints. In order to derive these equations let us consider the local geometry of the problem, see Figure 3.

The base body of MBS supposed to be connected with the absolute frame $O_0x_0y_0z_0$ ($AF$) fixed in the inertial space (body $A$ in our case is fixed platform with horizontal surface), $O_{\alpha}x_{\alpha}y_{\alpha}z_{\alpha}$ is the frame $BF_{\alpha}$ fixed in the body $\alpha \in \{A, B\}$. The outer surfaces $\Sigma_{\alpha}$ are defined by the equations $f_{\alpha}(r_{\alpha}) = 0$ ($\alpha = A, B$) w. r. t. appropriate $BF_{\alpha}$ whose axes are coincident to the principal central axes of inertia. In $AF$ these the equations read $g_{\alpha}(r_0) = f_{\alpha}[T^*_\alpha(r_0 - r_{O_{\alpha}})] = 0$ ($\alpha = A, B$). Here $r_{O_{\alpha}} = O_0O_A$, $r_{O_B} = O_0O_B$, $T_A$, $T_B$ are the orthogonal matrices determining orientation of the $BF_A$ and $BF_B$ w. r. t. the $AF$. An asterisk denotes the matrix transposition. The functions $g_A(r_0)$, $g_B(r_0)$ depend upon the time indirectly through the variables $r_A$, $r_B$, $T_A$, $T_B$.

The constraint object of our model has to compute at each current instant the points $P_A \in A$ and $P_B \in B$ implementing the minimal distance between the bodies. These points depend on relative orientation of the bodies. By virtue of above assumptions such points are to be evaluated in a unique way. Denote by $r_{P_A}$, $r_{P_B}$ the radii vectors of these points w. r. t. $AF$. The simple geometric reasons imply the following system of algebraic equations

$$
\begin{align*}
\text{grad } g_A(r_{P_A}) &= \lambda \cdot \text{grad } g_B(r_{P_B}), & g_A(r_{P_A}) &= 0,
\text{grad } g_B(r_{P_B}) &= \mu \cdot \text{grad } g_B(r_{P_B}), & g_B(r_{P_B}) &= 0.
\end{align*}
$$

The system (2) consists of eight scalar equations w. r. t. eight scalar variables: $x_{P_A}$, $y_{P_A}$, $z_{P_A}$, $x_{P_B}$, $y_{P_B}$, $z_{P_B}$, $\lambda$, $\mu$, where $\lambda$, $\mu$ are auxiliary variables. The equations (2) are in
use either without or with a presence of the contact of bodies $A, B$. In the latter case the condition $\mu \leq 0$ satisfies. If we want to simulate the point contact case then the equation $\mu = 0$ is to be added to the model. According to computational experience it is more reliable and convenient to use the equations of constraints (2) in a differential form. When using this latter, i.e. differential, form of a constraint one needs to set a consistent initial values for the variables $r_{PA}$, $r_{PB}$, $\lambda$, $\mu$ at the start time instant of simulation. Further we analyze several dynamical examples of mechanical contact.

4 POINT-CONTACT CASE

Point contact is a simplest case under verification. As it was already mentioned to satisfy the point contact between the bodies we should satisfy the condition $\mu = 0$ for equations (2). For simplicity and definiteness from now on we assume satisfied the contact condition for the point-contact case and for the patch-contact one as well.

After that two possibilities remain: (a) relative sliding of bodies $A$ and $B$ outer surfaces at contact; (b) relative rolling at the contact point. Consider them successively decreasing number of degrees of freedom (DOFs) for relative motion at contact. Denote by $F_A$ the force acting on the body $A$ from the body $B$. And by $F_B$ denote the force acting on the body $B$ from one of $A$ vice-versa. Each force cited acts at the point $P_A, \alpha = A, B$. In case of contact $P_A = P_B$. In addition, let us introduce auxiliary notations $F_{An} = (F_A, n_A)$, $F_{Ar} = F_A - F_{An}n_A$, $v_r = v_{PA} - v_{PB}$, $v_{rn} = (v_r, n_A)$, $v_{r\tau} = v_r - v_{rn}n_A$ where $n_A$ is the outer normal to the body $A$ surface at point $P_A$.

In case of bodies contact the condition $\mu = 0$ from above is equivalent to the kinematic one $v_{rn} = 0$. Cases of sliding and rolling differ one from another using conditions in a tangent plane. Implementation of the Coulomb friction model is assumed for the simplicity. Then one can obtain the vector force equation in the tangent plane

$$F_{Ar} - d \cdot F_{Ar} \frac{v_{r\tau}}{|v_{r\tau}|} - \kappa n_A = 0,$$

where $d$ is the coefficient of friction. Thus these latter three equations plus equation $v_{rn} = 0$ all combine the system of four scalar equations w. r. t. four scalars $F_{Ax}$, $F_{Ay}$, $F_{Az}$, $\kappa$.

When rolling the tangent velocity has to be zero $v_{r\tau} - \kappa n_A = 0$, and with equation $v_{rn} = 0$ we have four scalar equations once more for the same four variables of the MBS model.

Model equation (3) “works” properly in case of sliding if relative velocity is not very small. However the problem of regularization for the equation of constraint (3) arises when transposing from the state of rolling to one of sliding. It was found that one can apply here the known approximation for Coulomb’s friction using regularized expression for the tangent force

$$F_{Ar} - \kappa n_A = \begin{cases} d \cdot F_{An}v_{r\tau}/|v_{r\tau}| & \text{as } |v_{r\tau}| > \delta, \\ d \cdot F_{An}v_{r\tau}/\delta & \text{as } |v_{r\tau}| \leq \delta, \end{cases}$$
where one supposes that $\delta \ll 1$.

It is known [12] that in this case the solution of the regularized problem remains close to the solution of the original one on asymptotically large time intervals. Implementation and further simulation show this closeness holds with the very high degree of accuracy. Such an approach resolves completely the problem of modeling for accurate transitions between states of sliding and rolling.

### 4.1 Example of A. P. Markeev

Consider the translational-rotational motion of the homogeneous rigid body of an ellipsoidal shape on the rough horizontal surface [13]. Friction assumed of the Coulomb type and regularized as above with the small coefficient $d = 0.01$. The body is set at initial instant on its smallest semi–diameter endpoint and spinned fast enough. For this case one can repeat easily an experiment described qualitatively by A. P. Markeev: the body in finite time “stands” on its largest semi–diameter.

In case of our example body’s semi–diameters are close enough one to another: $a_1 = 1.2$, $b_1 = 1$, $c_1 = 1.3$. Axes of the body ellipsoidal surface coincide with ones of the central principal ellipsoid. Choosing initial data as follows: $\mathbf{r}(0) = (0, 1, 0)^T$, $\mathbf{v}(0) = (0.05, 0, 0)^T$, $\mathbf{q}(0) = (1, 0, 0, 0)$, $\mathbf{\omega}(0) = (0, -10, 2)^T$, one obtains the result cited above: the ellipsoid mass center “rises” progressively from the height of minimal semi–diameter to one of maximal semi–diameter, see Figure 4.

### 4.2 Example of the rattleback

Further consider dynamics of the rattleback playing a role of body $B$ on an immovable horizontal surface [14] playing a role of body $A$. Usually the rattleback, or wobblestone, or celtic stone, is assumed being rigid body bounded by paraboloidal or ellipsoidal surface.
Central principal axes of inertia are rotated w. r. t. the body outer shape axes of symmetry. We consider the case of an ellipsoidal surface.

The model described above has been developed using Modelica package of classes for 3D dynamics of MBS. The high quality of an approximation for the rattleback motions has been verified through different simulations performed. In particular, computations corresponding to case of Kane and Levinson [14] have been performed. Results of simulations are identical in all slightest details.

5 PATCH-CONTACT CASE

The point contact model unable often to explain dynamic effects while elastic bodies contacting. If we introduce the patch as a result of elastic interaction then we have to construct proper model of compliance at contact. There is a lot of models of such nature. The Hertz contact model [15] is one of the most popular one among engineers. This model provides a reliable procedure of normal elastic force computation for interaction of two rigid bodies.

Local geometric reduction in vicinity of contact is similar to above one. The only difference is absence of the point case constraint $\mu = 0$. For the patch case the condition $\mu \leq 0$ should take place permanently. In the current case we suppose the bodies $A$ and $B$ don’t create any obstacles for their relative motion. If 3D-regions bounded by the bodies outer surfaces do not intersect then the object of a contact, generates a zero wrench in the direction of each body. Simultaneously this object has to generate the radius vectors $\mathbf{r}_{P_A}$, $\mathbf{r}_{P_B}$ of opposing with each other points $P_A$, $P_B$, see Figure 3. One can find in [16] full detailed reduction and computational implementation of the Hertz model and its volumetric modification proposed by V. G. Vil’ke.

Elliptic patch arising in the Hertz model causes new effects in tangent friction forces behavior. The simplest friction model on this way is one of Contensou – Erismann [17, 18]. One can find in [19] details of this model computer implementation. One of the model key features is application of piece-wise continuous approximation for the wrench of friction forces distributed over the elliptic contact patch. Due to simplicity this approximation is efficient and has an implementation fast enough. Besides, verification of the model using the Tippe-Top example shows its satisfactory accuracy.

Unlike to previous considerations we construct here a so-called unrestricted contact problem. This latter means that we do not apply any preliminary assumptions about contact patch and normal force, what is usual in engineering practice. Instead we compute dynamically the patch and the normal force using the Hertz model (or its volumetric modification), and simultaneously we compute the total wrench of friction forces taking into account dimensions of contacting area permanently in time.
5.1 Example of the Tippe-Top

The general contact model including the Hertz model for normal forces and the Contensou–Erismann one for tangent forces has been verified here by two stages: (a) for the case of circular contact; (b) for the case of elliptic non-circular contact. The known Tippe-Top dynamical model was investigated as an example of the first case. All the parameters and initial conditions are exactly the same as in the paper [20] whose authors got these data in turn from the work [21].

The Tippe-Top body, supposed geometrically rigid, composed by two balls one of larger radius \( R = 1.5 \cdot 10^{-2} \text{m} \), and another, smaller, one of the radius \( r = 0.5 \cdot 10^{-2} \text{m} \). The top center of mass supposed resting at initial instant of motion. Besides the top itself, more accurately its larger ball, assumed without any initial penetration with the horizontal surface. The smaller ball is located on the upper hemisphere of the larger ball, and initially the top axis of symmetry bends w. r. t. vertical by the angle \( \theta_0 = 0.1 \text{rad} \). Initial angular velocity \( \omega_0 = 180 \text{s}^{-1} \) is directed along axis of the top symmetry.

In the model under development here we consider an unrestricted contact problem that is the normal force is computed from the Hertz (or V. G. Vil’ke) model with addition of some nonlinear viscous term. Simultaneously the contact area is computed too. Then all the data obtained are used to calculate the tangent force and the turning friction torque in frame of the simplified Contensou model.

Remarkably, a computational experiment showed the top revolution from “feet”, the larger ball in contact, to “head”, the smaller ball in contact, scenario well known in mechanics. Simultaneously, using the results of the paper [22] a verification procedure has been performed. Namely, exact formulae obtained by V. F. Zhuravlev in [22] for the friction force and for the turning friction torque, case (a), were applied to the top dynamics computer model implemented on Modelica language in frame of the unrestricted, in sense mentioned above, contact model. In the same dynamical frame the simplified Contensou model, case (b), as well as a linear-fractional Pade approximation for the friction force and torque, case (c), were also implemented. The results of the inclination angle evolution showed that revolution scenarios are mutually closest in cases (a) and (b). If we use the Pade approximations of higher order [24], then the resulting accuracy is improved. Results obtained in [20] were also completely verified.

Note in addition, one can easily obtain a behavior typical to the Tippe-Top, revolution to “head”, in frame of the “regularized” Amontons–Coulomb friction. One has to understand regularization in a sense proposed in the works [12, 23] and used above in case of the point contact. We only have to “bend” graph for the friction force dependence on the relative slip velocity in vicinity of zero replacing its discontinuity by the linear function. The more flat slope of the graph the sooner one can find out the Tippe-Top revolution effect. As the simplified Contensou model shows that just this slope appears in the corresponding graph for the friction force dependence on the velocity, this time in frame of the exact Contensou–Erismann model.
5.2 Example of the ball bearing

The dynamical model of the ball bearing was considered in a way similar to the paper [19] while the verification of second stage. This time the contact area is essentially elliptic one. The main goal for the numeric simulations was to compare two approaches: (a) the standard Hertz model for the normal force plus the Contensou simplified model for the friction forces; (b) the simplified model of V. G. Vil’ke for the normal elastic force plus the Contensou simplified model for the friction forces. As it was observed in [19] for the case of the regularized Coulomb friction force here dynamical models of the cases (a) and (b) differ one from another in a slightest degree too. Simultaneously, the model (b) is faster than (a) by 20% meaning the CPU time needed.

6 CONCLUSIONS

Brief summary of some results obtained in the paper could include the following issues:

(a) unified multibond graph representation of the MBS dynamics in a sufficiently simple way with the canonical junction structure is possible.

(b) The representation depicted in Figure 1 can be used as a guideline for constructing the consistent system of DAEs in a systematic way. In other words we can say that multibond graph constructs like ones of Figure 1 are to be used as a regular basis for more informal object-oriented approach.

(c) An object-oriented representation makes it possible to develop the constraints models adopted to the specific types of the bodies interconnections in a fast and effective manner implementing the corresponding bond graph formalisms in a more natural and informal way mainly by chains of inheritance for the behavior (equations) and properties thus gradually filling the complete multibond graph description.

(d) An acausal modeling accelerates the modeling releasing a developer from the problem of causality assignment if s/he takes into account some requirements like complementarity rules.

(e) It turned out an introduction of the component of the ODEs system for the elastic bodies outer surfaces tracking for the contact problem conserves an accuracy and simultaneously improves the reliability of the models. To implement the tracking in case of the complex shape surfaces we have to rearrange only one derived class at the end of the inheritance chain to define an equations for the gradients and Hessians of the surfaces $A$ and $B$ w. r. t. $LF$'s of the bodies.

(f) The algorithm of V. G. Vilke is more reliable and suitable for wide range of the contact area eccentricities simultaneously providing an accuracy of 0.5% with respect to the Hertz-point algorithm.

(g) The Tippe-Top “on head” revolution effect is caused completely by the dry friction force “regularization” in vicinity of zero value for the velocity of relative slip. Such a regularization takes place exactly in the Contensou–Erismann model. Numeric experiments showed if the slope of friction force graph in vicinity of the zero velocity in the regularized
Coulomb model is steep enough then the Tippe-Top effect either isn’t observed at all or arising during short time after a long evolution then vanishes quickly. And only noticeable decreasing of the slope mentioned immediately causes the top revolution on the “head” with the subsequent long precession in this position.

7 ACKNOWLEDGEMENTS

This work was performed with partial support of Russian Foundation for Basic Research, projects: 11-01-00354-a, 12-01-00536-a, 12-08-00637-a.

REFERENCES


A HIGH-RESOLUTION SCALABLE MESHLESS METHOD FOR COUPLED STEADY POROELASTIC ANALYSIS, BASED ON COLLOCATION WITH RADIAL BASIS FUNCTIONS

HENRY POWER*, DAVID STEVENS* AND K. ANDREW CLIFFE†

* Faculty of Engineering, Department of Mechanical, Materials and Manufacturing Engineering
University of Nottingham, University Park, Nottingham, UK
e-mail: henry.power@nottingham.ac.uk

† School of Mathematical Sciences
University of Nottingham, University Park, Nottingham, UK

Key words: Radial Basis Functions, Meshless, Poroelasticity, Finite Collocation

Abstract. This work describes the application of a novel meshless numerical technique, based on local collocation with radial basis functions (RBFs), to the solution of steady poroelastic problems. Its formulation allows scalability to large problem sizes, in contrast to traditional full-domain RBF collocation methods which are restricted to small datasets due to issues with numerical conditioning and computational cost. The proposed method is validated using a benchmark linear elasticity numerical example and a coupled steady poroelastic deformation problem, for which analytical solutions are known. Highly accurate solutions are obtained in each case, and convergence rates in excess of sixth-order are observed.

1 INTRODUCTION

The interaction between fluid flow and porous matrix deformation is known as hydromechanical coupling, where the porous matrix can deform as a result of either changes in external loads, or internal pore pressures. Poroelasticity describes the governing equation for porous matrix deformation in terms of a non-homogeneous Navier system of equations for linear elasticity, with the non-homogeneous term given by the gradient of the fluid pressure. Similarly, the stress tensor constitutive equation is given by the linear elastic stress tensor minus the fluid pressure. The fluid mass balance equation occupying the void space must also account for the motion due to the solid deformation, resulting in a set of coupled partial differential equations. By neglecting the compressibility of the pore fluid and soil matrix, and considering steady state conditions, the doubly-coupled problem reduces to a one way coupling with a forcing term proportional to the pressure gradient in the solid deformation governing equation.

In this work we implement a novel local RBF collocation method (RBF-FC approach), recently developed in [1], to solve the coupled problem of flow and solid matrix deformation in poroelasticity. The basic RBF collocation method for the solution of PDEs was originally described by Kansa [2, 3], and has been successfully applied to a wide range of PDEs. RBF collocation methods are attractive due to their meshless formulation, ease of implementation, high convergence rates, and flexibility to enforce arbitrary boundary conditions. However the use of globally supported basis functions leads to fully-populated collocation matrices, which
become increasingly ill-conditioned and computationally expensive with increasing dataset size. These limitations motivate the development of alternate methods to mitigate the computational cost and numerical conditioning issues, while maintaining the performance and flexibility of the full-domain formulation.

One of the most efficient ways to overcome the ill-conditioning problem is the use of a generalised finite difference method based on RBF collocation (RBF-FD method), where a small interpolation stencil is constructed around each data point within the domain, which connects the point to its neighbours. In this way a large number of highly overlapping local collocation systems are formed. In analogy to traditional finite difference methods, the local RBF systems collocate the unknown solution value at each internal node included within the system, with the governing PDE enforced by reconstruction to form a sparse global linear system that expresses the governing PDE operator in terms of surrounding nodal values. Unlike traditional polynomial-based FD methods, RBF-FD methods can operate effectively on irregular datasets. For more information on RBF-FD methods see, for example, [4, 5, 6, 7].

In [1] an alternative RBF localisation is proposed for the meshless solution of PDEs, which is referred to as the RBF finite collocation approach (RBF-FC). As with the RBF-FD approach, small overlapping collocation stencils are formed around each internal node. However, the proposed method does not use the finite difference principle to drive the solution of the PDE; the PDE and boundary operators are enforced only within the local collocation systems, and are not used to generate the sparse global system. The global system is instead obtained by reconstructing the (Dirichlet) value of the field-variable, thereby describing the solution at the system centrepoint in terms of the unknown solution values at surrounding nodes. In this way the enforcement of the governing PDE and boundary operators is handled entirely within the local systems; the solution is driven by collocation and not by finite difference reconstruction.

2 LINEAR POROELASTICITY

The theory of poroelasticity was essentially developed by Maurice A. Biot. In [8], he couples Navier’s linear elasticity equations with Darcy’s law for the flow through a porous medium. In the present work, steady state solutions of the Biot equations are considered for an isotropic material and incompressible fluid. Under these conditions, the equations modelling the displacement $u_i$ of the material and the pressure $p$ of the fluid can be written:

$$
\frac{\partial}{\partial x_i} \left( \frac{k(x)}{\mu} \frac{\partial (p - \rho g x_2)}{\partial x_j} \right) = 0
$$

where the pressure distribution, $p$, is defined by the steady Darcy flow, with an isotropic permeability $k(x)$, fluid viscosity $\mu$, gravity acceleration $g$ and where $x_2$ is the vertical coordinate, taken positive upwards.

The poroelastic stress-tensor is given by

$$
\sigma_{ij} = \lambda \varepsilon_{ik} \delta_{kj} + 2\mu \varepsilon_{ij} - \alpha p \delta_{ij}
$$

(2)
where $\alpha \in [0, 1]$ is the Biot coefficient, $\varepsilon_{ij}$ is the infinitesimal strain tensor

$$
\varepsilon_{ij} = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)
$$

and $\lambda$, $\mu$, are the Lame constants which, for plane-strain, are related to the Young’s modulus ($E$) and Poisson ratio ($\nu$) as:

$$
\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)}.
$$

The stress field satisfies the momentum equation

$$
\frac{\partial \sigma_{ij}}{\partial x_j} = 0
$$

Writing the stress tensor in terms of the effective stress as

$$
\sigma_{ij} = \bar{\sigma}_{ij} - \alpha p \delta_{ij}
$$

and substituting into equation (5), the following non-homogeneous equation is found:

$$
\frac{\partial \bar{\sigma}_{ij}}{\partial x_j} = \alpha \frac{\partial p}{\partial x_i}.
$$

In the above equation, the pressure field and its gradient are computed from the solution of the Darcy flow equation (1), and are coupled with the elasticity equation.

### 3 A NOVEL MESHLESS LOCAL COLLOCATION STRATEGY FOR HIGH CONVERGENCE SOLUTIONS OF PORO-ELASTICITY PROBLEMS

To facilitate the locally-driven solution of the PDE, the RBF-FC local collocation systems are somewhat different from those used in RBF-FD methods. Rather than collocating the solution value at all nodes within the local stencil, the solution value is collocated only over the periphery of the stencil.

![Collocation Stencils](image)

Figure 1: Comparison of collocation stencils for RBF finite difference, and the finite collocation approach.
- Black marks represent collocation of the unknown solution value (solution centres)
- Red crosses represent collocation of the PDE governing operator (PDE centres)

In this way the local stencils communicate only via their boundaries, and the resulting finite collocation method is more strongly related to traditional RBF domain decomposition.
overlapping approaches; e.g. [9]. Within the interior of the stencils the PDE governing operator is collocated. In the case that the stencil lies sufficiently close to a domain boundary, collocation of the boundary operator replaces collocation of the solution field along the appropriate stencil boundary. The difference between the stencils used in RBF-FD methods and those of the proposed finite collocation method is outlined in Figure 1.

3.1 Kansa Collocation for linear scalar problems

A radial basis function depends upon the separation distances of a set of functional centres, exhibiting spherical symmetry around these centres. There are several commonly used radial basis functions (see [10]), however we will use the multiquadric RBF throughout, with $m=1$.

$$\psi(r) = (r^2 + c^2)^m, \ m \in \mathbb{R}^+$$ (8)

The $c$ term is known as the ‘shape parameter’, and describes the relative width of the RBF functions about their centres. In practice, tuning of this parameter can dramatically affect the quality of the solution obtained, and much research has been directed towards finding effective methods of optimisation (see for example [11,12]). Since $c$ is a length scale it is appropriate to consider a non-dimensional alternative, $c^* = c/\Delta$, where $\Delta$ is typically related to the node separation distance in some way.

The Kansa RBF collocation method [2, 3], constructs the continuous solution $u(x)$ of the PDE from a distinct set of $N$ quasi-randomly distributed functional centres $\xi_d$:

$$u(x) = \sum_{j=1}^{N} \alpha_j \psi(\|x - \xi_j\|) + \sum_{j=1}^{NP} \alpha_{j,N} P_{m-1}(x), \ x \in \mathbb{R}^n$$ (9)

Here $P_{m-1}$ is the $j^{th}$ term of an order $(m-1)$ polynomial, under the constraint

$$\sum_{j=1}^{N} \alpha_j \psi^k(x_j) = 0, \ k = 1, ..., NP$$ (10)

Consider a typical linear boundary value problem

$$L[u] = f(x), \ \text{on } \Omega$$
$$B[u] = g(x), \ \text{on } \partial\Omega$$ (11)

where the operators $L[ ]$ and $B[ ]$ are linear partial differential operators on the domain $\Omega$ and on the boundary $\Gamma$, describing the governing equation and boundary conditions respectively. Collocating the interpolation formula (9) at $N$ distinct locations, $x_j$, coinciding with the functional centres $\xi_d$, leads to a system of equations

$$\begin{bmatrix}
L[\psi] & L[P_{m-1}] \\
B[\psi] & B[P_{m-1}] \\
P_{m-1} & 0
\end{bmatrix} \alpha = \begin{bmatrix}
f \\
g \\
0
\end{bmatrix}$$ (12)

which is fully populated and non-symmetric. This approach, known as Kansa’s method, has been applied to a wide range of problems with great success (see for example [13, 14, 15]).
3.2 Kansa collocation for non-homogeneous linear elasticity

For coupled multivariate PDEs the basic RBF collocation formulae (9), (10) must be modified slightly. The following demonstrates the application to linear elasticity, however the same principles may be used to construct a global collocation for any linear multivariate PDE.

The non-homogeneous Lamé-Navier equations for linear elastic deformation, equation (7), in terms of displacement, $u_i$, are given by

$$\left(\lambda + \mu\right)\frac{\partial^2 u_j}{\partial x_i \partial x_j} + \mu \frac{\partial^2 u_i}{\partial x_j^2} = -\rho b_i$$

(13)

for a given body force $b_i$ (in the case of poroelasticity $\rho b_i = \alpha \frac{\partial p}{\partial x_i}$) and material density $\rho$. Here $\lambda$ and $\mu$ are the Lamé coefficients as defined in (4). For a 2D plane-stress approximation, the value of $\lambda$ should instead be changed to:

$$\lambda' = \frac{E'v}{(1-v^2)}$$

(14)

Boundary conditions are applied either as a fixed displacement (Dirichlet condition), i.e.,

$$u_i = f_i$$

(15)

or as a prescribed surface traction (Neumann condition). The surface-traction operator at a surface with unit outward normal $n_j$ and an applied traction $\tau_i$, is given by:

$$\lambda n_i \frac{\partial u_i}{\partial x_k} + \mu n_j \left( \frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} \right) = \tau_i$$

(16)

The solution construction is similar to that of the single-variable formulation:

$$u_i(x) = \sum_{j=1}^{N} \alpha_{i,j} \phi_j^p(x) + \sum_{j=1}^{N} \alpha_{i,j} N_j^p(x)$$

(17)

Each variable $u_i$ is constructed using a common set of RBF functional centres $\xi_j$, and the associated RBF weighting function $\alpha_{i,j}$. The collocation system is generated by enforcing the governing equation (13) at each internal node, and enforcing the displacement (15) and surface traction (16) operators at Dirichlet and Neumann centres respectively. The resulting collocation system may be expressed as follows (excluding the polynomial terms for brevity):
The operators $L_{ij}$ and $B_{ij}$ represent the differential operators applied to $u_i$ within the $j^{th}$ equation of the PDE governing operator $L$ and the surface traction operator $B$ respectively. The functions $L_{ij}$ and $B_{ij}$ may therefore be expressed as:

$$
L_{ij} = \mu \delta_{ij} \frac{\partial^2}{\partial x_i^2} + (\lambda + \mu) \frac{\partial^3}{\partial x_i \partial x_j^2} \quad i,j = 1,2,3
$$

and

$$
B_{ij} = \lambda n_i \frac{\partial}{\partial x_j} + \mu n_i \frac{\partial}{\partial x_i} + \mu \delta_{ij} n_i \frac{\partial}{\partial x_i} \quad i,j = 1,2,3
$$

An equivalent operator for a Dirichlet boundary condition would be expressed as $B_{ij} = \delta_{ij}$.

The collocation matrix (18) represents a square and near-fully-populated linear system. When polynomial terms are included it is of size $N \times N = n(N+NP)$, where $n$ is the number of spatial dimensions, $N$ is the number of collocation points, and NP is the number of terms in the polynomial. To extend the collocation matrix to include the polynomial terms it is necessary to include them in the substitution of equation (17) into the governing and boundary operators (13), (15) and (16).

### 3.1 RBF Finite Collocation Formulation

By generating an RBF collocation around each of the $N_s$ stencils, formed around strictly interior nodes as indicated by Figure 1c, a series of $N_s$ RBF collocation systems are obtained:

$$
A^{(k)} \alpha^{(k)} = d^{(k)} \quad k = 1, \ldots, N_s
$$

Here $A^{(k)}$ represents the collocation matrix for system $k$, and will be formed as described by equation (12) or equation (18). The data vector $d^{(k)}$ contains the known boundary and PDE operator values, and also the unknown value of the solution field at the solution centres. Using the appropriate RBF reconstruction formula, i.e. equation (9) or equation (17), the approximate value of the field variable $u(x)$ may be computed for any $x$ within the support domain of system $k$. Expressing this computation as a vector product we have

$$
u^{(k)}(x) = H^{(k)}(x) \alpha^{(k)}$$

where $H^{(k)}(x)$ is identified as a reconstruction vector for system $k$ at location $x$. By reconstructing the value of $u$ at the system centrepoint, $x^{(k)}_c$; i.e. the node around which system $k$ formed (though not necessarily the geometric centre of the stencil), we obtain:
\[ u^{(k)}(x_{c}^{(k)}) = H^{(k)}(x_{c}^{(k)}) \alpha^{(k)} \]
\[ = H^{(k)}(x_{c}^{(k)}) [A^{(k)}]^{-1} d^{(k)} \]
\[ = W^{(k)}(x_{c}^{(k)}) d^{(k)} \] (23)

Here \( W^{(k)}(x_{c}) = H^{(k)}(x_{c}) [A^{(k)}]^{-1} \) is a stencil weights vector, expressing the value of the solution field \( u \) at the system centrepoint in terms of the entries in the data vector \( d^{(k)} \). This value of \( u \) at \( x_{c}^{(k)} \), as reconstructed by equation (23), appears as an unknown within the data vector of any systems which have within their stencil a solution centre located at \( x_{c}^{(k)} \); i.e. any systems which have this node on their periphery. Therefore, by performing the above reconstruction (23) for each local system \( k \), a series of \( N \) simultaneous equations are formed for the \( N \) unknown values of \( u^{(k)} \) at the system centrepoints. Solution of this sparse global system therefore provides the value of \( u \) at each of the \( N \) internal nodes. By feeding these values back into the local data vectors, \( d^{(k)} \), the local systems may be used to extract any other field values as may be required for post-process analysis (such as partial derivatives, or stresses in the case of linear elasticity). For more details about this highly convergent meshless collocation numerical scheme see [1].

In the sparse global assembly, the PDE governing operator and the corresponding boundary conditions of the problem have already been imposed within the local collocation systems. The number of non-zero matrix entries in each row corresponds to the number of solution centres in the associated local system; i.e. the number of nodes on the stencil periphery. Since only the peripheral nodes appear in the global system, the number of non-zero entries is smaller than for an equivalent Finite-Difference type method (see Figure 1); i.e. the sparse global matrix has a smaller bandwidth. With an appropriate sparse linear system solver the method may be scaled efficiently to very large datasets.

4 NUMERICAL RESULTS

Analytical results relating to rates of convergence and other numerical properties are extremely difficult to obtain for methods based on local RBF collocation. Therefore, in this work we focus on assessing the performance of the proposed finite collocation method qualitatively, using benchmark numerical problems with known analytical solutions. In each case we focus on analysing the solution accuracy and convergence rates, showing that high convergence rates may be reliably obtained.

In each numerical example we use a 5\( \times \)5 stencil configuration, with additional staggered PDE centres, as represented in Figure 1c. Larger stencil sizes may be used to obtain higher convergence rates at the expense of increased computational cost and somewhat increased sensitivity to basis function flatness (see [1]). The results presented below use a non-dimensional shape parameter of value \( c^* = 100 \), scaled against the local node separation.

4.1 Linear Elasticity Problem

Performance with the linear elasticity equation is assessed for a thin plate with a circular hole under uniform traction. The analytical expression for displacement and stress is given by
Here we choose material properties of $E=210\text{GPa}$ and $\nu=0.3$, representing mild steel, with a hole radius $a=1\text{m}$ and a far-field traction of $100\text{MPa}$. We exploit the symmetry of the problem to examine a single quadrant, constraining the solution domain to a 4m x 4m region. Over the hole we enforce the appropriate zero-traction condition, at the lines of symmetry we impose the symmetric condition, and at the far-field boundaries we enforce the analytical surface traction field, as defined by $\tau_i = \sigma_{ij} n_j$.

The dataset is generated by distributing $N+1$ nodes over the hole, and $N+1$ nodes in the radial direction (i.e. $1 \leq r \leq 4$). A simple RBF interpolation is then performed over the domain to map this cylindrical distribution onto the full Cartesian problem domain (see Figure 2a). The resulting stresses are largely concentrated in the region around the hole, and their accurate prediction is a challenging task (see Figure 2b).

The proposed method exhibits strong spatial convergence and provides highly accurate solutions. Figure 3 shows the $L_2$ relative error obtained from datasets of size $N=20, 30, 40, 60, 80$. In this log-log plot, the gradient of the curve represents the spatial convergence rate.
between successive datasets. The approximate convergence rates, obtained by line of best fit, are summarised in Table 1. The convergence rates are of orders between 6.5 and 7 for displacement and for all stress components.

<table>
<thead>
<tr>
<th></th>
<th>( u )</th>
<th>( \sigma_{11} )</th>
<th>( \sigma_{12} )</th>
<th>( \sigma_{22} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.90</td>
<td>6.79</td>
<td>6.75</td>
<td>6.56</td>
</tr>
</tbody>
</table>

Table 1: Approximate convergence rates (plate with a circular hole)

### 4.1 Poroelasticity Problem

To analyse the performance of the proposed RBF-FC method with a coupled poroelasticity problem we consider a porous cylindrical annulus subject to a large internal pressure. The solution is obtained in two stages; the first stage computes the pressure field and its Cartesian derivatives at each node within the domain, by solving Darcy’s equation (1). The second stage solves the inhomogeneous linear elasticity equation (7), where the inhomogeneous term is obtained from the computed pressure gradient.

For a cylindrical annulus with inner radius \( a \) and outer radius \( b \), with pressure \( p_i \) at \( r = a \) (i.e. internal pressure) and pressure \( p_o \) at \( r = b \) (i.e. external pressure), the analytical solution for the pressure field is as given by equation (24). We set the inner pressure as \( P_i = 100 \text{MPa} \) and the outer pressure as \( P_o = 10 \text{MPa} \).

\[
p = \frac{p_o - p_i}{\ln(b/a)} \ln(r) + \frac{p_i \ln(b) - p_o \ln(a)}{\ln(b/a)}
\]

(24)

We choose a dataset of size \((N+1)\times(N+1)\) with nodes distributed radially (see Figure 4a), and examine solution accuracy for \(N = 20, 30, 40, 60, 80\). The Darcy flow problem is solved to a very high level of accuracy in each case, as shown in Figure 6. The analytical expression (24) is predicted to an \(L_2\) relative error of \(3.4\times10^{-7}\) on the \(N = 20\) dataset, and \(2.9\times10^{-11}\) on the \(N = 80\) dataset. Errors found in the Cartesian gradients less than a factor of two higher.

![Fig 4: Darcy-flow: Dataset, boundary conditions and pressure-field representation](image-url)
Following computation of the pressure-field, the inhomogenous linear elasticity equation may be solved. An analytical solution for displacement and stress may be computed as:

\[
\begin{align*}
    u_r &= Ar + Br^{-1} + \xi_1 \left( \frac{r}{2} \ln(r) - \frac{r}{4} \right) + \xi_2 \frac{r}{2} \\
    u_\theta &= \sigma_{rr} = (\lambda + \mu) \left[ 2A + \xi_1 \ln(r) + \xi_2 \right] - 2\mu Br^{-2} + \frac{\xi_1 \mu}{2} - ap \\
    \sigma_{\theta\theta} &= (\lambda + \mu) \left[ 2A + \xi_1 \ln(r) + \xi_2 \right] + 2\mu Br^{-2} - \frac{\xi_1 \mu}{2} - ap \\
    \sigma_r\theta &= 0
\end{align*}
\]

where:

\[
\begin{align*}
    A &= \frac{1}{2(\lambda + \mu)} \left[ p_1 a^2 - p_0 b^2 \left( 1 - \frac{\alpha \mu}{\lambda + 2\mu} \right) + \frac{\alpha \mu (p_1 - p_0)}{2(\lambda + 2\mu) \ln(b/a)} \right] \\
    B &= \frac{a^2 b^2 (p_1 - p_0)}{2\mu (b^2 - a^2)} \left[ 1 - \frac{\alpha \mu}{\lambda + 2\mu} \right] \\
    \xi_1 &= \frac{\alpha (p_0 - p_1)}{(\lambda + 2\mu) \ln(b/a)} \\
    \xi_2 &= \frac{\alpha (p_1 \ln(b) - p_0 \ln(a))}{(\lambda + 2\mu) \ln(b/a)}
\end{align*}
\]

At the \( x = 0 \) and \( y = 0 \) boundaries we enforce symmetry (i.e. zero normal displacement and zero tangential traction), and the appropriate pressure-traction condition is applied at the inner and outer surfaces (see Figure 6a). We use parameters of \( E = 27.6 \text{ GPa} \) and \( \nu = 0.15 \), representing Weber sandstone. The resulting relative errors in the displacement and stress-fields are represented in Table 2.

Figure 5: Boundary conditions and stress representation for poroelastic deformation
Errors are low for each examined dataset, with strong reductions in error observed as the dataset is refined. On the finest \(N = 80\) dataset we observe errors of order \(10^{-9}\) to \(10^{-10}\). Errors for the stress field are, in each case, of similar magnitude to errors observed in the displacement field. The maximum relative displacement error is never more than a factor of four larger than the average relative error taken using the \(L_2\) norm. The spatial convergence is represented in Figure 6; taking a line of best fit average we see that the convergence rate is roughly seventh order for each of the examined fields (displacement, stress, and pressure).

5 CONCLUSIONS

A high-resolution meshless numerical solution has been described for coupled poroelastic analysis, based on the RBF finite collocation (RBF-FC) approach. The poroelastic analysis procedure begins with the solution of Darcy’s equation, in order to obtain the fluid-pressure field throughout the domain. From this solution the pressure gradient is computed at each internal node, and is used as an inhomogeneous term for the linear elasticity equation. By solving this equation with appropriate pressure-traction boundary conditions, the displacement and stress fields may be obtained for the full poroelastic problem.

The RBF-FC method allows highly accurate solutions to be obtained. The RBF-FC solution procedure has been demonstrated for a benchmark linear elasticity test case, for the solution of Darcy’s equation, and for a coupled poroelastic problem. In each of the cases examined the method is able to produce highly accurate solutions, even on relatively coarse datasets, and demonstrates better than sixth order convergence in each case.
ACKNOWLEDGEMENTS

This work was funded by the MUSTANG EU project (contract number 227286), which investigates the long-term geological storage of CO2

REFERENCES

LOCAL RADIAL BASIS FUNCTION COLLOCATION METHOD FOR SOLVING THERMO-MECHANICS OF HOT SHAPE ROLLING OF STEEL

U. HANOGLU* AND B. ŠARLER†*

*University of Nova Gorica
Laboratory for Multiphase Processes
Vipavska 13, SI-5000 Nova Gorica, Slovenia
email: umut.hanoglu@ung.si, www.ung.si

† Institute of Metals and Technology
Laboratory for Simulation of Materials and Processes
Lepi pot 11, SI-1000 Ljubljana, Slovenia
email: bozidar.sarler@imt.si, www.imt.si

Key words: Hot Shape Rolling, Thermo-Mechanical Modeling, Meshless Methods, Local Radial Basis Function Collocation Method, Steel.

Abstract. The aim of this paper is to demonstrate the suitability of the novel Local Radial Basis Function Collocation Method (LRBFCM) [1] in a coupled thermo-mechanical problem of hot shape rolling of steel. The physical concept of such a large deformation problem is based on a two dimensional traveling slice model [2], which assumes deformation and heat flow only in the perpendicular direction to rolling. The solid mechanics is, respectively, based on the steady Navier’s equation and the thermal field on the transient heat conduction equation. The displacement and traction boundary conditions are assumed in the mechanical model and Dirichlet and Neumann boundary conditions in the thermal model, both specific for hot shape rolling. The solution procedure is based on local collocation on a five noded influence domain with multiquadrics radial basis functions, augmented with the first order polynomials. The steel used in the calculations is assumed to have an ideal plastic behavior which obeys von Misses flow rule, defined by effective stress \( \sigma \) in terms of effective strain \( \varepsilon \), effective strain rate \( \dot{\varepsilon} \) and temperature \( T \). The LRBFCM results of hot shape rolling of steel for a continuous 5 stand rolling mill in Štore Steel company are presented for the case of rolling of a rectangular billet with initial dimension 80 x 95 mm to a circular bar with diameter of 60 mm. The advantage of the meshless method is in accuracy and straightforward node generation, that does not require any polygonisation. The paper presents one of the increasingly emerging examples of the use of the LRBFCM in industrial applications.

1 INTRODUCTION

Hot shape rolling of steel, that usually follows continuous casting of billets and blooms, provides different shapes of long products for various applications such as automotive
industry, construction, shipbuilding, railway, etc. The demand is constantly changing due to newly introduced designs and the production should be quickly adaptable. Respectively, computational modeling of the continuous casting and hot rolling gains its importance for better understanding, control, and better insight into these processes. It helps to improve the quality, productivity, safety and environmental impact of the production. The principal goal of rolling simulation is to connect the process variables, such as the rolling speed, temperature, rolling stand geometry, to the temperature, strain, strain rate, stress field in the billet, its microstructure with static and dynamic recrystallization, and calculation of the rolling torque and power. The modeling of rolling started with Hitchcock [3] where he solved the problem of the roll deformation. The current state of the rolling technology can be perceived from [4] and rolling modeling from [5].

The majority of the simulations in solid mechanics are done by using Finite Element Method (FEM) [6]. This method requires meshing as pre-processing which might be problematic and time consuming, especially in case of rolling, where the shape is constantly and drastically changing.

In the last decade, meshless numerical methods [7] started to represent an appealing alternative to the classical numerical methods, such as FEM. Meshless method is a numerical technique that uses a set of arbitrary distributed nodes, both on the boundary and within the computation domain, to represent the solution of physical phenomena. The main feature of meshless methods is omission of the polygonalisation between the nodes which can be remarkably demanding, particularly in realistic 3D geometrical situations. One of the simplest meshless methods, able to solve the fluid flow problems and solid mechanics problems is Local Radial Basis Function Collocation Method (LRBFCM). This method was first developed in [8] for elasticity problems and in [1] for diffusion problems. The idea behind this method is to approximate the function and its derivatives locally over a set of neighboring nodes using RBFs [9] and to use collocation for determining the expansion coefficients. The method has been recently applied to numerous scientific and engineering problems, connected with fluid mechanics [10-14] as well as solid mechanics [15-17]. In this paper the simulation of hot shape rolling is performed by using LRBFCM.

2 PHYSICAL MODEL

![Figure 1: Scheme of slice model of hot shape rolling](image-url)
2.1 Slice model assumptions

A scheme of the hot rolled billet is shown in Figure 1 with sketch of traveling slices and computational nodes. The temperature and the deformation field of the slice can be computed from the known time dependent boundary conditions. The slice time is on the other hand associated with the position in the rolling mill. The Cartesian coordinate $p_z$ measures the length from the beginning of rolling. The billet geometry in $z$ direction is assumed straight and the thermal and mechanical interactions in rolling direction are neglected. The $p_z$ coordinate can thus be considered parabolic, while the $p_x$ and $p_y$ coordinates are elliptic. In this way all fields at a given $p_z$ coordinate depend only on the slice history, including its cooling intensity and deformation through the rolls, as a function of time. The slices form at the $p_z$-start, longitudinal coordinate of rolling and travel in the direction of the $i_z$ base vector with the rolling speed $V_{roll}$, (see Figure 1). For calculating the cooling intensity of the slice as a function of time, a connection between the $p_z$ coordinate of the rolling mill and the slice history $t$ is needed

$$z(t) = \int_{t_{start}}^{t} V_{roll}(t') \cdot dt' + p_{z,start}; V_{roll}(t) = V_{roll}(t) \cdot i_z, V_{roll}'(t') = S(t') / S(t_{start})$$

where $t_{start}$ represents the initial time of a slice and $S(t_{start})$ and $S(t')$ represent the initial and the cross-section of the slice at time $t'$, respectively. In case when the rolling speed is constant and there is no deformation, we obtain the following simple connection between the $p_z$ coordinate of the rolling mill and the slice history $t(z) = t_{start} + (p_z - p_{z,start}) / V_{roll}$. A quadrant symmetry of the geometry and fields of the slice with domain $\Omega$ and boundary $\Gamma$ are assumed and the north-east quadrant (I-st quadrant) is computationally coped with, described by Cartesian coordinates $p_x$ and $p_y$ and base vectors $i_x$ and $i_y$.

2.2 Thermal model

The governing equation of the thermal model is

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T) + \dot{Q},$$

with $\rho, c_p, T, k, \dot{Q}$ standing for density, specific heat, temperature, time, thermal conductivity and heat generation rate due to deformation, respectively. The boundary temperatures of the travelling slice were obtained from the Robin type energy balance equations in point $p$

$$-k \frac{\partial T(p,t)}{\partial n_r} = h\left[T(p,t) - T^a_r(p,t)\right]; p \in \Gamma$$

(3)
at slice boundaries $\Gamma$, with $h$ standing for the heat transfer coefficient, $T_0^R$ standing for the reference temperature. On the two quadrant symmetry axis, $h$ is set to 0. The heat source due to the deformation is calculated as

$$Q = \int_{\Gamma} \bar{\sigma} d\bar{\varepsilon},$$  (4)

where $Q$ is the heat generated due to plastic work which is defined in terms of effective stress $\bar{\sigma}$ and effective strain $\bar{\varepsilon}$.

### 2.3 Mechanical model

The main aim of the mechanical model is to calculate the displacement field of the slice due to deformation by the roll, in order to get the new shape of the slice. The boundary is divided into natural $\Gamma_u$ and essential $\Gamma_i$ part $\Gamma = \Gamma_u \cup \Gamma_i$. The governing equation of the mechanical model is

$$L^T \sigma + b = 0,$$  (5)

where $L$ is the derivative operator, $\sigma$ is the vector of stresses, and $b$ is the body force vector, considered $b = 0$. Two dimensional plane strain model is assumed. The essential boundary conditions are considered between the slice and the roll and natural boundary condition are assumed at the boundary parts with no contact as well as along the symmetry lines. The natural boundary conditions are defined as

$$n_i \sigma_{ii} + n_j \sigma_{ij} = \bar{\tau}_i; \ i, j = x, y,$$  (6)

where in $n_i$ is component of unit normals and $\bar{\tau}_i$ is the prescribed shear stress component, equal to 0. The essential boundary condition is described as

$$u_i = \overline{u}_i; \ i = x, y,$$  (7)

where $u_i$ is the displacement and $\overline{u}_i$ is the prescribed displacement due to the action of the roll. The material is assumed to be ideally plastic which means that all the energy put into material turns into plastic deformation at yield stress. The yielding of a solid material is defined by von Mises flow rule. A non-linear plastic modulus $H_p$ is used instead of elastic parameters to relate the stresses with strains

$$\sigma_y = s_y = \frac{2}{3} H_p \epsilon_y,$$  (8)

where $s_y$ is the stress deviator tensor, $H_p = \partial \bar{\sigma}/\partial \bar{\varepsilon}$ is the plastic modulus [18] which is the derivative on the effective stress ($\bar{\sigma} = \sqrt{(3/2)\sigma_{ij}\sigma_{ij}}$) – effective strain ($\bar{\varepsilon} = \sqrt{(2/3)\epsilon_{ij}\epsilon_{ij}}$) curve. In metal deforming process, the stress deviator tensor can be assumed to be equal to the
components of the stress tensor. The strain vector \( \varepsilon \) can be written in terms of displacement vector as
\[
\varepsilon = \mathbf{L} \mathbf{u}.
\] (9)

Therefore, the strong formulation of deformation problem gives two individual balance equations in each principle direction on a 2D slice in terms of displacement \( u_i, i = x, y \) which are
\[
\frac{2}{3} H \frac{\partial^2 u_x}{\partial p_x^2} + \frac{1}{3} H \left( \frac{\partial^2 u_y}{\partial p_y^2} + \frac{\partial^2 u_x}{\partial p_x \partial p_y} \right) = 0
\] (10)
\[
\frac{2}{3} H \frac{\partial^2 u_y}{\partial p_y^2} + \frac{1}{3} H \left( \frac{\partial^2 u_x}{\partial p_x \partial p_y} + \frac{\partial^2 u_y}{\partial p_y^2} \right) = 0
\] (11)

3 SOLUTION PROCEDURE

3.1 Solution strategy

The coupled thermo-mechanical simulation is structured in the following way. First, nodes are generated on the un-deformed slice and temperature and deformation fields are calculated. Afterwards, the nodes are newly generated on the new, deformed shape of the slice, based on elliptic node generation [19]. The calculated values are interpolated to the new position of the nodes and the procedure is repeated, by taking into account that the calculated values of the old slice serve as initial values for the next slice. If there is a roll contact, the new shape is calculated with the mechanical model.
\[
p_i = p_{i0} + u_i; \ i = x, y
\] (12)

Afterwards, the temperature is calculated whether there is a contact or not. We assume that the time discretization is made in explicit Euler way.
\[
T = T_0 + \Delta t \left[ \nabla \cdot (k \nabla T) + \dot{Q} \right]_0
\] (13)

A detailed description of the LRBFCM solution of the thermal slice model is given in [11]. The slice is discretized into \( N_\Omega \) domain and \( N_\Gamma \) boundary nodes, in total \( N = N_\Omega + N_\Gamma \) nodes. A five noded sub-domain is associated with each of the nodes. The initial fields are interpolated on each of the subdomains by using collocation with radial basis functions, augmented with first order polynomials. In order to get the results for the field \( \phi \), it needs to be approximated with interpolation function \( \psi_n(p) \) and coefficients \( \alpha_n \)
\[
\phi(p) = \sum_{n=1}^{N_\Omega + N_\Gamma} \psi_n(p) \alpha_n,
\] (14)
where the interpolation function is defined by scaled multiquadrics radial basis functions inside five noded influence domains ($N_\omega = 5$), and first order polynomials ($N_p = 3$). A flowchart of the thermomechanical simulation is given in Figure 2.

\[
\psi_n(p) = \sqrt{\left(\frac{p_x - p_{x_{\text{on}}}}{x_{\text{max}}}\right)^2 + \left(\frac{p_y - p_{y_{\text{on}}}}{y_{\text{max}}}\right)^2 + c^2},
\]

\[
\psi_{n+1}(p) = 1, \quad \psi_{n+2}(p) = p_x - x_{\text{mea}}, \quad \psi_{n+2}(p) = p_y - y_{\text{mea}},
\]

where $x_{\text{max}}$, $y_{\text{max}}$, $x_{\text{mea}}$, $y_{\text{mea}}$ represent maximum distance between the five nodes in $i_x$ and $i_y$ directions, and mean position of the five nodes in $i_x$ and $i_y$ directions, respectively. $c$ is set to 32. The calculation of the coefficients $\alpha_n$ is elaborated in [20]. The derivatives of the functions, needed in the solution, are calculated from the derivatives of the radial basis functions.
The solution of the thermal model requires inversion of a matrix of the size \( N_n + N_p \) for each of the subdomains and no global matrix is formed.

### 3.2 Solution of mechanical model

The mechanical model is solved by expressing the displacements by radial basis functions

\[
\psi_n(p) = \sum_{n=1}^{N_n+N_p} \phi_n(p) \alpha_n, \quad \zeta = x, y
\]

The governing equation for node \( m \) becomes

\[
\frac{2}{3} H_{pm} \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p^2} \alpha_{xm} + \frac{1}{3} H_{pm} \left( \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p_x \partial p_y} \alpha_{xn} + \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p_x \partial p_y} \alpha_{yn} \right) = 0
\]

\[
\frac{2}{3} H_{pm} \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p^2} \alpha_{ym} + \frac{1}{3} H_{pm} \left( \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p_x \partial p_y} \alpha_{xn} + \sum_{n=1}^{N_n+N_p} \frac{\partial^2 \psi_{mn}}{\partial p_x \partial p_y} \alpha_{yn} \right) = 0
\]

The coefficients \( \alpha_{xn}; \zeta = x, y \) are expressed with the unknown values of the displacements \( u_{xm}; \zeta = x, y \) and a global \( 2N \times 2N \) sparse matrix is formed for calculation of the unknown displacements. The assembling of the respective global matrix is described in [21]. The heat generation rate is calculated as

\[
Q_m = \eta \left( \sigma_m^l + \sigma_m^{l-1} \right) \left( \bar{e}_m^l - \bar{e}_m^{l-1} \right) / 2(t^l - t^{l-1})
\]

where \( \eta \) is Taylor-Quinney parameter, defining the ratio of mechanical work turning into heat, \( m \) is node number and \( l \) is slice number.

### 3.3 Elliptic node generation

After each deformation step, the nodes on the boundary are redistributed with equal distance. After that, the internal nodes need to be distributed as orthogonal as possible with its neighboring nodes. The new arrangement of the nodes is done in two steps. First, Trans Finite Interpolation (TFI) [19] is used for aligning of the nodes in accordance with the boundary nodes. This is followed by Elliptic Node Generation (ENG) [19] which repositions them in an iterative way to make them more orthogonal. 100 iterations are allowed for this purpose. The performance of described LRBFCM has been assessed for large deformations by comparison with the commercial FEM code DEFORM [16].
4 NUMERICAL EXAMPLE

In the numerical example a continuous rolling schedule with 5 rolling stands is considered. The parameters, used in the simulation are given in Table 1 and the characteristics of the rolling mill are given in Table 2. The result are shown in terms of displacement field of an 95 × 80 mm initial billet shape, rolled into circular bar with radius \( r = 30 \) mm. Figure 3 represents the increase of the rolling speed due to the deformation, Figure 4 represents the node positions at the start of rolling and after each of the rolling stands. Figure 5 represents the displacement field after each of the rolling stands.

Table 1: Thermal and mechanical parameters used in calculations where \( A = 4.16 \), \( b = 0.23 \), \( c = 0.214 \).

\[ R \] is the gas constant and \( T^K \) is the absolute temperature.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat transfer coefficient to air</td>
<td>( h_{air} ) 20 W/m²K</td>
</tr>
<tr>
<td>Heat transfer coefficient to roll</td>
<td>( h_{roll} ) 16000 W/m²K</td>
</tr>
<tr>
<td>Thermal conductivity of steel</td>
<td>( k ) 29 W/mK</td>
</tr>
<tr>
<td>Specific heat of steel</td>
<td>( c_p ) 630 J/kgK</td>
</tr>
<tr>
<td>Initial rolling temperature</td>
<td>( T_{fur} ) 1250 °C</td>
</tr>
<tr>
<td>Initial rolling speed</td>
<td>( V_{roll} ) 0.76 m/s</td>
</tr>
<tr>
<td>Ambient temperature</td>
<td>( T_{am} ) 25 °C</td>
</tr>
<tr>
<td>Roll temperature</td>
<td>( T_{roll} ) 500 °C</td>
</tr>
<tr>
<td>Taylor-Quinney parameter</td>
<td>( \eta ) 0.9</td>
</tr>
<tr>
<td>Time step</td>
<td>( dt ) 10⁻⁴ s</td>
</tr>
<tr>
<td>Effective stress</td>
<td>( \sigma \left( \bar{e}, \dot{\bar{e}}, T \right) = AE^{\frac{b}{c}} \dot{\bar{e}} \exp \left( \frac{38000}{RT^K} \right) ) GPa</td>
</tr>
</tbody>
</table>

Table 2: Rolling schedule used in the simulation.

<table>
<thead>
<tr>
<th>Stand Number</th>
<th>Position (m)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45</td>
<td>OVAL 85 (H)</td>
</tr>
<tr>
<td>2</td>
<td>3.45</td>
<td>OVAL 85 (H)</td>
</tr>
<tr>
<td>3</td>
<td>6.05</td>
<td>ROUND 40 (V)</td>
</tr>
<tr>
<td>4</td>
<td>9.65</td>
<td>OVAL 55 (H)</td>
</tr>
<tr>
<td>5</td>
<td>12.65</td>
<td>ROUND 30 (V)</td>
</tr>
</tbody>
</table>
Figure 3: Velocity of slices in the rolling direction.

Figure 4: Initial shape of the slice and displacement vectors of the slice at the exit of each of the five rolling stands.
5 CONCLUSIONS AND FURTHER RESEARCH

In this paper a coupled thermo-mechanical slice model is given for simulation of hot shape rolling. The solution procedure is based on a fully meshless LRBFCM. This method can be applied to large deformations by using TFI and ENG between the deformation steps. The simulation is applied to five consecutive rolling stands of the continuous rolling mill, operating in Štore Steel Company, Slovenia. Ongoing research is focused on testing the simulations with realistic material properties and comparisons with the thermal measurements by infrared thermography and laser true dimensional measurements. A microstructure model, based on the point automata formulation [22], which would take into account the deformation of the grains and the static and the dynamic recrystallization is underway. It will be coupled to the represented macroscopic model.

6 ACKNOWLEDGEMENTS

This paper forms as a part of the project L2-3651, supported by Slovenian Grant Agency and Štore Steel Company.

REFERENCES


MESHLESS 2D DIRECT NUMERICAL SIMULATION AND HEAT TRANSFER IN A BACKWARD-FACING STEP WITH HEAT CONDUCTION IN THE STEP

ANDRÉS G. VIDAL¹, ALAIN J. KASSAB², EDUARDO A. DIVO³,⁴

¹ Department of Mechanical, Materials and Aerospace Engineering
University of Central Florida
Orlando, FL 32816, United States
e-mail: Andres.Vidal@knights.ucf.edu

² Department of Mechanical, Materials and Aerospace Engineering
University of Central Florida
Orlando, FL 32816, United States
e-mail: Alain.Kassab@ucf.edu

³ Department of Mechanical Engineering
Embry-Riddle Aeronautical University
Daytona Beach, FL 32114, United States

⁴ Department of Mechanical, Materials and Aerospace Engineering
University of Central Florida
Orlando, FL 32816, United States
e-mail: edivo@mail.ucf.edu

Key words: Meshless Methods, Computing Methods, Direct Numerical Simulation, Heat Transfer.

Abstract. A meshless direct pressure-velocity coupling procedure is presented to perform Direct Numerical Simulations (DNS) and Large Eddy Simulations (LES) of turbulent incompressible flows in regular and irregular geometries. The proposed method is a combination of several efficient techniques found in different Computational Fluid Dynamic (CFD) procedures. With this new procedure, preliminary calculations with 2D steady state flows show that viscous effects become negligible faster that ever predicted numerically. The fundamental idea of this method lays on several important inconsistencies found in three of the most popular techniques used in CFD, segregated procedures, as well as in other formulations. The inconsistencies found become important in elliptic flows and they might lead to some wrong solutions. Preliminary calculations done in 2D laminar flows, suggest that the numerical diffusion and interpolation error are much important at low speeds, mainly when both, viscous and inertia forces are present. With this competitive and efficient procedure, the solution of the 2D Direct Numerical Simulation of turbulent flow with heat transfer on a backward-facing step is presented. The thermal energy is going to be transferred to the fluid through conduction on the step, with both constant temperature and heat flux conditions in the back wall of the step. The variation of the local Nusselt Number through the
wall will be studied and its corresponding effect in the energy transfer to the fluid.

1 INTRODUCTION

Since the introduction of the Projection Method by Harlow and Welch [1], the science of Computational Fluid Dynamics (CFD) has become a fundamental tool for engineering calculations and design. Basically, the velocity-pressure coupling is done in a segregated way, one equation at a time. After the publication of this procedure, almost all numerical methods developed in Computational Fluid Dynamics (CFD), with some minor modifications, the original flow equations are transformed into a series of consecutive and explicit equations for velocity, pressure and mass correction, this last one needed to satisfy the mass balance.

For highly demanding problems such as Direct Numerical Simulation (DNS) and Large-Eddy Simulation (LES), the normal approach is to extend the current CFD procedures and perform some minor changes intended to reduce the so-called numerical diffusion error. Once again, almost all procedures transform the original flow equations into a system of segregated equations.

In spite of the improvement in all CFD techniques, even today the solution of complex elliptic problems, such as the backward-facing step or lid-driven cavity is still a mayor challenge. These two cases have produced by far the largest amount of differences in numerical results between procedures. Many authors have explained this effect as bifurcation of the solution. The idea of this work is to present a meshless localized RBF procedure to solve the flow equations in the original form, so that there is no simplification or approximation of any boundary condition. The velocity-pressure coupling procedure is the same one developed in [2]. The staggered point distribution approach (or grid) is selected and the RBF scheme is chosen to perform any necessary interpolation. Finally, in order to keep the numerical diffusion at a very low level, the well known flux-limiting scheme will be used in the discretization of the convection term.

2 PREVIOUS WORK

In recent years, several studies have been performed where the influence of geometry and/or flow conditions on the heat transfer are analyzed. For example, in [3], a LES simulation is performed to study the heat transfer and fluid flow of a turbulent separating flow past a backward-facing step. A fully collocated grid is used and compressible flow is assumed at low Mach numbers. The bottom wall downstream of the step was supplied with a uniform heat flux and different heat flux levels were analyzed. Dramatic variation of the wall temperatures in the recirculation zone was observed with a steep increment in wall temperature close to the step followed by a decrease in the convective heat transfer. Another conclusion is that the viscous sub-layer played a critical role in controlling the heat transfer rate. The Reynolds analogy was not valid in the recirculation region.

On the other hand, in [4], simulations of turbulent flow adjacent a 2D backward-facing step are presented to explore the effects of step height on separation and heat transfer. The stepped wall is kept at a constant heat flux and the Reynolds number is fixed at 28,000. The remaining walls are isolated. Here, a two-equation low-Reynolds model is employed. The primary and
secondary recirculation regions increase in size as the step size increases. The bulk
temperature increases more rapidly as the step height increases. Additionally, increasing the
step height causes the magnitude maximum kinetic energy to increase.

An interesting study can be seen in [5, 6], where a simulation is presented for a backward-
facing step flow and heat transfer inside a channel with ribs turbulators. The problem was
investigated for Reynolds numbers up to 32000. The effect of a step height, the number of
ribs and the rib thickness on the flow and thermal field were investigated. The effect of
turbulence was modeled by using a k-ε model with its wall function formulas. The obtained
results show that the strength and size of the recirculation zones behind the step are increased
with the increase of contraction ratio (i.e. with the in-crease of a step height). The size of
recirculation regions and the reattachment length after the ribs are decreased with increasing
of the contraction ratio.

In other works, such as [7], the effect on a pulsating flow and an oscillating wall is studied
with a possible application in chamber combustors. Results of steady state and transient
calculations are presented as well as the evaluation of several turbulence models. It was
observed that the variations in the excitation frequency of the inlet flow and wall vibrations
have an influence on the instantaneous heat transfer coefficient profile. However, significant
effect on the time mean value and position of the heat transfer peak is only visible for the inlet
velocity profile fluctuations with frequency approximately equal to the turbulence bursting
frequency.

In [8], a numerical study is presented where a locally turbulent oscillating jet is used to
evaluate the separation and reattachment. A three-equation turbulence model is used and
different forcing frequencies were evaluated. A constant heat flux was imposed in the bottom
wall. The time dependent distributions of the stresses indicated that heat transfer is
significantly enhanced at the most effective frequency.

An experimental study can be found in [9], in which the backward-facing step is controlled
by equipping a slit at the bottom cornet of the step. It was found that the heat transfer and
pressure drop characteristics are controlled by the flow ratio. When the suction flow was 0.6,
the highest performance was obtained.

Finally, in [10], the effect of a baffle in the entrance of the expansion is analyzed. Comparing
the results with and without the baffle, its presence improves the average Nusselt
number to a maximum of 190% for the heating step and 150% for the heating section.
Additionally, a slight movement of the baffle can cause a drastic change in the flow structure
and temperature distributions.

3 VELOCITY-PRESSURE COUPLING

The finite volume method proposed by Patankar [11], with SIMPLE and SIMPLER
techniques as velocity-pressure coupling procedures, is the most popular method in CFD.
These coupling schemes are used in most commercial and noncommercial CFD packages,
using finite volume, finite difference or finite element method as the main discretization
procedures.

However, these coupling procedures are known to produce significant numerical diffusion.
The most general procedure (SIMPLER) can be resumed as:
1. Discretize momentum equation ($\mathbf{p} = \rho \mathbf{u} - \mathbf{r} \cdot \mathbf{g}$):

$$\tilde{v}_p = \frac{1}{a_p} \sum a_{ab} \tilde{v}_{ab}, \quad \tilde{v}_p = -\tilde{v}_p + \frac{1}{a_p} \nabla \tilde{p}$$

(1)

2. Compute pressure by introducing Eq. (1) into continuity equation:

$$\nabla \left( \frac{1}{a_p} \nabla \tilde{p} \right) = \nabla \cdot \tilde{v}_p$$

(2)

3. Update pressure in Eq. (1) and solve for velocity.

4. Correct velocity to enforce mass continuity:

$$\tilde{v} = \tilde{v}^* + \tilde{v}' \quad \tilde{p} = \tilde{p}^* + \tilde{p}'$$

(3)

$$\nabla \left( \frac{1}{a_p} \nabla \tilde{p} \right) = -\nabla \cdot \tilde{v}^* + \nabla \cdot \tilde{v}'$$

(4)

where the term $\nabla \cdot \tilde{v}'$ is frequently neglected. The first comment that it is convenient to make to this procedure is that, the main coefficient $a_p$ is inside all partial derivatives in pressure and mass-correction equations. The structure of this coefficient is:

$$a_p = \nu C_L - a_s C_{DX} - v_s C_{DY}$$

(5)

with $C_L$, $C_{DX}$ and $C_{DY}$ the coefficients of the discretization scheme. For a non-uniform mesh, the coefficient $a_p$ is a function of the position. This coefficient will produce clearly numerical diffusion in Eqs. (1) and (2).

The only way that Eqs. (14) and (16) will not produce numerical diffusion is with a mesh of constant spacing and using central differencing in convection terms. Since $a_p$ gathers the diffusion and convection terms, central differencing for convection derivative will not have any coefficient. With a uniform mesh, the diffusion term of $\tilde{v}$ is zero and the Eq. (4) becomes exactly the same projection procedure of Harlow and Welch [1].

By updating velocity in SIMPLER, the procedure becomes the same algorithm [1], but in SIMPLE, updating pressure with $\tilde{p}'$ is updating pressure with the velocity potential. This explains why SIMPLE takes so many iterations to converge and why this procedure works only when velocity is corrected and not pressure (as initially inferred).

Another problem that SIMPLE and SIMPLER have is, in pressure equation (4), boundary conditions (pressure coefficient zero in all boundaries) imply that the viscosity of the fluid is infinite at the wall, inflows and outflows. The condition of viscosity infinite is correct at the wall but, at inflows and outflows is evidently incorrect. At inflows, this numerical change in
the viscosity of the fluid produces a force that helps the motion of the fluid but, at outflows, this change in viscosity produces a force that decelerates the fluid. This is one reason why pressure equation (4) has problems converging, unless a block-correction or multigrid algorithm is used.

A useful alternative that solves the problems associated to both segregated and direct full coupling procedures is presented in [12]. In general, this scheme uses the segregated grid arrangement in the same way as finite volume method.

The fundamental aspect of this coupling approach is that the velocity-pressure coupling is done with the momentum and continuity equations in the original form. After substituting finite differencing expressions for both, viscous and convection terms, as well as the pressure gradient, the momentum and continuity equations can be written in the form:

\[
\begin{align*}
\alpha_p^u u_p + C_p^u (p_p - p_E) &= b_p^u \\
\alpha_p^v v_p + C_p^v (p_p - p_N) &= b_p^v \\
D_p^u (u_p - u_w) + D_p^v (v_p - v_s) &= 0
\end{align*}
\]  

(6)

In the system (6), pressure and velocity components are located at different points, as expected in a staggered point distribution. The coupling of \(u\), \(v\) and \(p\) is performed by writing, in all possible ways, a linear system of 3 equations of the kind:

\[
\begin{bmatrix}
\alpha_{11} & 0 & a_{13} \\
0 & a_{22} & a_{23} \\
\alpha_{31} & a_{32} & 0
\end{bmatrix}
\begin{bmatrix}
u \\
0 \\
p
\end{bmatrix} =
\begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix} 
\]  

(7)

whose solution is:

\[
\begin{bmatrix}
u \\
p
\end{bmatrix} = 
\begin{bmatrix}
\frac{a_{23}a_{32}b_1 - a_{13}a_{32}b_2 + a_{13}a_{31}b_3}{a_{1}a_{23}a_{32} + a_{1}a_{31}a_{32} + a_{4}a_{31}a_{32}} \\
\frac{-a_{23}a_{32}b_1 + a_{13}a_{32}b_2 + a_{13}a_{31}b_3}{a_{1}a_{23}a_{32} + a_{1}a_{31}a_{32} + a_{4}a_{31}a_{32}} \\
\frac{a_{1}a_{23}b_1 + a_{1}a_{31}b_2 - a_{1}a_{23}b_2}{a_{1}a_{23}a_{32} + a_{1}a_{31}a_{32} + a_{3}a_{31}a_{32}}
\end{bmatrix} 
\]  

(8)

Having in mind the linear system (6) and its solution (7)-(8), the discretized system (6) may be rearranged in many different ways. The solution procedure presented in [2] re-writes Eq. (6) in all possible permutations of both velocity components adjacent to pressure (see [2] for full details).

The big advantage of this coupling procedure is that little memory is required to solve the corresponding linear system and that there are no errors in the boundary conditions. For the specific case of pressure, the problem is just an Initial Value Problem (IVP), and additionally in incompressible flows, only a reference for pressure is needed.

4 DISCRETIZATION PROCEDURE

The numerical procedure involves three different problems, the solution of the momentum, continuity and energy equation in the fluid:
\[
\frac{\partial \vec{v}}{\partial t} + \rho (\vec{v} \cdot \nabla) \vec{v} = -\rho \vec{g} - \nabla p + \mu \nabla^2 \vec{v} \\
\nabla \cdot \vec{v} = 0
\] (9)

\[
\frac{\partial T}{\partial t} + (\vec{v} \cdot \nabla) T = \alpha \nabla^2 T
\] (10)

In the solid, the only equation that has to be solved is the heat conduction:

\[
\frac{\partial T}{\partial t} = \alpha \nabla^2 T
\]

For the sake of simplicity, the general details of the discretization procedure are:

- Use a segregated grid arrangement for velocity components and pressure
- Locate the temperature in the same point as the pressure (fluid and solid)
- Discretize the diffusion term with second order finite differences
- Discretize the convection term with the Osher flux limiting scheme
- Use RBF to interpolate velocity components (needed in the flux limiting scheme)
- Use second order central differencing for pressure gradient
- Use second order central differencing for continuity equation
- Solve the resulting system (7) with the procedure explained in [2]

5 PROCEDURE FOR DIRECT NUMERICAL SIMULATION

5.1 Time integration scheme

The fundamental part of any DNS simulation is the right choice of the time integration scheme. Since the time scale for both (9) and (10) will be the same, the system of equations as a whole is a Differential Algebraic Equation, with some differential equations in time and others not (continuity).

Explicit schemes as Runge-Kutta and the classical predictor corrector Adams Bashford / Adams Moulton cannot be applied since there is way to compute the time derivative for pressure.

For this kind of differential systems, the only possible way is to use an implicit procedure. Here, the Adams Moulton is not suitable due to its limited stability region. Others such as Implicit Runge-Kutta (IRK) are hard to implement since their tableau is not diagonal dominant, fundamental condition for convergence of the linear system that has to be solved in each iteration.

The only suitable method is the so-called Backward-Differentiation Formulas (BDF) or implicit multistep, with a large stability region but computationally costly. Since in DNS, the inflow is going to be perturbed permanently (to produce turbulence), BDF of orders higher than 1 become unstable in certain cases and/or conditions and it is the procedure implemented here. Commercial packages such as Fluent, OpenFoam and Kiva report similar issues.
5.2 Inflow condition

The critical in any DNS simulation is the creation of perturbations in the inflow. The highly recommended Precursor Simulation has limited applicability because a periodic boundary condition must be implemented with the pseudo-spectral method as the only possible choice for solution, limited to cartesian coordinates only.

There are several other techniques used for the creation of inflow conditions but the main drawback is that a behavior is assumed in advance.

In order to avoid the inclusion of any predefined behavior, the white-noise technique will be used in this work. Since it is important to have perturbations with good statistical quality, the Mersenne Twister algorithm will be used [12]. This random number generator is very fast and it has a period of $2^{1937} - 1$, which is excellent for statistical studies.

The study of inflow perturbations on generation of turbulence will be performed with the following algorithm, trying to keep a strict control on the percentage of fluctuation. At any time step, the new inflow velocity will be computed by:

- Set an inflow velocity with the classical power profile to get a realistic boundary layer:
  \[
  \frac{u_{\text{inflow}}}{U_{\text{bulk}}} = \left[1 - \left(\frac{|s|}{100}\right)^2\right], \quad -1 \leq s \leq +1
  \]  
  \[(11)\]

- Select a set of random numbers (between -1 and +1) for every point in the inflow.

- Compute a correction factor $f_s$ so the average of the fluctuations is 1:
  \[
  f_s = \frac{A}{\int_A \sqrt{\left|\mathbf{x}\right|^2} dA}
  \]  
  \[(12)\]

- With $p_f$ the percentage of fluctuation, compute the velocity fluctuation of each point:
  \[
  v' = f_s p_f u_{\text{inflow}}
  \]  
  \[(13)\]

- Compute velocity correction $\delta v$ to enforce mass continuity:
  \[
  \delta v = V_{\text{bulk}} - \frac{1}{A} \int_A \left(1 + f_s p_f u_{\text{inflow}}\right) dA
  \]  
  \[(14)\]

Finally, and in order to avoid high frequency oscillations, the inflow velocity already describes will be damped with respect to the previous time step, as used in other procedures such as OpenFoam.

5.3 Outflow condition

The outflow condition is also important to avoid the propagation of perturbations back to the flow. In this work, the convective outflow condition will be used:

\[
\frac{\partial \phi}{\partial t} + V_n \frac{d\phi}{dn} = 0
\]  
\[(15)\]
with $V_n$ the average normal velocity of the corresponding outflow section and $\phi$ any scalar quantity, temperature and velocity components.

### 5.4 DNS parameters

The grid was built with 2 Kolmogorov deltas in the vertical direction and 12 Kolmogorov deltas in the direction of the flow for a Reynolds number of 5,000. The length of the expansion is 10 hydraulic diameters which is enough to hold completely the recirculation zone. The number of pressure points is about 166,000 for the expansion of 1.10 and over 200,000 for the expansion of 1.30. The delta time was small enough to keep the Courant number less than 1. At the inflow, a fluctuation factor of 20% was used and the damping factor was of 0.5.

### 6 BOUNDARY CONDITIONS

Additionally to the conditions for the DNS procedure itself, at the walls, no-slip condition is applied for velocity. As it was already explained, pressure needs no boundary conditions since is computed as an IVP, with one point fixed at a reference pressure.

For the thermal energy equation in the fluid, water at a dimensionless temperature of zero enters in the region and it is being heated by the step. The upper wall remains isolated and the lower wall that does not belong to the step remains isolated too.

The block, made of 304 carbon-steel has its left wall at a fixed dimensionless temperature of 1 degree and the lower wall remains isolated.

Finally, at the interphase that connects the block and the fluid, an energy balance is applied to compute the corresponding temperature:

$$
-k_S \frac{dT_S}{dn} = -k_F \frac{dT_F}{dn}
$$

Expressing the derivative of both fluid and solid, in terms of finite differences, the temperature of the interphase is:

$$
T_i = \frac{-k_S \sum_{ab} a^S_{ab} T^S_{ab} - k_F \sum_{ab} a^F_{ab} T^F_{ab}}{k_S a^S_p + k_F a^F_p}
$$

### 7 RESULTS AND DISCUSSION

Figure 1 shows the re-attachment length in terms for the expansion ratios of 1.1, 1.2 and 1.3 respectively. As expected, the reattachment is a linear function with a very small grow rate. This behavior is consistent with the experimental data published by Adams [13], in which a slight under prediction can be observed.

Figures 2 and 3 show the temperature distribution in the step. It is observed that as the aspect ratio increases, temperature gradients in the step become sharper due mainly to the increment in the recirculation of the flow.
Figure 1: Re-attachment for different expansion ratios

Figure 2: Temperature distribution in the step for an expansion ratio of 1.10
Figure 3: Temperature distribution in the step for an expansion ratio of 1.30

Figure 4: Nusselt number in the back wall for the expansion ratio of 1.10
Figure 4 shows the local Nusselt number in the back wall of the step. In the zone close to the lower wall, the gradient is small and becomes higher close to the corner (as expected). Even that there is no data to make a straight comparison, the shape and magnitude of the Nusselt number is consistent and in the same order of magnitude as it can be observed in references [3-10].

8 CONCLUSIONS

- A meshless procedure for DNS and LES simulations has been proposed.
- The re-attachment obtained for the expansion ratios from 1.10 to 1.30 shows good agreement with experimental data.
- The behavior of the Nusselt number in the back wall is consistent with the one observed in previous publications.

REFERENCES

ADVANCED MATERIAL AND STRUCTURAL BEHAVIOR IN INNOVATIVE FORMING PROCESSES – APPLICATION TO PROCESS-INTEGRATED RING ROLLING

R. Kebriaei, S. Reese∗

∗Institute of Applied Mechanics
RWTH Aachen University
Mies-van-der-Rohe-Str. 1, D-52074 Aachen, Germany
e-mail: {reza.kebriaei,stefanie.reese}@rwth-aachen.de, http://www.ifam.rwth-aachen.de

Key words: numerical modeling, ring rolling, powder coating, phase transformation

Abstract. The increasing demand for flexibility and economy in production has a significant impact on production methodologies. In the last decades, several innovative manufacturing processes were developed to meet these requirements. To carry out such a redesign in an efficient manner (which means by computational analysis), new computational models and methods are needed.

Process-integrated powder coating is a new kind of ring-rolling process. It takes advantage of the high temperatures and high forces of the ring rolling process. This is not only to increase the ring’s diameter, but also to integrate powder metallurgical multi-functional coatings within the same process. To improve the feasibility assessment of the proposed geometries and material combinations as well as to investigate important quantities such as e.g. the stress state in the rolling gap and the residual porosity of the powder metallurgically produced layer, the versatile application of the finite element method (FEM) is crucial. Therefore, a parameterized 3D finite element (FE) model is developed on the basis of a finite strain viscoplastic material formulation.

In order to increase the strength and wear properties of relevant steels, an appropriate heat treatment should be carried out. Therefore, an implicit numerical scheme is applied to investigate the thermomechanical-metallurgical response of multi-phase steel during phase transformation. The paper is concluded by a detailed description of the process simulation and a comparison of its results with experimental data.

1 INTRODUCTION

Ring rolling represents an incremental forming process which is used to manufacture precisely dimensioned seamless rings. Its first scientific developments were made in the 20th century (see [6]). Typical applications can be found in aerospace, automotive and
railroad industries [9], e.g. rings for railway wheels and tires [19]. In many applications, it is advantageous to equip the rolled ring with a wear resistant smart functional layer (see [4]). Examples are the rolls in crushing and briquetting mills used in mineral industries. There are several techniques (as e.g. thermal spraying, buildup welding, hot isostatic pressing) available to manufacture these coatings. The disadvantages of these methods are discussed in [10].

In process-integrated powder coating, the integration of the compaction process into the rolling stage is thought to break the limitations of alternative coating processes. The manufactured products can have a diameter up to 12 m and a height up to 2.8 m. Although this novel process is reasonably efficient with respect to energy, process time and costs, there exists some difficulties. The encapsulation has to maintain vacuum conditions and conventional rolling strategies are not applicable. To support the design of this new process and to predict the influence of several geometry and process parameters on the residual porosity in the layer, a parameterized FE model is developed on the basis of a finite strain viscoplastic material formulation.

After the rolling process, in order to increase the tensile strength and the wear properties of investigated steels, thermal treatment is initiated (see [7]). Since process temperatures in hot rolling are within the range of austenitizing temperatures for the investigated steels, controlled cooling can be conducted directly from process heat subsequent to the deformation process. A main factor of this procedure is cooling speed [1]. On the one hand, low-speed cooling leads to formation of a perlitic microstructure which has poorer mechanical properties than a martensitic or bainitic microstructure [14]. Furthermore, it cannot be transformed into those. On the other hand, fast-speed cooling leads to the formation of martensitic microstructure which comes along with a change in the crystal lattice causing a volume increase of the material. In this case, the subsequent residual stresses can be high enough to crack the coating. Therefore, the temperature should be controlled in such a way that the temperature range for transformation remains below the perlitic range (500-800 °C), but above the martensitic range (less than 200 °C) to enter the bainitic range before cooling to room temperature. Therefore, the paper deals with the numerical modeling of the heat treatment. This study concerns the investigation of transformation induced plasticity (TRIP) under various loading conditions. The applied tests enable us to modify Leblond’s transformation plasticity model [13]. Furthermore, a suitable implicit numerical scheme is implemented into a finite element code to investigate the thermo-mechanical-metallurgical response of multi-phase steel during the phase transformation also numerically.

The paper is structured as follows. In the next section the principles of the process-integrated powder coating are shown. Afterwards, the material model which describes the phase transformation will be discussed. The last section is devoted to the description of the process simulation and a comparison of its results with experimental data. The paper closes with some concluding remarks.
2 PROCESS-INTEGRATED POWDER COATING

The principle of the rolling process is sketched in Figure 1a. The mandrel pushes the ring towards the main roll which is driven by angular velocity. Friction between the ring and the main roll as well as between the ring and the mandrel lead to a rotation of the ring. By decreasing the radial rolling gap the ring grows in tangential and in axial direction. In the opposing axial rolling gap the height of the ring is controlled and reduced by the axial rolls.

![Figure 1: Process-integrated powder coating: (a) principle of ring rolling [17], (b) sectional view.](image)

The setup of the new process is depicted in Figure 1b. Here, a sheet metal is welded circumferentially around the outside of an unrolled ring blank. Powder layer material (metal matrix composite, MMC) is placed inside the resulting chamber.

3 CONSTITUTIVE MODELING

The model used to describe the compressible layer material is based on a finite strain elasto-plastic material formulation. Since the process takes place at high temperatures rate dependence is taken into account. A comprehensive description of the material model can be found in [3]. Here, we discuss the integration of heat treatment of the rolled ring into the subsequent cooling process which goes along with phase transformations.

3.1 TRANSFORMATION KINETICS

In this study, we use the modified version of the Johnson-Mehl-Avrami-Kolmogorov (JMAK) [8] equation which comprises two sequences of bainitic transformations

\[
p = \tilde{p}_1 \left[ 1 - e^{-\left(\frac{t}{\tau_1}\right)^{n_1}} \right] + \tilde{p}_2 \left[ 1 - e^{-\left(\frac{t}{\tau_2}\right)^{n_2}} \right]
\]

The parameters \(\tilde{p}_1\) and \(\tilde{p}_2\) describe the thermodynamic equilibrium fraction, which can be determined from the equilibrium phase diagram at a given temperature. The quantities \(\tau_1\) and \(\tau_2\) are the times needed to reach a perfect transformation, \(n_1\) and \(n_2\) are the Avrami
exponents. These parameters are determined by appropriate experimental investigations. The transformation rate \( \dot{p} \) is given by the first derivative of Eq.1 with respect to time.

\[
\dot{p} = n_1 \frac{\bar{p}_1 - p}{\tau_1} \left[ \ln \left( \frac{\bar{p}_1}{\bar{p}_1 - p} \right) \right]^{n_1-1} + n_2 \frac{\bar{p}_2 - p}{\tau_2} \left[ \ln \left( \frac{\bar{p}_2}{\bar{p}_2 - p} \right) \right]^{n_2-1}
\]

(2)

The JMAK equation is proposed for an isothermal condition. However, it is not directly applicable for the nonisothermal case. Therefore, we apply the isokinetic relationships of nonisothermal systems (see [18, 16]). Accordingly, the cooling curve is divided into a number of small time steps, and the amount of transformation at each time step is calculated using the JMAK equation. Finally, the martensitic transformation is taken account by using the general form of Koistinen-Marburger [12] equation as

\[
p = \bar{p}_m [1 - e^{-\left( \frac{M_s - T}{b} \right)^n}]
\]

(3)

Herein, \( \bar{p}_m \) denotes the volume fraction of the retained austenite. \( M_s \) designates the temperature where martensitic transformation starts. The quantities \( b \) and \( n \) are material constants which should be justified by experimental results.

### 3.1.1 TRANSFORMATION PLASTICITY

The plastic behaviour of steels during the transformations can be separated into two parts. The first is classical plasticity due to plastic flow from variations of the applied stress and the temperature (see [5]). Secondly, transformation plasticity arises due to plastic flow from variations of the phase proportions (see [2]).

Following this, the total strain

\[
\varepsilon = \varepsilon_{el} + \varepsilon_{th} + \varepsilon_p
\]

(4)

splits into elastic, thermal, and plastic parts, respectively. Each phase has different thermal properties. In order to take this fact and the rate of volumetric change due to the phase transformation into account, the rate of the thermal strain is defined by a linear mixture rule as:

\[
\dot{\varepsilon}_{th} = \sum_{i=1}^{n_p} [\dot{p}_i \alpha_i (T - T_0) + p_i \alpha_i \dot{T}] \mathbf{I}
\]

(5)

where, \( p_i \) is the volume fraction of phase \( i \). Additionally, \( \alpha_i \) denotes the thermal expansion coefficient, \( n_p \) is the total number of the involved phases, \( T \) and \( T_0 \) designate the current and initial temperature, and \( \mathbf{I} \) is the identity tensor. Furthermore, the material time derivative is shown by \( (\bullet) = d\bullet / dt \), the subscripts \( s \) and \( h \) denote the soft and hard phase and \( \Delta(\bullet)_{s \rightarrow h} \) designates the variation of \( \bullet \) from softer to harder phase.

The rate of the plastic strain

\[
\dot{\varepsilon}_p = \dot{\varepsilon}_{cp} + \dot{\varepsilon}_{tp}
\]

(6)

\[\text{142}\]
is decomposed into classical and transformation plasticity. Consequently, the rate of classical microscopic plastic strain

\[ \dot{\varepsilon}_{cp} = \dot{\varepsilon}_{\dot{\sigma}} + \dot{\varepsilon}_{\dot{T}} \]  

splits into two parts: the plastic strain induced from the deformation of the soft phase,

\[ \dot{\varepsilon}_{\dot{\sigma}} = \frac{3(1 - p_h)}{2\sigma_y} \frac{l(p_h)}{E} \sigma^D \dot{\sigma}_{eq} \]  

and the rate of the plastic strain induced from the variation of temperature:

\[ \dot{\varepsilon}_{\dot{T}} = \frac{3\Delta \alpha_{s\rightarrow h}}{\sigma_y} p_h \ln(p_h) \sigma^D \dot{T} \]  

In order to include the low values of volume fraction, the functions \( l(p_h) \) and \( k(p_h) \) are defined by [13] (see Table 1).

<table>
<thead>
<tr>
<th>p</th>
<th>0.00</th>
<th>0.125</th>
<th>0.250</th>
<th>0.500</th>
<th>0.750</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>k(p)</td>
<td>0.00</td>
<td>0.440</td>
<td>0.124</td>
<td>0.391</td>
<td>0.668</td>
<td>1.00</td>
</tr>
<tr>
<td>l(p)</td>
<td>0.00</td>
<td>2.53</td>
<td>4.00</td>
<td>2.76</td>
<td>1.33</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The accumulated plastic strain during the bainitic transformation in the austenitic phase is given by [15]

\[ \dot{\varepsilon}_{\dot{\sigma}} = \frac{2\Delta \varepsilon_{s\rightarrow h}^h}{1 - p_h} l(p_h) \frac{\sigma_{eq}}{\sigma_y} \ln(p_h) \dot{p}_h + \frac{l(p_h)}{E} \dot{\sigma}_{eq} - \frac{2\Delta \alpha_{s\rightarrow h}}{1 - p_h} p_h \ln(p_h) \sigma^D \dot{T} \]  

where \( \Delta \alpha_{s\rightarrow h} \) is the difference of the thermal expansion coefficients in the soft and the hard phase and E is Young’s modulus. The bainitic phase inherits the strain hardening of austenite which is defined by

\[ \dot{\varepsilon}_{\dot{\sigma}} = \frac{\dot{p}_h}{p_h} \varepsilon_{\dot{\sigma}} + \theta \frac{\dot{p}_h}{p_h} \varepsilon_{\dot{\sigma}} \]  

Herein, \( \theta \) is a parameter depending on the transformation, it is changed between 0 (without transformation) and 1 (perfect transformation). Additionally, we assume that during diffusional transformation, no strain hardening is included and \( \theta = 0 \). Finally, the stress tensor is obtained by

\[ \sigma = C[\varepsilon - \varepsilon_{th} - \varepsilon_p] \]  

Under the global yielding, there is no difference between classical plasticity and transformation plasticity. In this case, all involved phases are influenced by plastic deformation [11]. Therefore, conventional elastoplasticity can be applied.

An implicit numerical solution algorithm to calculate the plastic deformation of each phase is implemented into ABAQUS via an user material subroutine UMAT.
4 RESULTS

4.1 DENSIFICATION OF THE LAYER

To guarantee the functionality of the coating it is important to reach a high state of compaction inside the layer. Therefore, parameter studies are carried out. The powder material which is applied in the layer is a hot work tool steel 56NiCrMoV7 (see Table 2). Additionally, for the substrate (encapsulation) a cold work tool steel X220CrVMo13-4 is used.

Table 2: Chemical combination of the applied materials.

<table>
<thead>
<tr>
<th>Material</th>
<th>Chemical composition</th>
<th>Material</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>Cr</td>
</tr>
<tr>
<td>56NiCrMoV7</td>
<td>0.55</td>
<td>1.10</td>
</tr>
<tr>
<td>X220CrVMo13-4</td>
<td>2.30</td>
<td>12.50</td>
</tr>
</tbody>
</table>

Consequently, different radii for the main roll and the mandrel are chosen (see Table 3). The ratio of the main roll radius (RMR) with respect to the mandrel radius (RMA) is defined as $k$.

Figure 2: Relative density of compacted layer (hot work steel) after 15 seconds rolling: (a) $k = 1$, (b) $k = 2.7$, (c) $k = 3.5$, (d) $k = 7$.

Figure 2 (a)-(d) demonstrate the influence of $k$ on the relative density. It can be seen that the increase of $k$ leads to a small decrease of the relative density. This result can
Table 3: Variation of geometric parameters.

<table>
<thead>
<tr>
<th>k</th>
<th>Main roll</th>
<th>Mandrel</th>
<th>Ring</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>130</td>
<td>130</td>
<td>218.5</td>
</tr>
<tr>
<td>2.7</td>
<td>175</td>
<td>65</td>
<td>141</td>
</tr>
<tr>
<td>3.5</td>
<td>350</td>
<td>100</td>
<td>218</td>
</tr>
<tr>
<td>7</td>
<td>455</td>
<td>65</td>
<td>141</td>
</tr>
</tbody>
</table>

Figure 3: Influence of geometrical parameters on relative density: (a) relative density with respect to the time, (b) relative density with respect to \( k \).

be observed in Figure 3 (a). Figure 3 (b) shows the predicted influence of \( k \) on the relative density at the end of the rolling process. These results emphasize that with \( k = 1 \) the highest values for the relative density can be obtained. Working with \( k = 1 \) might not be realistic for a practical application. But it confirms the strategy to improve the compaction behavior by increasing the pass reduction at the main roll.

4.2 HEAT TREATMENT

For the material model validation several tests are done. In the first step, in order to get rid of the influence of any temperature gradient, isothermal annealing is done. Therefore, the steel is heated up to above the upper critical temperature (1050 °C) and this temperature is maintained for 10 minutes.

Then the temperature is cooled down below the lower critical temperature. Finally, it is cooled to the room temperature. As an example, the results related to the uniaxial tension tests are demonstrated (see Figure 4(a)-(b)). At the lower temperature we obtain a small deviation between experimental and modeling results for both materials. This difference becomes even smaller at higher temperatures. Additionally, we notice a decrease of the yield stress from lower to higher temperature. The main goal is to reach a bainitic microstructure formation. Therefore, the bainitic transformation is investigated. Figure 5(a)-(b) presents the percentage of austenite transformed to bainite. It can be seen that for the hot work steel at the temperature of 275 °C about 96% of austenite is transformed.
Figure 4: Temperature dependency of stress-strain curves: (a) hot work steel, (b) cold work steel.

Figure 5: Volume fraction of bainite: (a) hot work steel, (b) cold work steel.

to bainite which is quite close to a perfect transformation. However, the cold work steel shows a weaker transformation behavior. This is due to the fact that the recovery and the recrystallization of the grains reduce the dislocation density, which has a direct effect on the transformation. Consequently, we see a longer transformation time and weaker transformation behavior.

To see the influence of TRIP, the variation of the length of the specimen which is submitted to several compressive loads is investigated under isothermal conditions at the temperature of 375 °C (see Figure 6(a)).

By increasing the load we see an extra deformation. This extra deformation is due neither to elasticity nor to viscous effects. This is another plastic effect, which is referred to as TRIP. Finally, the validated material model is applied to estimate the bainitic volume fraction in the rolled ring.

The experimental conditions are high pressure water from the inner side and heat preservation from the other side of the ring as depicted in Figure 6(b). These conditions are also considered in the simulation. In comparison with the outer side, the inner side of
the ring has a lower temperature (see Figure 7(a)). This is due to the direct contact of the inner surface with high pressure water. Furthermore, the heat preservation from the outer side which assures the required transformation, maintains the higher temperature from the outer surface of the ring. The experimental investigations confirm that the required transformation time for the hot and cold work steel are about 5000 and 72000 seconds, respectively. Furthermore, the cooling rate is chosen as -0.2 K/s for the hot work steel and -0.3 K/s for the cold one. In order to have a better insight, we first study a solid ring. Figure 7(a)-(b) illustrates the corresponding temperature fields and bainitic volume fractions. It can be seen that the transformation is nearly perfect (see Figure 7(b)). In other words, the remaining volume fraction of martensite is about 5 per cent. This is due to a very fast cooling of the inner surface with high pressure water at the beginning of the cooling process. Finally the composite ring which consists of hot work steel (substrate) and cold work steel (layer) is investigated. The distribution of temperature over the ring is demonstrated in Figure 8(a). Again, the inner surface of the ring has a lower temperature.
and the outer surface has a higher temperature. However, since both materials have different thermo-mechanical material parameters, the distribution of the temperature is different than in the solid ring. Figure 8(b) demonstrates the volume fraction of bainite. A perfect bainitic transformation is observed in the substrate and the sheet metal. In contrast to the martensitic transformations, there is not a significant change in the volume of the substrate. This avoids the creation of cracks in the coating which is proved by experimental results. Moreover, the quantitative prediction of the crack in the coating can be carried out by appropriate damage modeling which is our future task.

5 CONCLUSIONS

In this paper a parameterized FE ring rolling model was presented that is applicable to the simulation of process-integrated powder coating by radial axial rolling of rings. Furthermore, we developed a material model which is originally introduced by [13]. Additionally, the modified version of the JMAK equation is used to describe the bainitic transformation. Moreover, mathematical modeling of transformation plasticity [14] in steels which is coupled with strain hardening phenomena is modified.

As a result, a coupled thermomechanical-metallurgical numerical model is developed which enabled us to see the response of generalized multi-phase steel during phase trans-
formations from austenite to bainite and martensite. An implicit numerical scheme is applied to investigate the thermomechanical response of multi phase steel during the phase transformation. Furthermore, the applied model is able to predict transformation-induced plasticity (TRIP) under various loading conditions in the different cases of transformations. The material model is validated by several tests and the results are compared with experimental data. Consequently, the validated material model is used to simulate the transformation in composite ring.

6 ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of this work by the Volkswagen Foundation under reference number I/81 247 Ak within the initiative ‘Innovative Methods for the Manufacturing of Multifunctional surfaces’. All the experimental results presented in this paper have been performed by the Chair of Materials Technology and the Chair of Production systems at the Ruhr University of Bochum (Germany). Therefore, we would like to thank Dr.-Ing. Heiko Moll (Chair of Materials Technology), Dr.-Ing. Axel Hoeffter, Dipl.-Ing. Tobias Husmann for providing the experimental data. Moreover, we would like to express our gratitude to Dipl.-Ing. Jan Frischkorn for his supports and scientific discussions.

REFERENCES


A NEW FRAME FOR CONSTITUTIVE MODELING OF VISCOELASTIC POLYMERS; ACCOUNTING THERMAL EFFECTS AND STRAIN INDUCED CRYSTALLIZATION

N. BILLON * AND J. BIKARD †

* Center for Materials Forming (CEMEF)
Mines-ParisTech, UMR 7635
BP 207 06904 Sophia Antipolis, France
e-mail: noelle.billon@mines-paristech.fr

† Technyl Innovation Center
R&D Solvay Engineering Plastics
Lab. New Tech – CRTL, 85 rue des Frères Perret BP 62 69192 Saint FONS, France
e-mail: Jerome.bikard@eu.rhodia.com

Key words: Time-dependent behaviour, Thermo mechanical coupling, Large-strain, Polymers, Inelasticity, Strain induced crystallization.

Abstract. Modelling the mechanical behaviour of polymers is a nontrivial task. Usual macroscopic approaches that decompose global deformation into three elementary components in a “a priori” manner often results in complex and “of limited efficiency” models. This study deals at promoting another concept of visco-hyper-elasticity for polymers close to Tg without arbitrary decomposition into “viscous” and “elastic” stresses or strains. To achieve that point, a hyper elastic model is extended to account for inelastic processes. Those latter are assumed to result in a kinetics of variation of internal variables that have to be accounted for in the energy balance at any time and that induce time effects in the writing. Variables of interest are related to de entanglement and/or strain induced crystallisation through specific kinetics laws. This version uses Edward-Vilgis’ network theory that depends on four physical parameters: the density of fixed network nodes, the density of sliding nodes or entanglements, a parameter ultimately related to chain extensibility and a parameter ultimately representative for level of freedom of entanglements. The stress is written in the framework of Irreversible Processes Thermodynamics (IPT) and in the frame of large strain approximations. Mechanical problem is coupled to thermal problem using Taylor-Quinney coefficient β. Equations are included in a finite difference code (using a θ-method) to calculate temperature and stress through the central section of a sample. Parameters identification is based on the minimisation of a two objectives cost function that accounted for average axial stress in the section and for surface temperature at this section. This latter was written in a mean-square approach.
1 INTRODUCTION

Constitutive models for polymers have to combine non linear processes that can be elastic or inelastic, highly sensitive to temperature and to strain-rate and that are, sometimes, ruled by discrete relaxations.

Usual macroscopic approaches that arbitrary decompose global deformation into three elementary components (reversible-instantaneous (elastic part), irreversible-not dependent upon strain-rate (plastic part) and irreversible-dependent upon strain-rate (viscous part), respectively) was extensively applied in that field but generally led to models of limited validity.

An exhaustive review would be a tedious task. Let’s only say that mathematical writings were built up, either on the basis of 1D-model, further extended to 3D case, or in the continuum thermodynamics frame, in which global behaviour is ruled by energy potentials and dissipation pseudo-potentials. Whatever the route is and despite of some mathematical differences, basic assumptions were equivalent in the most rigorous of those studies. Visco elasticity was considered combining elastic and viscous elementary elements in more or less sophisticated manners. Nonlinearity was addressed at two levels:

- Important strain-hardening in polymers (sometimes related to strain induced crystallization) was often reproduced introducing various hyper-elastic reversible elements.
- Nonlinear viscous effects were introduced through “non Newtonian pseudo plastic, Eyring or Carreau–like elements” or through phenomenological accountings for disentanglement. The dependence upon strain-rate and temperature was “post-introduced” through the dependence of given variables. Those approaches resulted in numerous parameters that are not that easy to identify without specific protocols and without pre-allocating one element or another to one macroscopic phenomenon.

The present study deals at exploring and promoting another concept hoping a reduction in the number of parameters. Starting point was the fact that it is possible to combine the general rubber hyper elastic theory with some evolution of internal variables, potentially induced by microstructure alteration, to model time effects in the constitutive model. This route was already suggested in a totally independent manner in the past. More precisely, our general goal was to model visco-hyper-elasticity and visco-plasticity of polymers above Tg without arbitrary a priori decomposition into “viscous” and “elastic” stresses or strains. Therefore, a hyper elastic model was extended to account for inelastic processes within the frame of continuum thermodynamics. Dissipative processes were assumed to result of a kinetics of variation of internal variables that had to be accounted for in the energy balance at any time and induce time effects in the writing.

Model was validated in its 1-D form using experimental observations on different polymers. However, it was developed in a much more general 3D form, consistent with the large strain mechanical approach and with continuum thermodynamics.
Present paper intends to remind fundamental aspects of this approach and to illustrate its efficiency for various materials (amorphous, semi crystalline or experiencing stain induced crystallization).

2 THEORETICAL

2.1 Constitutive model [1]

Main hypothesis is that any polymer could be modelled with an equivalent entangled network, at least above or close to $T_g$. A hyper elastic strain energy density, $w$, could then be defined that rules the behaviour of the polymer. $w$-potential depends on temperature, $T$, and on some variables that reflect the microstructure of the material ($m_1$, $m_2$, ..., $m_n$) as well as on elastic extension ratios, $\lambda_i^e$, or on associated invariants.

During loading, microstructure of the material is altered (disentanglement, crystallisation or crystalline reorganisation), which results in evolution of associated $m$-variables. This represents a potential change in the energy of the material that has to be compensated for thermodynamics principles to be obeyed. Therefore, some local elastic recovery in the network, $\partial\lambda_i^{re}/\partial t$, have to occur. This is the source of inelastic strain-rate, $\partial\lambda_i^{re}/\partial t = -\partial\lambda_i^{re}/\partial t$, in the model, i.e. inelastic processes are the fraction of elastic extension that has to be released to compensate change in energy due to microstructural processes. The concept is summarised in Eq. (1):

$$\sum_{j=1}^{n} \frac{\partial w}{\partial m_j} \frac{\partial m_j}{\partial t} = -\frac{1}{\beta'} \sum_{i=1}^{3} \frac{\partial w}{\partial \lambda_i^e} \frac{\partial \lambda_i^e}{\partial t} = 1 \sum_{i=1}^{3} \frac{\partial w}{\partial \lambda_i^a} \left( \frac{\lambda_i^e \partial \lambda_i^a}{\lambda_i^e \partial t} \right)$$

with $\beta' \geq 1$

where $\partial m_j/\partial t$ is the kinetics of the evolution of $m_j$-variable. At this point principal elastic extension (e indexes) and principal inelastic extension (a indexes) are assumed to be collinear. $\beta'$-parameter, express the fact that part of the released energy could be dissipated into heat (thermo mechanical coupling) or stored via some permanent changes in the microstructure (plastic-like phenomenon).

Edwards-Vilgis’ model was chosen as primary network (Eq. (2)) as its efficiency was intensively demonstrated. According to this approach four constitutive parameters are defined. Chains are assumed to have a finite extensibility, which is ultimately controlled by one positive parameter $\alpha$ (0 in the case of a free Gaussian chain). Chains are linked by permanent nodes of density $N_c$ per unit volume and slip links (entanglements) of density $N_s$ per unit volume. Finally, a positive “slipperiness factor”, $\eta$, is ultimately defined and related to the degree of mobility of slip links. A zero $\eta$-value corresponds to permanent nodes. Initial $\eta$-value was chosen close to 0.2343 following. So elastic phenomenon are modelled through Eq. (2):
\[ w = \frac{N_s}{2} \left( \frac{\Lambda_{\alpha\eta}}{X_{\alpha\omega} Y_{\eta\lambda}} Z_{\eta\lambda} + \ln(Y_{\eta\lambda}) + \ln(Y_{\alpha\alpha}) \right) + \frac{N_c}{2} \left( \frac{1 - \alpha^2}{X_{\alpha\alpha}} I_1^e + \ln(X_{\alpha\alpha}) \right) \]

\[ Y_{\eta\lambda} = 1 + \eta I_1^e + \eta^2 I_2^e + \eta^3 X_{\alpha\alpha} = 1 - \alpha^2 I_1^e ; Z_{\eta\lambda} = I_1^e + 2\eta I_2^e + 3\eta^2 \; \Lambda_{\alpha\eta} = (1 + \eta)(1 - \alpha^2) \]

\[ I_1^e = \lambda_1^e \lambda_2^e + \lambda_3^e \lambda_2^e + \lambda_3^e \lambda_3^e ; I_2^e = \lambda_1^e \lambda_2^e + \lambda_1^e \lambda_3^e + \lambda_2^e \lambda_3^e ; I_3^e = \lambda_1^2 \lambda_2^2 \lambda_3^2 = 1 \]

where \( N_s \) and \( N_c \) are \( N_s kT \) and \( N_c kT \), respectively, with \( k \) the Boltzmann’s constant and \( T \), the absolute temperature.

Time dependent effects were introduced in this model through the alteration of parameters, in example an increase in \( \eta \) (disentanglement). Rate of changes were related to the energy that is available in the material in the sense that the more stressed the slip-links the faster the processes (Eq. (3)):

\[ \frac{\partial \eta}{\partial t} = \mathcal{Z}(f_s) > 0 \]

\[ f_s = \left( \frac{\Lambda_{\alpha\eta} X_{\alpha\omega} Y_{\eta\lambda}}{\eta_{\eta\lambda}} + \ln(Y_{\eta\lambda}) + \ln(Y_{\alpha\alpha}) \right) - 3 \left( \frac{1 - \alpha^2}{1 - 3\alpha^2} + \ln(1 + \eta) \right) + \ln(1 - 3\alpha^2) \]

\[ \mathcal{Z} = \xi (\exp(f_s - \psi) - 1) \text{ if } f_s > \psi, \mathcal{Z} = 0 \text{ if } f_s \leq \psi \]

According to the framework of Irreversible Processes Thermodynamics (IPT) and in the frame of large strain approximations the stress is given in Eq.(4) / Eq. (5):

\[ \sigma = 2F_e \left( \frac{\partial \omega}{\partial I_i^e} \frac{\partial I_i^e}{\partial C_e} + \frac{\partial \omega}{\partial I_2^e} \frac{\partial I_2^e}{\partial C_e} + \frac{\partial \omega}{\partial I_3^e} \frac{\partial I_3^e}{\partial C_e} \right) F_e^{-T} - p' I \]

where \( p' \) is an arbitrary pressure due to incompressibility (compressible model will be further presented), \( F_e \) is the elastic extension tensor and \( C_e \) is the Cauchy-Green stress tensor;

Parameters can depend on temperature and strain-rate but it was possible to combine those two dependences in a unique dependence upon equivalent strain-rate at reference temperature, \( T_{ref} \), as classically defined in the usual WLF’s approach [1]. To summarise parameters depends on the fictitious strain-rate:

\[ \dot{\varepsilon} = a_{T/T_{ref}} \dot{\varepsilon}_{eq} = a_{T/T_{ref}} \sqrt{\frac{2}{3}} D : D \]

where \( D \) is the total strain-rate tensor. \( a_{T/T_{ref}} \) is the so-called shift factor:
\[
\log \frac{d_T}{d_{T_{\text{ref}}}} = - \frac{C_1(T - T_{\text{ref}})}{C_2 + (T - T_{\text{ref}})}
\]  

(7)

\(C_1\) and \(C_2\) are material characteristics that depends on \(T_{\text{ref}}\), an arbitrary reference temperature.

2.2 Thermomechanical coupling

Thermomechanical coupling was introduced in energy equation (Eq. (8)) in a classical manner accounting for dissipation potential, \(\Phi_{\text{int}}\), related to inelastic strain:

\[
\rho C_p \frac{\partial T}{\partial t} = \lambda \Delta T + T \frac{\partial \sigma}{\partial T} : D_e + \Phi_{\text{int}}
\]

where \(D_e\) is the elastic strain-rate tensor, \(\lambda\) is the heat conductivity, \(C_p\) is the heat capacity and \(\rho\) the specific mass of the polymer.

Dissipation potential was related to disentanglement (Eq. (3)) and is rewritten as:

\[
\Phi_{\text{int}} = A^{D_v} : D_v + A^\eta \eta \geq 0
\]

(9)

\[A^{D_v} = 2C_e \frac{\partial \psi}{\partial C_e} : D_v\]

\[A^\eta = -\frac{\partial \psi}{\partial \eta} \]

where \(D_v\) is the inelastic strain-rate tensor. \(A^{D_v}\) and \(A^\eta\) are the thermodynamics forces associated to the inelastic deformation process, respectively.

This approach is consistent with Clausius-Duheim inequality and is compatible with the Generalized Standard Materials (GSM) approach. Novelty relies on the fact that the inelasticity results from the consumption of some elastic energy to promote changes in the internal variables.

Eq. (9) depicted the case of an athermal process where there is no heat dissipation. A Taylor-Quinney coefficient \(\beta\) was introduced in the intrinsic dissipation in a general manner making not any assumption concerning its value (Eq. (10)):

\[
\Phi_{\text{int}} = A^{D_v} : D_v + A^\eta \eta = \beta (D_e, \alpha_r \dot{\epsilon}_{eq}) A^D_v : D_v \geq 0
\]

(10)

where \(\beta(A^{D_v} : D_v)\) is the energy transformed into heat. Two extreme values for \(\beta\) exist:

- \(\beta\) close to 1 (but not 1): inelasticity is mainly converted into heat;
- \(\beta=0\) there is no heat dissipation, inelastic energy is stored in changes in microstructure only.

In consequence, the energy equation can be rewritten as Eq. (11):
2.3 Coupling with strain induced crystallization

Ability of some initially amorphous polymers (such as PET, PLA) to develop semi-crystalline microstructure under mechanical loading above glass transition has been intensively studied. Parallel to crystallisation a spectacular strain-hardening is observed, which an important phenomenon to reproduce in constitutive models is. This latter point was addressed through the models presented here.

A precise experimental protocol, involving exact thermo-mechanical descriptions of the deformation as well as coolings as rapid and as controlled as possible [2], allowed us to conclude that this crystallization does not obey the classical schematic of simple two steps process: nucleation and growth of the spherulites. According to these measurements crystallisation is a more progressive phenomenon with intermediate stages: orientation, lateral packing of the molecules and lamellae formation. Mechanical strain hardening is not due to the appearance of well defined crystals and crystallinity ratio as measured after cooling is not a relevant descriptor for the entities responsible for strain hardening. As a matter of fact, the increase in stress occurs during the very first stage of the crystallisation when only some traces of micelle-like entities can be detected.

It could then be concluded that the occurrence of strain hardening can be due to the extension of chains and the increase in the interaction between molecules. So a concept of locking up of an amorphous network was chosen as a relevant concept for mechanical modelling.

Therefore, within the frame of Edwards & Vilgis’ potential, it was assumed that the density of permanent node increased as a function of elastic energy in the material. Parallel to that extensibility of chains increased as a consequence of a lost of conformational freedom. Obviously, disentanglement still exists in not crystalline zones through an increase of $\eta$. Eq. (1) could then be rewritten as:

$$ \frac{\partial w}{\partial \eta} + \beta \sum_{i=1}^{3} \frac{\partial w}{\partial \alpha_i} = \frac{1}{\beta} \left( \sum_{i=1}^{3} \frac{\partial w}{\partial \lambda_i} \right) $$

with $\beta \geq 1$ (12)

Model can then be tracked as previously and allowed combining inelastic processes due to amorphous phase as well as that induced by the texturation.

3 IDENTIFICATION OF PARAMETERS

Experimental data consisted in tensile and shearing tests where local strain field and surface temperature field were measured. DIC on random painted patterns were used for the former an IR camera allowed the latter. From those measurements local strain and temperature were deduced in the necking zone allowing the extraction of true stress vs. true
strain and surface temperature vs. true strain curves that was used to identify parameters of the model.

To achieve that point the above equations were included in a finite difference code (using a θ-method) to calculate temperature and stress through central section of the sample [3]. Concerning temperature a through the thickness 1D-modelling was used to model the experiments with a convection with surrounding air boundary condition.

In addition, this numerical model was associated to an inverse analysis, based on the simplex algorithm available in MATLAB® software. The identification was based on the minimisation of a two objectives cost function that accounted for average axial stress in the section and for surface temperature at this section. This latter was written in a mean-square approach.

In consequence, parameters to be identified were: mechanical parameters \( N_c, N_s, \alpha, \eta \) and parameters controlling evolution of \( \eta, N_c \) and \( \alpha^2 \), parameter \( \beta \) and their dependence upon \( \dot{\varepsilon} \) and, finally, \( h \), the convection coefficient. The thermal conductivity, the heat capacity and the density of the material were issued from the literature and were assumed to be constant.

4 CONCLUSION

Using such approach, it is possible to reproduce both the mechanical and thermal properties of the polymer over the wide range of experimental sets using only 12 parameters.

The use of equivalent strain-rate at a reference temperature made it possible to model the behavior of various

It was also possible to reproduce the drastic strain hardening of PET under tensile test performed above Tg (Figure 1) and the change in the inelasticity regime when “crystallization” occurs.

![Figure 1 Typical uploading unloading true stress vs. true strain (Henky’s strain) at 86 °C and 0.01 s⁻¹. Dots represents experimental data, lines correspond to calculation with the model.](image-url)
REFERENCES


COMPARISON OF A FLUID AND A SOLID APPROACH FOR THE NUMERICAL SIMULATION OF FRICTION STIR WELDING WITH A NON-CYLINDRICAL PIN

P. Bussetta*, N. Dialami†, R. Boman*, M. Chiumenti†, C. Agelet de Saracibar†, M. Cervera† and J-P. Ponthot*

* University of Liège, Department of Aerospace & Mechanical Engineering, Non Linear Computational Mechanics, Building B22/3, Chemin des Chevreaux, 1; B-4000 Liège, Belgium
e-mail: P.Bussetta@ulg.ac.be; R.Boman@ulg.ac.be; J-P.Ponthot@ulg.ac.be

† International Centre for Numerical Methods in Engineering (CIMNE), Campus Norte UPC, 08034 Barcelona, Spain
e-mail: miguel.cervera@upc.edu; micleo@cinm.upc.edu; agelet@cinm.upc.edu

Key words: Friction Stir Welding (FSW), Finite Element Method, Remeshing, Arbitrary Lagrangian Eulerian (ALE) formalism

Abstract. Friction Stir Welding (FSW) process is a solid-state joining process during which materials to be joined are not melted. As a consequence, the heat-affected zone is smaller and the quality of the weld is better with respect to more classical welding processes. Because of extremely high strains in the neighborhood of the tool, classical numerical simulation techniques have to be extended in order to track the correct material deformations. The Arbitrary Lagrangian Eulerian (ALE) formulation is used to preserve a good mesh quality throughout the computation. With this formulation the mesh displacement is independent from the material displacement. Moreover, some advanced numerical techniques such as remeshing or a special computation of transition interface is needed to take into account non-cylindrical tools. During the FSW process, the behaviour of the material in the neighborhood of the tool is at the interface between solid mechanics and fluid mechanics. Consequently, a numerical model of the FSW process based on a solid formulation is compared to another one based on a fluid formulation. It is shown that these two formulations essentially deliver the same results in terms of pressures and temperatures.
1 Introduction

Friction Stir Welding (FSW) is a relatively recent welding process, which was developed at the Welding Institute (UK) in 1991. FSW is a solid-state joining process. It means that during welding the materials to be joined are not melted. The joining is constituted by mechanical intermixing of the two materials. A rotating non-consumable tool is inserted between the two workpieces and displaced along the welding direction (see figure 1). The tool is composed of two parts: a pin and a shoulder. The pin is introduced into the welded joint to mix deeply the two materials together. The aim of the shoulder is to contain the material around the pin. The part of the welding joint where the rotational velocity of the tool and the advance velocity add up is named the advancing side. The other part, where the two velocities are in opposite directions, is called the retreating side. The friction between the rotating tool and the workpieces as well as the plastic deformation in the neighborhood of the tool increase the temperature in the welded zone. But, during the process, the temperature is always smaller than the melting temperature of the materials. So, the heat-affected zone is smaller and the quality of the welding is higher than in more classical welding processes.

![Diagram of FSW process](image)

(a) General view  (b) View of the tool

Figure 1: Scheme of the FSW process.

In spite of the important number of applications of FSW, the phenomena happening during welding are still not very well understood. Therefore, the investigations on this process and especially regarding numerical simulations are still very active [1, 2, 3]. The mechanical intermixing submits the material in the neighborhood of the tool to extremely high strains. Consequently, classical numerical simulation techniques have to be extended in order to track the correct material deformations. One of these possible extended techniques is the Arbitrary Lagrangian Eulerian (ALE) formulation [4, 5, 6].
This formulation is used to control the mesh motion regardless of the real material displacements. The ALE formulation is also used to maintain a good mesh quality during the computation. For that sake, in the general case of a non-cylindrical pin, the ALE formulation has to be enhanced and advanced numerical techniques such as a remeshing operator step is needed.

This paper presents and compares two different numerical models of the FSW process. The first model is based on a solid approach written in terms of nodal positions and nodal temperatures. The second model of FSW process is based on a fluid approach written in terms of the velocity, the pressure and the temperature fields. Both models use advanced numerical techniques such as remeshing and the ALE formulation.

2 2D Numerical modelling of FSW process

The FSW process is modelled in two dimensions under the plane strain hypothesis. To model this welding process, the displacement of the tool is split into an advancing movement (actually assigned to the work-pieces but, in the opposite direction) and a rotation (imposed to the tool).

In other words, the centre of the pin is fixed and a constant velocity is imposed to the plates (see figure 2(a)). The tool is described by a classical Lagrangian mesh. Then, in relation with the distance to the centre of the tool, three zones of the plates are identified. In the closest zone around the pin, the material is submitted to extremely high strains. This region is called the thermo-mechanically affected zone (TMAZ). Due to high deformations, the use of a Lagrangian formalism would lead very quickly to mesh entanglement. Thus, in this region, the ALE formulation is employed. On top of this, the ALE formulation allows the model to take into account non-circular pin shapes. In this zone, the mesh has the same rotational speed as the pin (red region in figure 2(b)). In the furthest zone from the tool, the grey zone in figure 2(b), the Eulerian formulation is used. Thus in this region, the mesh is fixed. The ring connecting region 1 and region 3 is a transition zone (white region numbered 2 in figure 2(b)). In such a model, the quality of the mesh does not change during the simulation except in the transition zone. So, to overcome this problem, two different numerical techniques are proposed (see section 2.2.2).

2.1 Thermomechanical formulation

The numerical models presented here are based on the finite element method. The thermomechanical equations are split in a mechanical part and a thermal part. First, the mechanical equations are solved for the temperature field computed at the latest increment. Then, the thermal equations are solved on the frozen resulting geometrical configuration. In this paper, two numerical formulations are compared. The first one is based on a solid mechanics approach. It is written in terms of nodal positions and temperatures. The second one is based on a fluid mechanics approach. The equilibrium
is written as a function of nodal velocities, pressures and temperatures.

**Solid approach** In the solid approach, the finite element used are linear quadrilaterals. The position and temperature fields are computed at each node of the elements. The stresses and the internal variables are computed at each quadrature point of the element (4 Gauss points). To overcome the locking phenomena, the pressure is considered constant over the element and computed only at a central quadrature point.

**Fluid approach** The fluid approach is based on a stabilized mixed linear velocity-pressure finite element formulation. A triangular mesh is used for the domain discretization. The velocity, the pressure and temperature fields are computed at each node of the element. The deviatoric stresses and the other internal variables are computed at each quadrature point of the element. This approach is exposed in more details in [1, 2, 7, 8].

2.2 Numerical simulation strategy

2.2.1 Arbitrary Lagrangian Eulerian formulation

In region 1 and region 3 in figure 2(b) the Arbitrary Lagrangian Eulerian formulation is used. Indeed, the Eulerian formulation (used in the region 3) is a particular case of
the ALE formulation. With the ALE formalism, the mesh displacement can be decoupled from the material displacement [1, 2, 4, 5, 6]. The ALE formulations used in the two approaches are different.

**Solid approach**  The ALE formulation used in the solid approach is described in more details in [4, 5, 6]. To simplify the solution procedure and remain competitive against Lagrangian models, the system of ALE equations is solved using an operator-split procedure. First, for each time step, the classical Lagrangian formalism is used. During this Lagrangian step the mesh sticks to the material until an equilibrated Lagrangian configuration is iteratively obtained. The second step, also called the Eulerian step, is divided into two substeps: first the nodes of the mesh are relocated to a more suitable position, thus defining a new mesh. In the case of region 1 and region 3 the position of the relocated nodes is known because the mesh velocity of these regions is imposed. Then, the unknowns and the internal variables are transferred from the old mesh to the new one [6].

**Fluid approach**  The ALE formulation used in the fluid model is not based on a operator-split like in the formulation presented for the solid approach. In this fully coupled formulation [1, 2], the equilibrium state is computed at each time increment without remeshing and remapping steps. The system of equations solved includes the convective terms due to the velocity of the mesh relative to the material. In the TMAZ, region 1 in figure 2(b), the velocity of the mesh is imposed and the velocity and pressure of the material are directly computed. In the case of the Eulerian formulation (region 3 in figure 2(b)), the mesh does not move during the computation.

2.2.2 The transition zone

**Solid approach**  In the solid approach, the transition zone is a ring with a finite thickness (region 2 in figure 2(b)). In this region, the evolution of the rotational speed of the mesh, which differs from the material velocity, is linearly interpolated between the ALE region and the Eulerian zone. As the mesh distortion grows with time, a remeshing operation is periodically required. For one full rotation of the pin, the remeshing process is applied 30 times. The time interval between two successive remeshings is thus constant.

The remeshing operation can be divided into two steps. First, a better-suited mesh, called the new mesh, is created. In the case of the transition zone, the relatively simple geometry of this region allows an easy generation of the new quadrangular mesh.

Then, to carry on the computation over this mesh, the state variables from the mesh before remeshing, called the old mesh, has to be transferred to the new one. Each field used to define the equilibrium state is transferred independently from the other ones. The data transfer method used in this paper is called the Finite Volume Transfer Method (FVTM) with linear reconstruction of the fields. In [9, 10], this transfer method is presented in
more details and the comparison with some of the remapping algorithms most commonly used in the literature brings to light the advantages of this method.

**Fluid approach** In the fluid model the transition zone (region 2 in figure 2(b)) is limited to a circle (zero thickness). Each node of the mesh on this circle is duplicated. One node is linked to the ALE region (numbered 1) and the other one to the Eulerian region (numbered 3). The coupling between both regions is performed using a specific node-to-node link approach. At every mesh movement step, for a given node of the ALE region, the corresponding node of the Eulerian one is found and a link between the two nodes is created. Afterwards, the boundary conditions and the properties of the plate nodes are copied to the corresponding TMAZ nodes within the link. The time step can be conveniently chosen such that the two interface meshes (ALE and Eulerian) are always compatible. In this case the ALE mesh would slide precisely from one Eulerian interface node to the next one at each time step.

2.2.3 Thermomechanical constitutive model

In both models, the constitutive model of the tool is thermo-rigid. So, no mechanical fields are computed over this material. However, from the point of view of the thermal equations, the tool has a classical thermal behaviour as far as heat conduction is concerned. In addition, the material behaviour of the plates is modelled as thermo-visco-plastic using a Norton-Hoff constitutive model:

\[ S = 2\mu D \left( \sqrt{3} \left( \frac{2}{3} D : D \right) \right)^{m-1} \]  \hspace{1cm} (1)

where \( m \) and \( \mu \) are the rate sensitivity and viscosity parameters respectively. Both are temperature dependent. \( S \) is the deviatoric part of the stress tensor and \( D \) is the deviatoric part of the strain-rate tensor.

**Solid approach** In the solid model, the value of the variation of the pressure \( (dp) \) is computed thanks to the variation of the volume \( (dV) \) and the bulk modulus \( (K) \):

\[ dp = K dV. \]  

In addition, with the solid approach, it is possible to replace the Norton-Hoff constitutive model with a thermo-elasto-visco-plastic one, see e.g. [11]. With this kind of constitutive model, it is possible to compute the residual stresses.

**Fluid approach** In the fluid model, the material is assumed to be incompressible and this constraint is incorporated into the equations to be solved.
3 Comparison of numerical results based on the two approaches

In this example, the section of the pin is an equilateral triangle (figures 3 and 4(a)). The mesh used with the solid model is presented at the beginning of the computation in figure 3. The dimensions of the tool are:

- radius of the circumscribed circle to the pin: 3 mm.

The width of the two plates is 50 mm and the simulated length is 100 mm. The simulated region of the plates is a square with a side of 100 mm. The centre of the pin is on the centre of this square (see figures 3 and 4(a)).

The most important parameters of the considered FSW process are:

- rotation speed: 40 RPM
- welding speed: 400 mm min\(^{-1}\).

The thermo-mechanical properties of the plates are the following:

- density: 2700 kg m\(^{-3}\)
- bulk modulus: 69 GPa (used only with the solid approach)
- thermo-mechanical Newton-Hoff law (presented in the part 2.2.3), with \( \mu = 100 \, MPa, m = 0.12 \),
- heat conductivity: 120 W m\(^{-1}\) K\(^{-1}\)
- thermal expansion coefficient: \( 1 \times 10^{-6} \) K\(^{-1}\)
- heat capacity: 875 J kg\(^{-1}\) K\(^{-1}\)

The thermo-mechanical properties of the tool are the following:

- density: 7800 kg m\(^{-3}\);
- heat conductivity: 43 W m\(^{-1}\) K\(^{-1}\);
- heat capacity: 460 J kg\(^{-1}\) K\(^{-1}\).

The total time of the simulation is 15 seconds which corresponds to 10 revolutions for the pin.

Figures 4(b), 4(c) and 4(d) show the evolution of the pressure computed by the two models at three reference points defined in figure 4(a). Points 1 and 2 move with the mesh, because these points have the same rotational speed as the pin. On the other hand,
4 Conclusion and future works

The phenomena happening during the Friction Stir Welding (FSW) process are at the interface between solid mechanics and fluid mechanics. In this paper, two different methods are presented to simulate the FSW process numerically. One model is based on a solid approach which computes the position and the temperature fields and another one is based on a fluid approach written in terms of velocity, pressure and temperature fields. Both models use advanced numerical techniques such as the Arbitrary Lagrangian Eulerian formalism or remeshing operations or an advanced stabilization algorithm. These advanced numerical techniques allow the simulation of the FSW process with non-circular tool
Figure 4: Evolution of the pressure computed by the two models at the three reference points

shocks. The presented example (with a triangular pin) shows that the two formulations essentially deliver the same results. More investigations are still needed to understand the small differences between the two models. While the fluid model is more efficient from a computational point of view, the model based on the solid approach has the advantage that it can be used to compute the residual stresses (the thermo-visco-plastic constitutive model can be replaced with a thermo-elastic-visco-plastic one).

Acknowledgements

The Belgian authors wish to acknowledge the Walloon Region for its financial support to the STIRHETAL project (WINNOMAT program, convention number 0716090) in the
Figure 5: Evolution of the temperature computed by the two models at the three reference points.

References

kinematic framework for the numerical modelling of friction stir welding. *Computers


stir welding process simulation: Calibrations from welding experiments. *In-
Figure 6: Temperature field (in °C) at the end of the computation (time : 15 s) obtained with the solid model.


NUMERICAL INVESTIGATION OF CONTINUOUS FIBER GLASS DRAWING

Q. CHOUFFART∗, P. SIMON† AND V. E. TERRAPON∗

∗Department of Aerospace and Mechanical Engineering
University of Liege
1 Chemin des chevreuils, 4000 Liège, Belgium
e-mail: vincent.terrapon@ulg.ac.be - web page: http://www.mtfc.ulg.ac.be/

†3B - The Fibreglass Company
Rue de Charneux 59, 4651 Battice, Belgium
e-mail: philippe.simon@3b-fibreglass.com - web page: http://www.3b-fibreglass.com

Key words: Fiberglass drawing, glass forming, fiberization, multiphysics simulation

Abstract. The physics of glass fiber drawing is studied through numerical simulations and experimental measurements, with a focus on the fluid region, from the hole tip at the bushing plate to the glass transition point. The influence of the different heat transfer mechanisms is investigated to understand their respective impact on fiberization, such as fiber radius attenuation and internal stresses. Numerical predictions are then compared to experimental data measurements obtained from a dedicated fiberization unit. Numerical and experimental results show a good agreement. In particular, it is found that the ambient air temperature and the radiation have an important impact on the fiber cooling rate. Moreover, for a prescribed fiber diameter, internal stresses are lower when operating at a higher temperature.

1 Introduction

Due to their good mechanical properties, glass fibers are frequently used for the reinforcement of composite materials in many engineering fields. The industrial process to manufacture these fibers consists in continuously drawing and cooling a glass melt into fibers. The liquid glass, which is melted upstream in a furnace, flows through a bushing plate, thus forming long filaments. The bushing plate is typically composed of 1'000 to 8’000 holes, each with a diameter of about 1 to 2 mm. The different holes have a cylindrical protuberance on the lower side of the bushing plate, which is hereafter simply referred to as “tip”. These filaments are then quickly cooled by fins and water sprays to reach glass transition, and undergo solidification before being wound at high velocity (see figure 1). During this process, the fiber diameter is strongly attenuated (by up to
two orders of magnitude). The final fiber diameter (typically 10 to 34 µm) is controlled by the drawing velocity and the glass flow rate. It has been observed during industrial production that this continuous process is highly sensitive to small disturbances that can lead to fiber breaking. The fiber breaking problem, common across the glass fiber industry, strongly limits the efficiency of the process. In particular, the break of a single fiber requires taking the specific bushing position off-line during several minutes, and leads to a large quantity of glass waste, which can typically amount to up to 10% of the total glass production. Therefore, the overall objective of the present research is to understand the physical mechanisms underlying fiber breaking and to identify a strategy to reduce it. In this context, the first step is to characterize the physics governing the glass fiber forming through numerical modeling of a single fiber.

Different studies have already been performed on the physics of fiber glass drawing, but unfortunately many of the results still remain confidential within industrial companies. During the 1960’s, Glicksman [1] used a one-dimensional Newtonian fluid model to understand the behavior of the fiber during the manufacturing process. Neglecting two-dimensional effects was justified by the very small slope of the free surface in the central region, i.e. far away from the tip. His theoretical study was compared to experimental data and good agreement was found, thus validating the one-dimensional assumption for the central region. However, this simplification is not adequate to accurately represent the region very close to the tip. Huynh and Tanner [2] solved numerically the two dimensional problem using a finite elements method. They showed that radiation is the dominant mode of cooling in the vicinity of the tip exit and they proposed to use a non-constant effective emissivity in order to take into account the radius attenuation along the axial component. Purnode [3, 4] also used a two-dimensional finite elements model and relied on both steady and unsteady simulations. His results confirmed that radiation is the main heat transfer mechanism near the tip exit, while convection becomes dominant further away. Moreover, surface tension was shown to be destabilizing while convection had a stabilizing effect. Von der Ohe [5] performed a numerical parametric study for different glass types and operating windows. She pointed out that the drawing speed and cooling conditions of the surrounding air are important, in particular for the final product properties. On the other hand, the experimental study by Rekhson [6] on a bushing plate with about hundred holes indicates that fiber break is mostly due to inhomogeneous heat patterns on the bushing plate. More recently, Liu [7] investigated numerically the variations of the final fiber diameter due to variabilities in the process and identified the ambient air and furnace temperatures as their main cause. These findings led to the development of a control system to reduce the variation of the fiber diameter due to both these effects by adjusting the winder velocity.

Although these models led to relative good agreement with experimental measurements, their accuracy could still be improved. In particular, internal radiation is often neglected and constant emissivity is typically assumed along the fiber surface. Additionally, the environment around the fiber is generally considered as isothermal although
experimental measurements demonstrate that it is not the case. These simple models are extended here to account for variations in the surface emissivity for radiative, and variations in the ambient air temperature for convective heat losses. A sensitivity analysis on some key parameters is performed and numerical predictions are compared to experimental data from the literature and from measurements performed on a dedicated fiberization unit.

2 Problem statement and governing equations

Glass fiber drawing involves various complex physics. It is generally accepted that a forming fiber can be divided into three regions as shown in figure 1. The first region is the forming shape at the tip exit where the largest radius attenuation occurs. As the temperature is high, the viscous glass melt flows as a free-surface jet and the heat transfers are essentially dominated by radiation. The typical length of this region is a few millimeters from the tip exit. The “draw-down region” then starts where the slope of the free surface becomes very small and ends at glass transition. In this region, the viscosity increases dramatically until the glass fully solidifies. Its length is typically about a few centimeters, and convection is the main cooling mechanism as a large heat exchange is promoted by the high drawing velocity. Finally, the last region is defined as the region after the transition where the material is in a glassy state. The fiber motion is there that of a solid rod translating axially. Note that the defining characteristics of these three regions change smoothly and continuously, so that their boundaries are only qualitative.

This work focuses on the first two regions. The model used is based on following assumptions: the glass material is modeled as an incompressible Newtonian viscous fluid, the flow is assumed to be axisymmetric, and only the steady state is considered. The liquid glass can be described by the basic governing equations of mass, momentum and energy conservation:

\[
\nabla \cdot \mathbf{v} = 0, \tag{1}
\]

\[
\rho \left( \mathbf{v} \cdot \nabla \right) \mathbf{v} = \nabla \cdot \mathbf{\sigma}, \tag{2}
\]

\[
\rho c_p \mathbf{v} \cdot \nabla T = \mathbf{\sigma} : \nabla \mathbf{v} + \nabla \cdot \left( k \nabla T \right) - \nabla \cdot \dot{\mathbf{q}}_r, \tag{3}
\]

where \( \mathbf{v} \) is the velocity vector, \( T \) the temperature, \( \mathbf{\sigma} \) the stress tensor, and \( \dot{\mathbf{q}}_r \) the radiative heat flux. The density \( \rho \), the specific heat \( c_p \) and the conductivity \( k \) are considered constant as their temperature dependence is low in the temperature range considered. For a Newtonian fluid the stress tensor is equal to \( \mathbf{\sigma} = -p \mathbf{I} + 2 \eta \mathbf{D} \), where \( p \) is the isotropic pressure and \( \mathbf{D} = 1/2(\nabla \mathbf{v} + \nabla \mathbf{v}^T) \) the strain-rate tensor. The dependence of the viscosity \( \eta \) on the temperature \( T \) is given by Fulcher’s law

\[
\eta = 10^{-A} + T_{0}^{B}, \tag{4}
\]

where \( A \), \( B \) and \( T_{0} \) are three constants that depend on the material considered. Note that the velocity and stress fields only depend on the temperature through the viscosity.


Figure 1: Schematics of the drawing process for a single fiber, computational domain (indicated by the dotted frame on the left) and boundary conditions.

2.1 Boundary conditions

The geometry is composed of the tip and the fiber. The problem being axisymmetric, only a slice is included in the computational domain. At the tip inlet, the volumetric flow rate is imposed based on the height \( H \) of the glass column and the viscosity of the melt [8]:

\[
Q_0 = \frac{\pi}{8\eta} \left( -\frac{\partial p}{\partial z} \right) R_0^4 = \frac{\pi}{8\eta} \left( \frac{\rho g H}{h} \right) R_0^4,
\]

(5)

where \( g \) is the gravity constant and \( R_0 \) and \( h \) are the tip radius and length, respectively. It is well-accepted that the viscosity at the tip should be around 1000 Poise to ensure fiberization, which corresponds to a temperature called \( T_3 \) (3 for \( \log_{10} \eta = 3 \)). As a result, a temperature \( T_0 \) around the value of \( T_3 \) (generally between \( \log_{10} \eta = 2.5 \) and \( \log_{10} \eta = 3 \)) is used as boundary condition for the tip walls, where a no-slip condition is imposed for the velocity field. At the outlet, a constant drawing velocity \( v_f \) is imposed in the axial direction. The final fiber diameter is thus directly controlled by the inlet volumetric flow rate and the drawing velocity.

Along the free surface, the interface condition is given by

\[
n \cdot \sigma_g - n \cdot \sigma_a = \gamma n \cdot (\nabla \cdot n),
\]

(6)

\[
v \cdot n = 0,
\]

(7)

where \( \gamma \) is the surface tension assumed constant, \( \sigma_g \) and \( \sigma_a \) the interface stress for glass and air, respectively, and \( n \) the surface normal. Note that \( \sigma_a \) is neglected here.

Both convective and radiative heat losses take place at the free surface. The heat flux leaving the fiber can thus be written as
\[ \dot{q}_s = \dot{q}_{\text{conv}} + \dot{q}_{\text{rad}} = h_c (T_s - T_{\text{ext}}) + \epsilon \sigma \left( T_s^4 - T_{\text{ext}}^4 \right), \]  

(8)

where \( h_c \) is the convective heat transfer coefficient, \( T_s \) the surface temperature, \( T_{\text{ext}} \) the ambient temperature around the fiber, \( \epsilon \) the emissivity of the fiber surface and \( \sigma \) the Stefan-Boltzmann constant.

The convective cooling is induced by the high drawing velocity of the fiber (of the order of 20 m/s) that entrains ambient air. Many previous studies have considered a constant ambient air temperature. However, experimental measurements show a strong gradient of \( T_{\text{ext}} \) along the fiber. This is due to the heating of the air in the vicinity of the bushing plate (whose temperature is around \( T_3 \)) and its entrainment downwards by the fiber. In the present model, the variations of the ambient temperature \( T_{\text{ext}} \) along the vertical coordinate \( z \) are considered by fitting experimental measurements of \( T_{\text{ext}} \). The convective heat transfer coefficient \( h_c \) is approximated by Kase and Matsuo correlation for a thin cylinder moving axially [9]

\[ h_c = 0.42 \frac{k_a}{D} \left( \frac{D}{\nu_a} \right)^{0.334}, \]  

(9)

where \( k_a \) is the air conductivity and \( D(z) \) the fiber diameter, and the Reynolds number is defined as \( \text{Re} = \frac{v_z D}{\nu_a} \), with \( \nu_a \) and \( v_z \) the air kinematic viscosity and the fiber axial velocity at the surface, respectively. This relation has been used in many previous studies and seems to be adequate [3, 5, 6]. Note also that the dependence of the air properties (e.g., \( k_a, \nu_a \)) on the temperature \( T_{\text{ext}} \) is included, and dry air is assumed.

The glass melt is a semi-transparent medium for some range of wavelengths and, thus, contributes in addition to conduction to internal heat transfers. Consequently, an accurate radiation model would be required to obtain the precise internal temperature profile and surface radiative flux. In order to simplify the model and reduce CPU time, internal radiation is modeled here through an effective emissivity \( \epsilon = \epsilon(D, T) \) that takes into account the variations of the fiber diameter \( D \) and surface temperature \( T_s \). At the same time, the radiative internal heat flux \( \dot{q}_r \) in Eq. (3) is neglected. The emissivity \( \epsilon \) decreases with decreasing \( D \) and increasing \( T \). Consequently, the change in emissivity along the fiber corresponds to a balance between these two effects. The effective emissivity is approximated by a cubic and linear fit of experimental measurements for \( D \) and \( T \), respectively.

The governing equations are solved numerically with ANSYS POLYFLOW [10] using a finite-element computational fluid dynamic method. The free surface is treated with an Arbitrary Lagrangian-Eulerian (ALE) formulation.

3 Experimental approach

A dedicated fiberization unit has been developed to study experimentally the drawing of a single fiber. Figure 2 shows the overall apparatus. Glass cullets are melted at a
temperature of up to 1450°C inside an electrically heated platinum-rhodium container. The temperature inside the furnace is monitored through three thermocouples located at different heights. The furnace temperature is continuously maintained at a constant temperature based on measurements from the middle thermocouple. The forming fiber is drawn from one single cylindrical tip by a rotating drum. The fiber diameter can be measured on-line through a laser diffractometer, from off-line optical measurements of fiber samples, or from measuring off-line the total mass of fiber produced. The take-up velocity of the rotating drum is also measured on-line. A camera with a macro lenses is used to take high-resolution digital photographs of the fiber meniscus, which are then post-processed to obtain the fiber diameter. Finally, the temperature of the ambient air around the fiber is measured through a thermocouple at different axial locations $z$. Measurement uncertainties on the velocity and final fiber diameter are of the order of 0.2% and 1 $\mu$m, respectively.

4 Results and discussion

4.1 Validation

The numerical predictions are first validated through comparisons with experimental studies from the literature and from measurements obtained on the dedicated fiberization unit described above. In particular, the experimental data from Glicksman’s case M5 [1] are used. We focus here on the meniscus attenuation, as it represents the easiest macroscopic quantity to measure experimentally, and is often the only one available in the literature. Additionally, the variation of the fiber diameter is also a good proxy for the velocity gradients. For the second set of experimental data, the fiber diameter is obtained from high-resolution photographs of the fiber free-surface. Note that only the upper part of the cone shape is measured in this case. The respective process conditions are summarized in table 1 and results are shown in figure 3(a). A good agreement can
be observed for both cases, especially in the upper part of the cone shape. Nonetheless, some discrepancies can be seen for the lower part (Glicksman’s case). This could be due to larger relative errors in the measurements as the fiber radius becomes very small in this region. At such a small radius, the fiber is represented by only a few pixels on the photograph so that the identification of the free-surface becomes much less accurate. Furthermore, the uncertainties of glass properties can provide error on simulation results.

Table 1: Experimental conditions for the two validation cases: volumetric flow rate $Q_0$, tip diameter $D_0$, tip temperature $T_0$, and drawing velocity $v_f$.

<table>
<thead>
<tr>
<th></th>
<th>$Q_0$ [m$^3$/s]</th>
<th>$D_0$ [mm]</th>
<th>$T_0$ [$^\circ$C]</th>
<th>$v_f$ [m/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Glicksman’s M5 case [1]</td>
<td>$3.171 \cdot 10^{-9}$</td>
<td>1.72</td>
<td>1227</td>
<td>25.88</td>
</tr>
<tr>
<td>2) Present experiment</td>
<td>$4.720 \cdot 10^{-10}$</td>
<td>1.20</td>
<td>1308</td>
<td>1.55</td>
</tr>
</tbody>
</table>

![Simulations](image1.png)
![Present experiment](image2.png)

Figure 3: (a) Comparison between the predicted and measured fiber radius $R(z)$ as a function of the distance $z$ from the tip for the two experiments of table 1. (b) Computed surface temperature $T_s$ and measured air temperature $T_{ext}$ as a function of $z$ for experiment 2. The dotted line represents an exponential fit of the measured $T_{ext}$.

4.2 Heat transfer analysis

Heat transfers play a critical role in the process, as the glass material is quickly cooled from the melt temperature to ambient temperature. Moreover, the thermal history, and in particular the cooling rate, determine the microscopic structure of the glassy state, and thereby, the properties of the solid glass. As such, the process and the fiber properties
are very sensitive to variations in the environment. For instance, it has been frequently observed that changes in the environment properties can lead to a large increase of the fiber breaking rate.

Unless otherwise stated, the following results are based on the second experiment of table 1 with an effective emissivity $\epsilon(D, T)$ and a variable external temperature $T_{\text{ext}}$. Figure 3(b) shows the temperature $T_s$ along the fiber surface as a function of the distance from the tip, illustrating the rapid cooling over a short distance. In the same figure, the measured external temperature $T_{\text{ext}}$ demonstrates that the assumption of a constant $T_{\text{ext}}$ is not adequate, as this temperature varies from about 650°C at the tip to room temperature further away. The normalized temperature profile $\Delta T(r, z) \equiv T(r, z) - T_s(z)$ at different $z$ locations is shown in figure 4(a). One can observe that the maximum temperature difference across a fiber section first increases to reach a maximum value of about 8°C at a distance $z \approx 2.5$ mm, and then slowly decreases to become much lower than 1°C. However, the fiber radius $R(z)$ experiences simultaneously a strong attenuation, so that the radial temperature gradients increase continuously.

The relative contributions of the radiative flux $\dot{q}_{\text{rad}}$ and convective flux $\dot{q}_{\text{conv}}$ to the total heat flux $\dot{q}_s$ at the fiber surface are shown in figure 4(b). These results confirm that radiation dominates in the first region of the forming fiber very close to the tip. Convection then becomes the major mechanism for heat loss when the fiber radius approaches its final value. The total heat flux shows two maxima corresponding to the maximum radiative and convective heat fluxes, respectively.

Figure 4: (a) Computed normalized temperature profile $\Delta T(r, z) \equiv T(r, z) - T_s(z)$ as a function of the dimensionless radial component $r/R$ at different $z$ locations for experiment 2. (b) Computed radiative, convective and total surface heat fluxes along the axial component $z$ for experiment 2.
The thermal history can be measured by the cooling rate

\[
\dot{T}_c = \frac{dT}{dt} = v_z \frac{\partial T}{\partial z},
\]
which represents the change in temperature over time experienced by a glass particle. Because internal diffusion in the axial direction is relatively small, the cooling rate averaged over a fiber cross-section is related to the surface heat loss and radius attenuation by

\[
\langle \dot{T}_c \rangle (z) \approx -\frac{2\dot{q}_s(z)}{\rho c_p R(z)},
\]
where \(\langle \cdot \rangle\) represents the average over the cross-section. In order to assess the influence of the effective emissivity, two additional simulations have been performed with a constant emissivity \(\epsilon = 0.35\) and \(\epsilon = 0.5\), respectively. The first value corresponds to a value often found in the literature, while the second is a better approximation for the glass type considered here. The cooling rates at the fiber surface obtained with both constant emissivity values and with the effective emissivity \(\epsilon(D, T)\) are compared in figure 5(a). A variation of about 10% of the maximum cooling rate is observed, where its maximum value is obtained using an effective emissivity. It is worth noting that those largest variations are observed in the convection-dominated region, although the emissivity is only related to radiation. In this case, changes in radiation lead to a different radius attenuation \(R(z)\), and, through Eq. (11), to a different cooling rate. This demonstrates the tight coupling between all physical phenomena and the importance of accurate modeling.

The influence of the environment surrounding the fiber is investigated by considering different external temperatures \(T_{ext}\). In particular, the surface cooling rates \(\dot{T}_{c,s}\) for a constant temperature \(T_{ext} = 650^\circ C\) (case 1), a variable external temperature with constant air properties, corresponding to \(T_{ext} = 650^\circ C\) (case 2), and a variable external temperature with air properties depending on temperature (case 3) are compared. Note that in the case of a variable external temperature, \(T_{ext}(z)\) is given by an exponential fit of the measured values as shown in figure 3(b). Results are summarized in figure 5(b) and show that the environment properties have a strong impact on the cooling rate. In particular, the cooling rate increases when the air temperature variation is considered since, in this case, the temperature difference \(T_s - T_{ext}\), and thereby the convection, are larger. This effect is intensified when changes in air properties are taken into account. This shows that an accurate description of the external temperature is very important.

4.3 Stresses

For a Newtonian fluid, the extra-stress tensor \(\tau = 2\eta \mathbf{D}\) depends on the viscosity and the velocity gradients. The normal stress in the axial direction is thus defined as

\[
\tau_{zz} = 2\eta(T) \frac{\partial v_z(z)}{\partial z},
\]
and is shown in figure 6(a). A rapid increase in $\tau_{zz}$ can be observed, which is related on the one hand to the rapid increase of the axial velocity $v_z$ with the radius attenuation (up to four orders of magnitude), and on the other hand, to the increase in viscosity as the glass cools down ($\eta \approx 10^{12}$ Pa·s around the transition). As the fiber radius approaches its final value $R_f$ and the glass approaches the transition temperature, the velocity gradients tend to zero but the viscosity experiences a dramatic increase, so that the final stress approaches a finite value. The normalized radial profile of the stress $\tau_{zz} - \tau_{zz,s}$ at different $z$ locations is shown in figure 6(b). Note that in the vicinity of the tip, the axial normal stress is larger at the centerline than at the surface. The opposite is observed further away from the tip since the temperature at the surface is there lower, and thereby the viscosity much higher, than at the centerline. The stress is also sensitive to the process conditions. Figure 6(c) shows the final surface stress at the outlet, where $R = R_f$, for different final diameter values (i.e., different drawing velocities). As expected, the stress increases as the diameter is reduced (approximately as $1/D_f^2$). Additionally, as the drawing velocity increases, the velocity gradients become larger, which leads to a thinner meniscus; the cooling rate increases more rapidly, as does the viscosity. Finally, figure 6(d) shows the final surface stress for different values of the tip temperature $T_0$ (i.e., different volumetric flow rates) and drawing velocity such that the final diameter remains $11 \, \mu$m. It can be seen that the final stress is lower when the fiberization temperature $T_0$ is increased.
Figure 6: Computed axial normal stress $\tau_{zz}$ for experiment 2. (a) Surface stress as a function of the axial coordinate. (b) Radial profile of the normalized stress $\tau_{zz}(r, z) - \tau_{zz,s}(z)$ at different axial positions $z$. (c) Final surface stress as a function of the final fiber diameter (i.e., at different drawing velocities). (d) Final surface stress as a function of the tip temperature $T_0$ (i.e., at different flow rates).

5 Conclusion

The physics of glass fiber drawing has been investigated experimentally and numerically. The numerical results have been compared and validated with data from the literature and from a dedicated fiberization unit. Based on the numerical model, it has been demonstrated that radiation dominates in the region very close to the tip. Nonetheless, a variation of emissivity has also an impact on the fiber cooling far away from the tip, as the cooling rate depends not only on the surface heat loss but also on the attenuation of the meniscus. It was also shown that prescribing a correct external temperature is very important as the cooling rate, and thereby all other properties, are very sensitive to the
heat loss. Finally, the internal stresses have been computed and their dependence on some key parameters has been computed. Since the viscosity increases as a logarithmic law, stresses are extremely sensitive to small temperature variations. It was shown that the stresses increase when the fiber diameter and/or the fiberization temperature are reduced. In conclusion, simulations can provide a useful tool to gain insight into the physics of fiber glass drawing and to devise new strategies for adjusting the process operating window.

Future work will focus on improving the models for internal radiation and the surrounding environment, which might include numerical simulations of the air flow. The unsteady case will be considered. Additionally, the transition phase should also be included into the model. Finally, this knowledge will be used to characterize the physical mechanisms of fiber breaking.

REFERENCES


TOOLING MATERIALS AND SOLUTIONS FOR THIXOFORMING STEEL

AHMED RASSILI* AND JACQUELINE LECOMTE-BECKERS*

*Thixo Unit ULg, Bd de Colonster, 4
B-4000 Liège, Belgium
e-mail: a.rassili@ulg.ac.be

Keywords: Thixoforging, steel, high melting point, tool material.

Abstract. The aim of this work is to evaluate the thermal and mechanical loadings applied to the tools during steel thixoforming process in order to determine appropriate tool materials and solutions. This evaluation was realized thanks to experimental trials and to the finite elements simulations. The effect of these loadings on the tool’s failure modes are highlighted and compared to the ones observed in classical forming processes. Beyond this, the failure modes of different tool materials and solutions are presented. The tested materials are hot-working tool steels. Other possibilities and tool coating or surface treatments are discussed as well.

1 INTRODUCTION

Due to high slug temperature (usually higher than 1350°C), tools surfaces reach very high temperature. In hot forging, this temperature could already reach 500°C [1-3]; in thixoforming, tool temperature increase could reach 700°C and even higher. Such a temperature is higher than classical tool steels annealing temperature and could lead to a fall of the mechanical properties. In order to minimize the thermal shocks, dies are usually pre-heated from 40 to 350°C in hot forging, but this doesn’t prevent the temperature from increasing.

Thixoforming process, as hot forging is composed of three sequential steps:
• Brutal contact of high temperature slug on the tool. If needed, tool closing could be done before or after this step.
• Forming step during which mechanical constraints are applied to the tool.
• Part ejection and tool cooling.

In production, these steps are repeated in a cycle. Tool damaging could be due to different mechanisms: fatigue cracking following thermomechanical loading cycle, microstructure evolution or scaling due to hot working, geometrical modification generated by wearing or plastic deformation. These mechanisms are commonly known as: (1) abrasive wearing, (2) thermal fatigue, (3) mechanical fatigue, (4) plastic deformation [4]

In thixoforming, thermomechanical loadings are quite different as forming loads are lower but thermal loads are higher. The failure modes could be different too or simply accentuated.
Their determination is the first aim of this paper as well as a set of solutions either in terms of materials or coatings.

2 EXPERIMENTAL AND SIMULATION INVESTIGATIONS

2.1 Tool Design

The tool used during this work (Figure 1) forms an axisymmetric H shape part. The deformation is a compression followed by an important reverse than a direct extrusion. Due to small thickness of the walls, this geometry highlights the thermal and stress effects occurring during forming. The dies and the punches are instrumented by thermocouples in order to measure their inner thermal fields. At the beginning of cycle, the tool is open and the punches are out of the dies. When heating is done, robot puts the slug in the lower die and moves back. Then, the upper part of the tool moves down to close it and the two punches form the part. It is also possible to form it with the upper punch alone if the lower one is already inside the die at the beginning of cycle or consolidated to this one. This tool is also used to determine friction parameters [5].

2.2 Tool Material

Tool has been made of two different tool materials and compared to classical Z38 hot working tool steel. The latter has a good thermal shocks resistance thanks to the presence of chrome, molybdenum and vanadium. It is commonly used as die material in hot forging [6]. The X38CrMoV5 composition is given on Table 1.

Table 1: Mass composition of X38CrMoV5 hot working tool steel [6]

<table>
<thead>
<tr>
<th>C</th>
<th>Cr</th>
<th>Mn</th>
<th>V</th>
<th>Ni</th>
<th>Mo</th>
<th>Si</th>
</tr>
</thead>
<tbody>
<tr>
<td>[%]</td>
<td>0.40</td>
<td>5.05</td>
<td>0.49</td>
<td>0.47</td>
<td>0.20</td>
<td>1.25</td>
</tr>
</tbody>
</table>
Chrome, molybdenum and vanadium precipitate carbides which increase wearing resistance. Chrome and molybdenum delay the softening due to annealing. Chrome and vanadium inhibit the grains coarsening during austenitizing and chrome and silicon increase scaling resistance.

Nevertheless, this steel grade looses a part of its mechanical properties at high temperature. Table 2 gives the mechanical properties of X38CrMoV5 for four working temperatures for a material previously oil-quenched from 1040°C after two tempering at 640°C. At 600°C, yield strength is nearly divided by two. Extrapolated until 800°C, Rp0.2 falls to 400MPa, so lower than the locking force applied on the dies. Moreover, the austenitizing beginning temperature (830°C) is close to the working one [7].

Table 2: Mechanical properties of X38CrMoV5 at different working temperatures [7]

<table>
<thead>
<tr>
<th>Temperature [°C]</th>
<th>Rm [MPa]</th>
<th>Rp0.2 [MPa]</th>
<th>A [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1400</td>
<td>1170</td>
<td>12</td>
</tr>
<tr>
<td>400</td>
<td>1170</td>
<td>1020</td>
<td>13</td>
</tr>
<tr>
<td>500</td>
<td>1050</td>
<td>900</td>
<td>18</td>
</tr>
<tr>
<td>600</td>
<td>810</td>
<td>700</td>
<td>25</td>
</tr>
</tbody>
</table>

Figure 2: Yield strength of the investigated tool materials compared to Z38
Figure 2 shows the yield strength of the two investigated tool materials compared to Z38. At this stage of analysis, we could notice that these two materials will exhibit a more interesting behaviour at high temperature as their yield strength doesn’t drop drastically at 600 °C especially for material C.

### 2.3 Modeling

The Finite Elements code Forge2011© was used for the simulations. The constitutive law used in this work is quite simple and mainly driven by the liquid fraction, and so the temperature. Thus, the structure of the raw material and its evolution are not explicitly represented. Even if this is a limitation of the calculation results, the error on the flow behavior is small for high solid fraction. Thermal exchanges are already taken into account by the FE code.

The constitutive law is a classical Spittel one (which is the default law used by the solver) when material temperature is lower than solidus and a modification of this Spittel equation when the material temperature is higher than solidus. The modification induces a linear decrease of the consistency by multiplying it by a factor going from one to zero between the solidus and the liquidus. There is then a smooth transition between semi-solid and solid behavior during cooling.

The constitutive law is

\[
\sigma = A e^{m_1 \varepsilon} e^{m_2 \dot{\varepsilon}}
\]

for \( T < T_{\text{solidus}} \) and

\[
\sigma = A \left( \frac{T_m - T}{T_{\text{liq}} - T_{\text{sol}}} \right) e^{m_1 \varepsilon} e^{m_2 \dot{\varepsilon}}
\]

for \( T_{\text{solidus}} < T < T_{\text{liquidus}} \)

### Table 3: Values of the constants used in equations (1) and (2)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2707.108</td>
</tr>
<tr>
<td>m1</td>
<td>-0.00325</td>
</tr>
<tr>
<td>m2</td>
<td>-0.00325</td>
</tr>
<tr>
<td>m3</td>
<td>0.1529</td>
</tr>
<tr>
<td>m4</td>
<td>-0.05494</td>
</tr>
<tr>
<td>( T_{\text{sol}} )</td>
<td>1315°C</td>
</tr>
<tr>
<td>( T_{\text{liq}} )</td>
<td>1480°C</td>
</tr>
</tbody>
</table>

In these equations, \( \sigma \) is the stress, \( \varepsilon \) is the strain, \( \dot{\varepsilon} \) is the strain rate, \( T \) is the temperature, \( T_{\text{liq}} \) is the liquidus temperature, \( T_{\text{sol}} \) is the solidus temperature and \( A, m1, m2, m3 \) and \( m4 \) are
constants depending on the steel grade. For 100Cr6 steel, the values of the constant parameters are given in table 3. The values of A and m1 to m4 come from the database of Forge2011© and the values of T_{liq} and T_{sol} have been obtained by Differential Scanning Calorimetry (DSC) [8].

3 RESULTS AND DISCUSSIONS

3.1 Mechanical loading

In the case of thixoforming, mechanical loadings are about ten to twenty times lower than in hot forging [9, 10]. Figure 3 shows the Von Mises equivalent stresses, calculated by the Forge2011© software, at the end of forming inside the lower part of the tool. The simulated forming is a 100Cr6 steel slug symmetrically deformed with a tool speed of 170mm/s. It appears that maximum stress, for the areas in contact with the semi-solid material, is around 260MPa. This maximum stress is located in the center of the punch top surface. If the punch temperature reaches a value for which its material yield stress is lower than 260MPa, there would be a deformation of this punch.

The simulation did not take into account the locking force applied to the dies to keep the tool closed. In the present case, this force is 2000kN. Depending on the locked surface, the pressure could be as high as 630MPa.

Figure 3: Von Mises equivalent stresses inside the tool at the end of forming

3.2 Thermal loading

In hot forging, slugs are usually heated at a temperature higher than 1000°C. Their contact with the dies could heat these ones up to 500°C. In thixoforming, the working temperatures are still higher, until more than 1400°C. Tools surfaces are then subjected to very high temperature. The H shape tool has been designed in order to be instrumented by
thermocouples (TC1 and TC2 in Figure 2). The measures of these thermocouples allowed validating the temperature fields calculated by simulation, as shown on Figure 4. The aim of the comparison is to show that simulation investigations could be of great help in order to determine better solutions for tool materials prior to any costing experimental investigations.

Figure 4: Comparison of the temperatures inside the tool measured by thermocouples (dashed) and calculated by Forge2011© (continue)

Figure 5 shows the temperature fields inside the lower part of the tool for the whole forming process. Simulation shows that, at the moment of the ejection, the surface temperature could reach 1100°C in the punch and about 580°C in the lower die. In this case, if the stresses are important, because of galling during the ejection for example the tool could easily be damaged.

Simulation has been run for tool material C. In the case of another tool material, and thus another thermal conductivity, the surface temperature should be different. In the case of a lower thermal conductivity, the surface temperature would be higher, which will be interesting from the forming point of view as the flowing material temperature would stay higher during a longer time and thus, the forming load would be lower. At the opposite, from the tool point of view, this higher surface temperature would increase the risk to overrun the tool material yield stress and to damage the tool. Thermal stresses, coming from thermal gradients, depend of these gradients value and of the thermal dilatation coefficient. The effects of these thermal stresses are shown in figures 6, 7 and 8.
A. Rassili and J. Lecomte-Beckers.

3.3 Wearing

Figure 6 shows the area of maximum wearing. As in hot forging, they are located where sliding speeds are the higher, thus mainly at the punch edge. On the die, there is not any wearing at the joining plane level as the tangential speed is equal to zero on this area in the case of a symmetric deformation. As the working temperature is higher in thixoforming, the tool wearing resistance is lower than in hot forging.
3.4 Hardness

Hardness of the tooling’s lower punch has been measured after 50 cycles of forming. Figure 7 shows the hardness values measured on the punch and the line along which the measured have been made.

![Figure 7](image)

**Figure 7**: Profile comparison on the punch after 26 (a) and 50 forming cycles (b). Hardness measured along the lower punch after 50 parts forming (section cut A – B of Figure 8) (c)

The graph shows that the lower and middle parts of the punch have kept their original hardness, around 43-45 HRC. But the hardness of the corners has noticeably increased to 51 HRC due to tempering after annealing. The area between has a hardness between 32 to 39 HRC. This softening is due to annealing occurring at high temperature. This means that the punch is more easily deformable. Moreover, some marks are visible on the punch surface, due to galling and abrasive wearing.

3.4 Mechanical resistance

Figures 7(a – b) and Figure 8 give profile measures of the upper punch (Figure 7) and die (Figure 8). These profiles have been measured on different points in order to limit the impact of local damages. The profile of the punch lateral surface (Figure 7) shows significant modification of the shape. It seems that mechanical resistance of the punch is not high enough to avoid plastic deformation.

![Figure 8](image)

**Figure 8**: Profile measures on the die before (left) and after 50 parts forming (right)
A. Rassili and J. Lecomte-Beckers.

The top surface of the die shows important modifications (6 to 10 tenth of millimeter) at the joining plane level. This is due to the friction of the hottest contact zone which induces important stresses (about 300MPa). Around this joining plane, the temperature could reach 570°C (Figure 5).

4. CONCLUSIONS AND REMARKS

Up to now, thixoforming tool lifetime is still the main lock to the technology industrialization. Due to high working temperature, mechanical features of the hot work tool steels classically used in hot forging strongly decrease. In particular, hardness and yield stress are too low to guarantee the tooling integrity.

Plastic deformation is the main issue. It is due to mechanical and thermal stresses. Compared to hot forging, mechanical stresses are clearly lower but thermal stresses are higher. However, at industrial production rate (6-12 parts per minute), the working temperature should be higher but temperature variation would be lower, so the thermal fatigue should be lower than in the case of laboratory study. To minimise the thermal loadings, tool materials and solutions exit and give good results as shown in Figure 6. An important point is also to minimise the contact time between tool and semi-solid steel in order to minimize the tool temperature. Parts ejection must then be as fast as possible to decrease thermal loading.

5. ACKNOWLEDGEMENT

The authors gratefully acknowledge the University of Liège and the Walloon Region for their financial support as well as collaborators from CRM, MK and SIRRIS.

6. REFERENCES


COUPLING LEVEL SET METHODS WITH BOUNDARY INTEGRAL METHODS FOR FREE SURFACE POTENTIAL FLOWS

M. Garzon*, L. J. Gray† AND J.A. Sethian††

* Dpto. of Applied Mathematics
Universidad de Oviedo
Calvo Sotelo s/n, 33007 Oviedo, Spain
e-mail: maria.garzon.martin@gmail.com

† Bergen Software Service International
Post Box 2921, 5825 Bergen, Norway
e-mail: len@bssi-tt.com

†† Dept. of Math., U.C. Berkeley and Math. Dept., Lawrence Berkeley National Lab., CA, USA
e-mail: sethian@math.berkeley.edu

Key words: Potential flows, Level Set methods, Boundary Integral Methods, Wave Breaking, Drop formation

Abstract. Non viscous and rotational free fluid flows in moving domains can be used to model various physical phenomena such as wave breaking, droplet and bubble breakup dynamics and electrostatically charged drop distortion, among others.

The classical lagrangian fully non linear potential flow model can be recasted in a complete Eulerian formulation using the Level Set approach. The coupled system of partial differential equations is approximated using a Boundary Integral method, to calculate the moving front velocity, and the Level Set method to update the position of the front. The numerical methods involved are: a linear Boundary Element method for the laplacian equation and first order upwind schemes for the Level Set hyperbolic equations. This novel approach has turn out to be very robust and versatile, due to the capability of the Level Set method to handle topological changes of the moving domain.

Full details of how to arrive to the Eulerian formulation of the model equations will be presented and the numerical schemes of the coupled system will be also described. Several computational results will be addressed: The propagation and shoaling of a solitary wave over a sloping beach and the collapse of an infinite fluid column under the action of surface tension forces. Recent results on the evolution and distortion of charged droplets and subsequent Coulombic fission will be presented in more detail.
1 INTRODUCTION

In this paper we present a class of problems in the field of fluid mechanics that can be modeled using the potential flow assumptions, that is, inviscid and incompressible fluids moving under an irrotational velocity field. While these are significant assumptions, in the presence of moving boundaries, the resulting equations is a non linear partial differential equation, which adds considerable complexity to the computational problem. In the literature this model is often called the fully non linear potential flow model (FNPFM). Several interesting and rather complicated phenomenon are described using the FNPFM, as for example, Helle-Shaw flows, jet evolution and drop formation, sprays and electro sprays, wave propagation and breaking mechanisms, etc, see [1], [2], [3].

Level Set Methods (LSM) [4] are widely used in fluid mechanics, as well as other fields such as medical imaging, semiconductor manufacturing, ink jet printing, and seismology. The LSM is a powerful mathematical tool to move interfaces, once the velocity is known. In many physical problems, the interface velocity is obtained by solving the partial differential equations system used to model the fluid/fluids flow. The LSM is based on embedding the moving front as the zero level set of one higher dimensional function. By doing so, the problem can be formulated in a complete Eulerian description and topological changes of the free surface are automatically included. The equation for the motion of the level set function is an initial value hyperbolic partial differential equation, which can be easily approximated using upwind finite differences schemes.

Recently, the LSM has been extended to formulate problems involving the transport and diffusion of material quantities, see [5]. In [5] model equations and algorithms are presented together with the corresponding test examples and convergence studies. This led to the realization that the nonlinear boundary conditions in potential flow problems could also be embedded using level set based methods. As a result, the FNPFM can also be formulated with an Eulerian description with the associated computational advantages. Several difficult problems that have been already approximated using this novel algorithm are wave breaking over sloping beaches [7], [8]; the Rayleigh Taylor instability of a water jet [9]; a two inviscid fluid system of different densities [10]; and the evolution of charged and uncharged droplets under the action of an electrical field [12].

2 POTENTIAL FLOW EQUATIONS

In this section we first we present the general conservations laws applied to a fluid volume $V$ and then, by introducing the required assumptions, we derive the potential flow model equations. Denote by $\rho = \rho(P,t)$ the volumetric mass density of the continuous medium at point $P$ and at time $t$ and $u$ the velocity field, the mass conservation law is

$$D_t \rho + \rho \text{div} \ u = 0.$$  \hfill (1)

For the conservation of momentum, Newton’s law is applied to a fluid volume $V$:

$$D_t \int_V u \ \rho dV = \int_V \mathbf{g} \ \rho dV + \int_{\partial V} \tau (ds).$$  \hfill (2)
The term in left hand side of this equation is the rate of change with time of the momentum associated with volume $V$ when dragged by the continuous medium. The first term in the right hand side corresponds to the volumetric forces inside $V$, generated by a vector field per unit mass $g$, usually the gravitational field. The second term represents the “contact” forces applied by the rest of the medium over the part in $V$. The Cauchy’s tensor $\tau$ is a linear operator field that is obtained from specific relationships which depend on the material, the so called constitutive relations. We are interested in inviscid fluids which verify the Pascal’s law: $\tau(ds) = -pds$, where $p$ is the pressure scalar field. Green’s formula,

$$\int_{\partial V} -p \, ds = \int_V -\nabla p \, dV,$$

shows that contact forces may be computed as a kind of volume forces with density $-\nabla p$. For a small volume $\delta V$ dragged by the fluid, equation (2) can be written:

$$D_t(u \rho \delta V) = (g \rho - \nabla p) \delta V.$$  \hspace{1cm} (3)

Due to the mass conservation law, $D_t(\rho \delta V) = 0$, equation (3) leads to the Euler equation:

$$D_t u = \partial_t u + \partial_u u = g - \frac{1}{\rho} \nabla p.$$  \hspace{1cm} (4)

If $g$ is a uniform field it comes from the gradient of a potential function:

$$g = -\nabla U(P), \quad U(P) = -g \cdot (P - O),$$

where $P - O$ is the position vector of the point $P$.

Now we introduce the potential flow restrictions. Assuming constant fluid density and an irrotational flow regime, $\text{curl } u = 0$, there exists an scalar field $\phi$ such that $u = \nabla \phi$ and Eqn.(1) becomes

$$\Delta \phi = 0.$$  \hspace{1cm} (5)

Outside of the fluid domain, and separated by a free boundary, there is a gas at pressure $p_a$ that, for now, is assumed to be constant.

Using the vectorial relationship $\nabla u^2/2 = \partial_u u + u \times (\text{curl } u)$, $u = \nabla \phi$ and $\text{div } u = 0$ we have

$$\nabla \left( \partial_t \phi + \frac{1}{2} u^2 + \frac{p}{\rho} + U \right) = 0.$$  \hspace{1cm}

Performing the first integration,

$$\partial_t \phi + \frac{1}{2} u^2 + \frac{p}{\rho} + U = C(t),$$

where $C(t)$ is an arbitrary function of time, which can be chosen in such a way that the previous relation can be written:

$$\partial_t \phi + \frac{1}{2} u^2 + \frac{p - p_a}{\rho} + U = 0.$$
Now using the obvious relation $\partial_t \phi + u^2 = \partial_t \phi + \partial_u \phi = D_t \phi$, we finally obtain

$$D_t \phi - \frac{1}{2} u^2 + \frac{\rho - \rho_a}{\rho} + U = 0. \quad (6)$$

Now, let $\Omega_t$ be a closed 3D moving fluid domain, $\Gamma_t$ the free surface boundary at time $t$ and $Q$ the position vector of a fluid particle on the front ($Q$ will be precisely defined in the next section). To close the system we have to add the conditions at the moving boundary:

- The kinematic boundary condition, which states that the particles in the free front move with velocity $u$, $D_t Q = u$.

- The Dynamic boundary condition, which comes from the conservation of momentum when we impose the continuity of the stress tensor across the free boundary,

$$D_t \phi = f,$$

being

$$f = -U + \frac{1}{2} u^2 - \frac{\rho - \rho_a}{\rho}.$$

Therefore, the Lagrange formulation of the potential flow equations is:

$$u = \nabla \phi \text{ in } \Omega_t$$
$$\Delta \phi = 0 \text{ in } \Omega_t$$
$$D_t Q = u \text{ on } \Gamma_t$$
$$D_t \phi = f \text{ on } \Gamma_t \quad (10)$$

Outside the fluid domain we may consider several possible scenarios, as for example:

1. An infinite exterior fluid at rest.

2. An infinite exterior moving fluid of different density (an inviscid two fluid system).

3. A dielectric gas under the action of a uniform electric field at the far field.

For each of these scenarios the corresponding model equations in the exterior domain have to be coupled with system (7)-(10), and a rich variety of dynamics arises depending upon the expression of $f|_{\Gamma_t}$ in each particular case.

Next, the boundary conditions posed on the moving front are going to be embedded in higher dimension equations using the level set techniques. This procedure will lead to the complete Eulerian formulation of the potential flow equations.
3 THE LEVEL SET METHOD

The Level Set method is a mathematical tool very adequate to follow interfaces which move with a given velocity field. The key idea is to view the moving front as the zero level set of one higher dimensional function called the level set function. One main advantage of this approach comes when the moving boundary changes topology, and thus a simple connected domain splits into separated disconnected domains.

Let be $\Gamma_t$ the set of points lying in the surface boundary at time $t$. This surface is defined through the zero level set of the scalar field $\Psi = \psi(P,t)$:

$$\Gamma_t = \{ Q | \psi(Q,t) = 0 \}.$$  \hfill (11)

To identify the fluid particles, the configuration at $t_0$ (reference configuration) is used:

$$\Gamma_{t_0} = \{ Q_0 | \psi(Q_0,t_0) = 0 \}.$$  \hfill (12)

The particle movement is specified through the function

$$Q = R(Q_0,t),$$  \hfill (13)

which gives the position $Q \in \Gamma_t$ of the fluid particle $Q_0 \in \Gamma_{t_0}$. The particle $Q_0$ velocity is calculated using the convective derivative $D_t$ (“following the particle”):

$$u = D_t Q = \frac{d}{d\epsilon} R(Q_0, t + \epsilon) \bigg|_{\epsilon=0}.$$  \hfill (14)

According to definition (11), we have $\psi(R(Q_0,t),t) = 0$. Deriving with respect to time and applying the chain rule, we obtain

$$\partial_t \Psi + u \cdot \nabla \Psi = 0,$$  \hfill (15)

which has to be completed with the value of the level set function at time $t = 0$, usually set to be the signed distance function to the initial front,

$$\Psi(P,0) = s(P)d(P),$$

being $d(P)$ the distance from the point $P$ to the surface at the initial configuration $\Gamma_0$, $s(P) = -1$ if $P \in \Omega_0$ and $s(P) = +1$ if $P \notin \Omega_0$.

Now, if we take a fixed 3D domain $\Omega_D$ that contains the free surface for all times, we can define the initial value problem for the level set function $\Psi$ posed on $\Omega_D$:

$$\partial_t \Psi + u \cdot \nabla \Psi = 0 \text{ in } \Omega_D$$

$$\Psi(P,0) = s(P)d(P) \text{ in } \Omega_D$$  \hfill (16) (17)

Equation (16) moves all the level set of $\Psi$, not just the zero level set, and in many physical applications the front velocity is just defined for points lying on the free boundary.
Therefore for this equation to be valid on the whole domain we have to extend the velocity $u$ off the front [6].

Equation (9) can be directly formulated as the level set Eq. (15). For the velocity field $u(Q,t)$, the trajectory of a fluid particle at initial position $Q_0$ is given by the solution of

\begin{align}
D_t Q &= u(R(Q_0, t), t), \\
R(Q_0, 0) &= Q_0.
\end{align}

(18)

Next, let $G(P,t)$ be a functions defined on $\Omega_D$ such that for every $Q \in \Gamma_t$

$$G(Q, t) = \phi(Q, t),$$

(19)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Extension of the velocity potential off the front}
\end{figure}

It is important to remark here that $G(P, t)$ is an auxiliary functions defined in $\Omega_D$ that can be chosen arbitrarily, the only restriction is that it equals $\phi(Q, t)$ on $\Gamma_t$. Figure 1 gives an interpretation of this property for a moving curve in 2D. Applying the chain rule in identity (19) we obtain

$$\partial_t G + u \cdot \nabla G = f$$

(20)

which holds on $\Gamma_t$. Note that $u$ and the right hand side of Eq. (20) are only defined on $\Gamma_t$, and thus, in order to solve these equations over the fixed domain $\Omega_D$, these variables must be extended off the front.

The system of equations, written in a complete Eulerian framework, is

\begin{align}
\mathbf{u} &= \nabla \phi \quad \text{in} \quad \Omega_t \\
\Delta \phi &= 0 \quad \text{in} \quad \Omega_t \\
\Psi_t + \mathbf{u}_{\text{ext}} \cdot \nabla \Psi &= 0 \quad \text{in} \quad \Omega_D \\
G_t + \mathbf{u}_{\text{ext}} \cdot \nabla G &= f_{\text{ext}} \quad \text{in} \quad \Omega_D
\end{align}

(21)-(24)

Here the subscript “ext” denotes the extension of $f$ and $\mathbf{u}$ onto $\Omega_D$.

The numerical approximation of this coupled system of PDEs Eqs.(21)-(24) can be described in two basic steps. The time derivatives in the level set equations are approximated using a standard first order backward Euler scheme and the space derivatives a first
order upwind scheme. At each time step, Eqn. (22) is solved using its Boundary integral formulation and a Boundary linear element approximation is used. The complete details of the numerical approximation can be found in [7], [9], [11].

4 EXAMPLES OF POTENTIAL FLOW MODELS WITH MOVING BOUNDARIES

4.1 The wave breaking problem

The coupled level set/extension potential equations for breaking waves is set in two dimensions. Let \( \Omega_t \) be the 2D fluid domain in the vertical plane \((x, z)\) at time \( t \), with \( z \) the vertical upward direction (and \( z = 0 \) at the undisturbed free surface), and \( \Gamma_t \) the free boundary at time \( t \) (see Figure 2).

![Figure 2: The domain](image)

For this case, surface tension forces will not be considered and thus the expression of \( f \) in Eqn. (24) is

\[
f = \frac{1}{2} u^2 - gz.
\]

Here we have to add the boundary condition on the rest of the fluid boundary, \( \phi_n = 0 \) on \( \Gamma_b \cup \Gamma_1 \cup \Gamma_2 \).

A complete numerical convergence study for the wave breaking problem can be found in [7], [8], where it has been also studied the influence of the beach bottom profile on the wave breaking characteristics. Here, in Fig. 3 we just show the case of a solitary wave with initial height \( H_0 = 0.6 \) shoaling, turning and breaking over a mild constant slope of 1:22.

4.2 The Rayleigh Taylor instability

A significant challenge in the numerical solution of free boundary problems is when the domain undergoes topological changes. This is the case of the Rayleigh-Taylor instability of a fluid column of length \( L \), in which an small initial perturbation of wave number \( k = \frac{2 \pi}{L} \) will lead to the fluid overturning, pinch-off and subsequent cascade of drop formation. To model this problem we have used the axisymmetric version of the coupled level
Figure 3: Wave shape at various times for a beach slope of 1 : 22

set/extension potential equations and the cylinder geometry in the \((r, z)\) plane is depicted in Fig. 4.

Figure 4: Cylinder geometry in the \(r - z\) plane.

Here, surface tension forces have to be considered, being the dominant forces in this scenario. The pressure jump across the free boundary is \(p = p_a + \gamma \kappa\), where \(\gamma\) is the surface tension coefficient and \(\kappa\) is twice the mean curvature of the moving surface. As the effects of gravity are neglected, the expression of \(f\) in Eqn. (24) is

\[
f = \frac{1}{2} u^2 - \frac{\gamma}{\rho} \kappa.
\]

Periodic boundary condition on the lateral walls of the cylinder are imposed.

An extensive numerical study and validation of the collapse of a fluid column is presented in [9]. In Fig. 5 we show the fluid column overturning, just prior to pinch-off. After separation the satellite drop undergoes its own instability and it pinches-off in its middle point. This originates capillary waves propagating on the satellite surface, which lead to a cascade of drop formation, including fast separations and reconnections. Comparison with Laboratory experiments and self similar scaling laws can also be found in [9].
4.3 Drop distortion under the action of an electrical field

For this problem we consider a perfectly conducting inviscid fluid droplet of density \( \rho \) surrounded by a dielectric gas of permittivity \( \epsilon \) and exposed to the action of an uniform electric field at infinity. Here we have to couple the potential fluid equations that govern the interior drop dynamics with the electrostatic assumptions for the exterior gas. As the ambient medium is considered uniform and uncharged, the electric field is the gradient of an electric potential \( W \),

\[
\begin{align*}
E & = -\nabla W \quad \text{in } \Omega_2(t) \quad \text{(25)} \\
\Delta W & = 0 \quad \text{in } \Omega_2(t) \quad \text{(26)} \\
W & = W_0(t) \quad \text{on } \Gamma_t \quad \text{(27)} \\
W & = -E_\infty z \quad \text{at infinity} \quad \text{(28)}
\end{align*}
\]

The electric stresses at the free surface elongate the drop against the restoring effect of the surface tension and the pressure jump across the free surface \( \Gamma_t \) is given by

\[
p = p_a + \gamma \kappa - \frac{\epsilon}{2} E_n^2, \quad \text{(29)}
\]

where \( E_n = -\nabla W \cdot \mathbf{n} \), \( \mathbf{n} \) the unit normal vector pointing from the interior to the exterior domain. In this case the right hand side expresion for Eqn. (24) is

\[
f = \frac{1}{2} \kappa^2 - \frac{\gamma}{\rho} \kappa - \frac{\epsilon}{2\rho} E_n^2.
\]

The complete model can be written in its non-dimensional form, where the only parameter left is the non-dimensional electric field intensity. Theoretical and experimental studies show that there exists a critical value \( E_c \approx 0.3 \) such that for \( E_\infty < 0.3 \) the drop oscillates with known frequency, while for \( E_\infty > 0.3 \) the drop is distorted in the direction of the electric field. A cone singularity (the taylor cone) appears from which thin jets are ejected. In Fig. 6 we depict the drop profiles at various times: the drop elongates and finally pinches-off and jet discharge occurs. Laboratory photographs, kindly provided by Grimm and Beauchamp, are shown on the left for two different electric field intensities; on the right the corresponding numerical profiles are also shown. A complete study of electrostatically driven jets from non-viscous droplets will be presented elsewhere [12].

5 CONCLUSIONS

- The Eulerian formulation in 3 dimensions of the fully nonlinear potential flow model has been derived using the Level Set technique, which automatically includes topological changes of the free boundary. The numerical approximation of the coupled Level Set-Boundary Integral equations uses very simple first order numerical schemes, which are proved to be enough to capture complex non viscous fluid dynamics.
- Very interesting phenomena, such as wave breaking, the collapse and pinch-off of a fluid column and the distortion of droplets under the action of an electric field, are successfully simulated using this novel algorithm. We briefly show some results and address to the proper references.

REFERENCES


Figure 5: Fluid column overturning and pinching followed by a cascade of drop formation, top to bottom
Figure 6: Laboratory photograph left; computational profiles right at indicated times
PARALLEL APPLICATION ON HIGH PERFORMANCE COMPUTING PLATFORMS OF 3D BEM/FEM BASED COUPLING MODEL FOR DYNAMIC ANALYSIS OF SSI PROBLEMS

M.C.GENES*

* Faculty of Engineering, Civil Engineering Department
Zirve University
Kızılhisar Campus, 27260 Gaziantep, Turkey
e-mail: cemal.genes@zirve.edu.tr, www.zirve.edu.tr

Key words: Soil-Structure Interaction, High Performance Computing, Coupled Method, Finite Element, Boundary Element.

Abstract. Implementation of an improved parallel computation algorithm into a coupled model based on Finite Element and Boundary Element Methods for analysis of three-dimensional Soil-Structure Interaction (SSI) problems on High-Performance Computing (HPC) platforms is presented. The model and the parallel computation algorithm are developed for the linear analysis of large-scale three-dimensional SSI problems. The finite element method is used for modeling the finite region and the structure, and the Boundary Element Method is used for modeling the soil extending to infinity. The parallelization of the model is performed by the calculation of the impedance coefficients on the interaction nodes between the near- and the far-fields. The performance of the parallel computation algorithm is represented by elapsed timing measurements according to the number of processors. The efficiency of the proposed parallel algorithm of the coupled model is validated with one numerical example that confirm the consistent accuracy and applicability of the parallel algorithm by considerable time saving for large-scale problems.

1 INTRODUCTION

In many fields of engineering, Finite Element Method (FEM) and the Boundary Element Method (BEM) are frequently applied analysis tools. Each one of these methods has its specific areas of applications. The FEM, is especially well suited for the analysis of problems involving in-homogeneities or non-linear behavior of solid bodies, [1, 2]. The BEM has advantages, if stress singularities on unbounded media are present. BEM has further advantages, if incident wave fields need to be considered in dynamic problems, [3-5]. Therefore, over the last 25 years, much progress has been made in finite/boundary element coupling methods to solve problems involving sub-regions with different characteristics. During this coupling, the respective advantages of both methods are used. In structural mechanics, these coupling methods are usually used for assessing the dynamic responses of stiff, heavy and embedded structures, such as nuclear reactors, gravity dams, tunnels, liquid-storage tanks and buildings with the soil media surrounding their foundations. Thus, soil-structure interaction (SSI), when incorporated, allows realistic prediction of the coupled
behavior of the soil and structure. An arbitrarily shaped non-homogeneous body subjected to
dynamic loads requires the use of the FEM and BEM together with unbounded boundary. 
BEM is a semi-analytical method and requires the fundamental solutions pertaining to the 
region, provided that the radiation condition at infinity is satisfied, [6-7].

Application of parallel computation to engineering problems is a relatively recent 
development that started approximately 15 years ago. The algorithms are widely used in 
ingineering applications for large-scale computationally intensive problems. Widely used 
parallel algorithms are based on partitioning a computational domain and then assigning each 
partitioned domain to a separate computer processor, thus reaching a solution using many 
processors concurrently, [8, 9]. The parallel implementation of BEM codes have been studied 
by many prior researchers [10-14] and are well summarized in Davies[15]. More recent 
studies are found in Cunha and et al.[16], Bird and et al.[17], and Park and Heister[18]. Cunha 
and et al.[16] applied the standard and portable libraries for the parallelization of BEM codes; 
Bird and et al.[17] used a coupled BEM/scaled boundary FEM formulation to analyses linear 
elastic fracture mechanic problems; Park and Heister [18] proposed a parallelization 
procedure for the analysis of unsteady BEM problems. Most of these studies have worked on 
a structural problem or parallel implementation itself. However, in the study of Park and 
Heister[18] the simulation code for the free-surface problem has the unique dynamic grid 
characteristic.

The SBFEM is an alternative and effective method for modeling systems with finite and 
infinite extension having non-homogeneous material properties, [19-22]. Genes and 
Kocak[23, 24] applied the SBFEM to large-scale systems on high performance computing 
platforms. In this study, the linear equation solver of the system equations obtained from 
BEM for impedance analysis is used as portable library called The Scalable Linear Algebra 
Package (ScaLAPACK1) and the impedance analysis is also parallelized by distributing the 
impedance analysis work according to the number of the interaction nodes between the near-
and far-field to the slave processors. Hence, the applicability of BEM/FEM coupling 
procedure to the high performance computing platforms for large-scale 3D SSI problems is 
demonstrated. The gained experience and the coded intelligent routines for the parallelization 
of SBFEM [23] are implemented to the parallelization of SBFEM/BEM/FEM [26] coupling 
and BEM/FEM coupling. In the proposed SSI model, where the boundary at the interface 
between the near-field (the structure and surrounding soil medium), and the medium which is 
extending to infinity, is modeled by the BEM. The structure and the surrounding soil medium 
are modeled by FEM. The dynamic stiffness matrix of the boundary is combined with the 
dynamic stiffness matrix of the finite medium by using the substructure method (SM). In this 
coupled model, best attributes of these two methods are combined.

Layered finite medium can be discretized by using a parametric soil model, where soil is 
composed of sub-layers, represented by some parameters as proposed by Kocak and 
Mengi[27]. However, in this study, to propose a more generic model, layered infinite medium 
is modeled by using 3D BEM formulation for linear elastodynamics[16]. The proposed model 
and parallel algorithm is verified by studying two examples that are analyzed by the coupled 
FEM/BEM and FE/SBFEM models. It is found that, depending on the number of processors

1The ScaLAPACK is a set of library for distributed memory MIMD (Multiple Instruction Multiple Data) parallel 
computers developed by the ScaLAPACK project, [25].
used, the proposed parallel algorithm can decrease the computation time considerably. The obtained results of the model are compared with the results given in the literature and good agreements are noted, [28-30]. Comparisons showed that the Parallelized Coupled FEM/BEM Model can be used in SSI analyses of large-scale structures efficiently and accurately. The results also demonstrate the importance and the advantages of using parallelization for SSI problems that are complex in geometry and have non-homogeneous unbounded media.

2 PHYSICAL MODELS AND NUMERICAL APPROACHES

2.1 Structural Dynamics with FEM

In this study, the dynamic response of structures is described by the equation of motion resulting from FEM formulation [1] in the time domain as,

\[ \mathbf{M} \ddot{\mathbf{U}} + \mathbf{C} \dot{\mathbf{U}} + \mathbf{K} \mathbf{U} = \mathbf{F}(t) \]  

(2.1)

where; \( \mathbf{M} \): mass matrix, \( \mathbf{C} \): damping matrix denoting inner or structural damping, \( \mathbf{K} \): static stiffness matrix, \( \mathbf{F}(t) \): time dependent dynamic load acting on the structure caused by the external harmonic or transient vibrations or seismic excitation, \( \dot{\mathbf{U}} \): acceleration vector, \( \dot{\mathbf{U}} \): velocity vector, and \( \mathbf{U} \): displacement vector. Equation of motion can be written as below in the frequency domain by ignoring the damping matrix and taking into account the structural damping,

\[ \{(1 + 2iz)\mathbf{K} - \omega^2 \mathbf{M}\} \mathbf{U}^f = \mathbf{F}^f \]  

(2.2)

where; \( z \): hysteretic damping, \( \omega \): frequency, \( i \): imaginary number, and \( f \): frequency space.

In this study, all the formulations are derived in frequency space. For the sake of simplicity, the superscript \( f \) will be omitted. The term in curly brackets in Eq. (2.2) is referred to as the dynamic stiffness matrix and will be represented as \( \mathbf{S} \). For the finite region in Figure 1, Eq. (2.2) can be written in matrix form as,

\[
\begin{bmatrix}
S_{ii} & S_{is} \\
S_{si} & S_{ss}
\end{bmatrix}
\begin{bmatrix}
\mathbf{U}_i \\
\mathbf{U}_s
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{F}_i \\
\mathbf{F}_s
\end{bmatrix}
\]  

(2.3)

where; subscripts; \( i \): interaction nodes, and \( s \): non-interaction nodes.

2.2 Formulation of Elastodynamic Problems by Boundary Element Method

The detailed formulation of boundary element (BE) equation for elastodynamic problems in time and frequency domains is presented in the literature, [31-35]. The BE equation can be written as in Eq. (2.4) for the elastodynamic analysis of the 3D body shown in Figure 2 in Fourier transform domain.

\[
c\mathbf{u}(A) = \int_S \mathbf{G}(A,P)\mathbf{t}(P)dS + \int_S \mathbf{H}(A,P)\mathbf{u}(P)dS + \int_V \mathbf{G}(A,P)\mathbf{f}(P)dV
\]  

(2.4)

where; \( S \) indicates the surface; \( V \) volume of the body. \( \mathbf{H} \) and \( \mathbf{G} \) are the matrices obtained by the integration of the first and second fundamental solutions of BEM over the each boundary element. The \( \mathbf{u} \), \( \mathbf{t} \) and \( \mathbf{f} \) are the displacement, traction and volume force vectors, respectively.
Also, $A$ and $P$ indicates the used constant and integration points during the evaluation of the integrals. Equation (2.4) shows, the relation between the displacements of point $A$ (Fig. 3), and the defined integrals on the body volume ($V$) and the boundary surface ($S$).

The BE equation given in Eq. (2.4) for the elastodynamic problems, can be used for the calculation of the unknown displacements and traction vectors on the boundary.

The resultant system of equation can be obtained by some manipulations of Eq. (2.4) as:

$$AX=F$$  \hspace{1cm} (2.5)

where; $A$ is the coefficient matrix on the unknown quantities, and $F$ is the force vector obtained from the multiplication of known quantities and their coefficient matrix.

The solution of Eq. (2.5) gives the unknowns on the boundary of the body. In this study, Eq. (2.5) is used for the calculation of the impedance coefficients on the interaction nodes between the structure and the far-field. The parallelization of the model is performed on the calculation of the impedance coefficients by using Eq. (2.5) and the parallel solution of Eq. (2.5) by using ScaLAPACK.
2.3 BE Formulation for Dynamic Analysis of Soil Medium with Two Layers

The BE formulation for the dynamic analysis of homogeneous soil medium, can be modified for the dynamic analysis of nonhomogeneous soil medium with two layers. The detailed formulation is given in the study of Tanrikulu [37]. The resultant system equation of the composite body composed of three materials can be obtained as:

\[
\begin{bmatrix}
\bar{H}_1 & \bar{H}_{1L} & 0 & \bar{H}_{1c} & -\bar{G}_{1c} & -\bar{G}_{1L}
\end{bmatrix}
\begin{bmatrix}
\bar{u}_1
\bar{u}_{1L}
\bar{u}_c
\bar{t}_1
\bar{t}_{1L}
\end{bmatrix}
=\begin{bmatrix}
\bar{G}_1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\bar{t}_1
\bar{t}_2
\end{bmatrix}
\tag{2.6}
\]

The Eq. (2.6) can be written as a resultant system equation as in Eq. (2.5).

3 FEM/BEM COUPLING PROCEDURE FOR THE LAYERED SOIL AND PARALLELIZATION

The FEM/BEM coupling procedure for the layered soil (FE Eq. 2.3 and BE Eqs. 2.5 or 2.6), is not straightforward as the coupling procedure for the homogeneous soil [38, 26]. Therefore, with the help of the substructure method, impedance analysis must to be performed for the interaction nodes between the soil and the structure. The resulting impedance matrix is combined with the FEM Eq. (2.3). In this study, the impedance analysis is performed on the layered soil media by using the BEM. For the BEM analysis, an algorithm presented by Tanrikulu[37] for layered soil medium is used in the parallel algorithm proposed in this study. The definition of the impedance analysis can be explained on the model given in Figure 4. In other words, the impedance matrix is the inverse of the compliance matrix. The compliance matrix can be obtained by applying unit load in each degree of freedom of the interaction nodes to retrieve the displacements. The impedance matrix \( S^\infty \) is the contribution to the structure of the layered soil medium extending to infinity. The impedance matrix can be combined with the FEM equation by using the substructure method. The combined equation for the Coupled FEM/BEM model is given as:

\[
\begin{bmatrix}
S_{xx} & S_{xi}
S_{ix} & S_{ii} + S_{i\infty}
\end{bmatrix}
\begin{bmatrix}
U_x
U_i
\end{bmatrix}
=\begin{bmatrix}
F_x
F_i
\end{bmatrix}
\tag{3.1}
\]

Impedance analysis for a SSI system with \( n \) degrees of freedom of the interaction nodes between the structure and the layered soil, needs \( n \) times repeated analysis of BEM equation. This is very time consuming analysis. Therefore, in this study a parallelization algorithm is proposed and applied. The proposed parallel algorithm calculates the impedance matrices for small frequency steps of an interval. If the number of nodes between the structure and the unbounded soil medium \( m \), the total number of degrees of freedom on the interaction nodes is \( n=3m \). This means that, at each frequency step, \( n \) number of system of equations obtained by BEM on the interaction nodes have to be analyzed. The proposed parallel algorithm performs this analysis by partitioning the columns of the impedance matrix to the processors equally for the calculation of the impedance matrix (Fig. 5).
The parallel program assigns the load of each processor as,

$$n_{load} = \frac{m}{nproc}$$  \hspace{1cm} (3.2)

where, \(m\): the number of interaction nodes, \(nproc\): the number of slave processors. Consequently, each processor calculates the impedance vectors on \(n_{load}\) number of nodes for each degree of freedom. The calculated impedance vectors are collected on the master processor and combined to obtain the full impedance matrix. The flowchart of the parallel program is presented in Figure 6. In the parallel program, one of the processors is assigned as the master and the others as the slaves. In the workflow of the parallel computation procedure, the master program first defines the computation load, then calculates the dynamic stiffness matrix of the FE region and defines the parameters of the far-field for BEM formulation. Then, it sends the data related to far-field and the computation load of each processor to the slave processors. Each slave processor computes the impedance coefficients at certain number of interaction nodes by using BEM. The BEM equation is solved by using ScaLAPACK library by the slave processor on sub-slave processors (Fig. 7). Figure 7 shows the communication scheme between the processors for a cluster has 16 processors. The impedance matrix \(S^\infty\) is combined at the master processor by receiving the calculated columns of \(S^\infty\) from slave processors. The dynamic stiffness matrix of FE region is combined with impedance matrix by using the compatibility of the displacements and the equilibrium of the forces at interaction nodes as in Eq. (3.1) by the master processor. The solution of this linear system of equation is performed by LAPACK library [39].

Parallelization and fast matrix inversion are critical for dynamic analysis problems such as this as matrix inversion step tends to dominate computational time in most cases. The fully populated coefficient matrix provides a challenging problem for fast inversion schemes as one can not take advantage of any banded structures in general. Therefore, solving the linear system effectively may be a key factor to enhance the performance of the code. The ScaLAPACK is a set of library for distributed memory MIMD (Multiple Instruction Multiple Data) parallel computers developed by the ScaLAPACK project [25]. ScaLAPACK provides routines for dense and banded systems of linear equations; linear least squares problems, and eigenvalue and singular value problems. Since the ScaLAPACK routines are portable to any distributed memory computer using either Message Passing Interface (MPI) or a Parallel
Virtual Machine (PVM), the parallelization is directly carried out by applying the routine of ScALAPACK on the linear system \((\textbf{AX}=\textbf{F})\) obtained from the BEM calculation.

![Diagram 1](image1.png)

**Figure 6.** Parallel solution flowchart for the Coupled FE-BE Model of layered soil

![Diagram 2](image2.png)

**Figure 7.** Communication scheme between the processors for a cluster with 16 processors
4 EFFICIENCY OF THE PARALLEL PROGRAM

In the parallel program, the master processor partition the BE computation load to the slave processors and collects the calculated impedance coefficients from the slaves. Next, the master processor adds the complete impedance matrix to the coefficient matrix of the FEM equation and solves it by using the LAPACK library. Initial testing of the efficiency of the improved program was performed by comparing the elapsed time for the BE meshes as shown in Figures 8a and 8b. The meshes have 384 and 1526 boundary elements. The durations are presented in Table 1 for the calculation of the first 5 frequencies of solution interval. As can be seen from these durations for different number of processors in Table 1, the proposed parallel algorithm decreases the computation time at least by a factor of approximately 15. By using the elapsed timing given in Table 1, the speedup curves are plotted in Figure 9. This figure exhibits the efficiency of the method with respect to the minimum number of processors, and as expected, when the number of processors increases to a certain number, the computation time decreases accordingly. It also can be seen from the Figure 9 that for larger problems the speedup is better, which can be explained by the problem becoming more computation bound, since for the platform used in this study, CPU speed is faster than that of message passing.

![Figure 8. Boundary Element Meshes with 384 elements (Mesh I) and 1536 element (Mesh II)](image)

![Figure 9. Speed-up curves for Mesh I and Mesh II](image)

![Figure 10. Rigid Square Foundation rest on layered soil media](image)
Table 1. Timings for Mesh I and Mesh II according to number of processors

<table>
<thead>
<tr>
<th></th>
<th>Timings according to number of processors (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Mesh I</td>
<td>262</td>
</tr>
<tr>
<td>Mesh II</td>
<td>895</td>
</tr>
</tbody>
</table>

5 NUMERICAL EXAMPLES

In this section, one well known example problem is solved by the proposed parallel program for the presented SSI model and compared with other methods presented in the literature.

As stated before, in the present model, infinite nonhomogeneous soil region is modeled by boundary elements. The infinite regions, calculated by BEM, are plotted separately in order to demonstrate the interface between finite and infinite regions. During the analysis, the matching nodes of finite and infinite regions are automatically combined by the program.

5.1 Rigid Square Foundation on Elastic Nonhomogeneous Soil

The parallel solution algorithm and the presented model for SSI analysis in frequency domain are verified by comparing vertical, horizontal, and rocking impedances obtained for a rigid square foundation on layered nonhomogeneous soil (Fig. 10). The results are compared with those by Wong and Luco [28]. The rigid square foundation is considered completely connected to the soil media so that, the interaction surface between the structure and the soil medium is displaced compatibly with the rigid foundation. The origin of the coordinate system $X_1 X_2 X_3$ (O) is located at the center of the interaction surface (Fig. 10).

The BE mesh given in Figure 8a is used to compare the results obtained from Coupled FEM/BEM model in this study with those by Tanrikulu[37]. In this study, the impedance matrix is calculated by using the same BE solution algorithm presented by Tanrikulu[37] with the difference that, in this study, for the Coupled FEM/BEM model, the solution steps presented in the flowchart given in Figure 6 are used. In this problem the best speedup is obtained when 10 processors are used at the same time. For more consistent comparison with the literature, the BE mesh given by Tanrikulu[37] is also used. The horizontal, vertical and rocking impedances for the layered medium with different Poisson’s ratios and material properties are presented in Figures 11-14. As seen in Figures 11 and 12, the present model demonstrates quite good match with the others, Wong and Luco[28]. In addition, Figure 13 shows the applicability of the model to layered soils.

![Figure 11. H, V and R impedance-frequency for three elastic materials ($\nu_1 = \nu_2 = \nu_3 = 0.45$)](image-url)
6 SUMMARY AND CONCLUSIONS

In this article, a computational model is presented for large-scale SSI analysis of layered media using the Parallelized Coupled FEM/BEM Model. In the proposed model, the finite region, which might be considered as the structure, is modeled by the FEM. On the soil-structure interface, the boundary at the bottom of the finite media which is extending to infinity is modeled by the BEM. The analyses are conducted in the frequency domain.

Dynamic stiffness matrices pertaining to related regions of the SSI system are calculated by FEM and BEM, and combined by using sub-structuring method. Depending on the scale of the problem, the number of the slave processors can be determined by the user for the impedance analysis of the boundary by using BEM. The impedance matrix obtained by BEM for the contribution of the far-field is symmetric and the summation with the FEM equation is easy. Two example problems are solved to verify and investigate the applicability of the model and the coded parallel algorithm. It is found that, depending on the number of processors used, the proposed parallel algorithm can decrease the computation time at least by a factor of approximately 15 for the given example meshes. The obtained results of the model are compared with the results given in the literature and good agreements are noted. Comparisons showed that the Parallelized Coupled FEM/BEM Model can be used in SSI analyses of large-scale structures efficiently and accurately. The results also demonstrate the
importance and the advantages of using parallelized coupled models for analyzing complex structures and non-homogeneous unbounded media. The proposed model and the developed program can be implemented to the non-linear analysis of SSI problems under transient or seismic loads.

Acknowledgements

The author is thankful to Drs. Y. Mengi and A.H. Tanrikulu for their permission to use their BEM program in the coupled model and comparisons. Also, the support, and the invaluable contribution by Dr. S. Kacin and Dr. H.R. Yerli are greatly appreciated. This study was performed at Mustafa Kemal University as a part of the project (Tubitak, ID:106M258) sponsored by The Scientific and Technological Research Council of Turkey.

References


APPLICATIONS OF AN INCOMPRESSIBLE FLUID-RIGID BODY INTERACTION ON PROGRESSIVE MOVING-GRID FINITE-VOLUME METHOD

Shinichi ASAO* and Kenichi MATSUNO†

*Department of Mechanical Engineering, College of Industrial Technology, 1-27-1, Nishikoya, Amagasaki, Hyogo, 661-0047, Japan
  e-mail: asao@cit.sangitan.ac.jp, web page: http://www.sangitan.ac.jp

†Department of Mechanical System Engineering, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto, 606-8585, Japan
  e-mail: matsuno@kit.ac.jp - Web page: http://www.kit.ac.jp

Key words: Incompressible Flow, Moving Mesh, Finite Volume Method, Interaction Simulation

Abstract. The purpose of this paper is to introduce the Trans-mesh method and the Moving computational domain method as a progressive Moving-grid finite-volume method. In the Trans-mesh method, the bodies can move freely in a main mesh that covers the entire flow field. On the other hand, in the Moving Computational Domain method, the whole of the computational domain including bodies inside moves in the physical space without the limit of region size. These methods are constructed based on the four-dimensional control volume in space-time unified domain such that the method assures to be divergence-free in the space-time unified domain and thus satisfies both the physical and geometrical conservation laws simultaneously. The methods are applied to a falling sphere by gravity in an infinite long bending pipe and a trajectory of a flying ball over ground in incompressible fluid. The results indicate that these methods are promising in simulating the interaction of incompressible fluid-rigid body.

1 Introduction

Today, one of the interesting problems in CFD is a moving boundary problem. Especially, in the case of fluid-bodies interaction is interesting on engineering. For this problem, the body-fitted coordinate is usually applied. When multiple bodies move a long distance in the flow field, a body-fitted coordinate system hardly adapt the mesh to the motion of the bodies. This decreases an efficiency and accuracy. To overcome
this problem, we have proposed a Trans-mesh Method \[1\]. In the Trans-mesh method, a body moves freely in a stationary mesh while mesh planes are added, removed, and/or shifted in the main mesh satisfying both physical and geometric conservation laws. Next, we consider a flow around a sphere in the long pipe. It is necessary to make the computational mesh for whole of the pipe. Then, a huge number of computational mesh is needed. As the result, this simulation by using traditional method spends a lot of time. We have proposed the Moving Computational Domain method \[2\]. This method is kind of moving mesh method. The feature of this method is to move computational domain with the body. The Moving Computational Domain method can consider the region without limit. The only necessary assumption is that the conditions just in front of the computational domain should be known a priori, such as, stationary fluid state or uniform flow and so on. As these flow solvers, we modified the Moving-grid Finite-volume method \[3\]. The method is constructed based on the four-dimensional control volume in space-time \((x, y, z, t)\) unified domain such that the method satisfies the divergence-free character in the \((x, y, z, t)\) space and both the physical and geometrical conservation laws simultaneously \[4\]. Due to the use of four-dimensional control volume, the method has a lot of merits or freedom. The purpose of this paper is to introduce applications using the Trans-mesh method and Moving Computational Domain method. The methods are applied to a falling sphere by gravity in an infinite long bending pipe and a trajectory of a flying ball over ground in incompressible fluid.

2 Trans-mesh Method and Moving Computational Domain Method

2.1 Governing equations

The governing equations are the continuity equation and the incompressible Navier-Stokes equations. These are written as follows:

\[
\nabla \cdot \mathbf{q} = 0, \quad (1)
\]

\[
\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{E}_a}{\partial x} + \frac{\partial \mathbf{F}_a}{\partial y} + \frac{\partial \mathbf{G}_a}{\partial z} = - \left( \frac{\partial \mathbf{E}_p}{\partial x} + \frac{\partial \mathbf{F}_p}{\partial y} + \frac{\partial \mathbf{G}_p}{\partial z} \right) + \left( \frac{\partial \mathbf{E}_v}{\partial x} + \frac{\partial \mathbf{F}_v}{\partial y} + \frac{\partial \mathbf{G}_v}{\partial z} \right). \quad (2)
\]

Here \(\mathbf{q}\) is the velocity vector, \(\mathbf{E}_a, \mathbf{F}_a, \) and \(\mathbf{G}_a\) are advection flux vectors in the \(x, y,\) and \(z\) directions, respectively, \(\mathbf{E}_v, \mathbf{F}_v,\) and \(\mathbf{G}_v\) are viscous flux vectors, and \(\mathbf{E}_p, \mathbf{F}_p,\) and \(\mathbf{G}_p\)
are pressure flux vectors. The elements of the velocity vector and flux vectors are:

\[
\begin{pmatrix}
u \\
v \\
w
\end{pmatrix}, \quad
\begin{pmatrix}
u^2 \\
vw \\
u
\end{pmatrix}, \quad
\begin{pmatrix}
uw \\
v^2 \\
vw
\end{pmatrix}, \quad
\begin{pmatrix}
uw \\
vw \\
w^2
\end{pmatrix},
\]

\[
\begin{pmatrix}
p \\
0 \\
0
\end{pmatrix}, \quad
\begin{pmatrix}
0 \\
p \\
0
\end{pmatrix}, \quad
\begin{pmatrix}
0 \\
0 \\
p
\end{pmatrix}.
\]

(3)

Where \(u, v, \) and \(w\) are the velocity component of the \(x, y,\) and \(z\) directions respectively, \(p\) is the pressure, and \(Re\) is the Reynolds number. The subscripts \(x, y,\) and \(z\) indicate derivatives with respect to \(x, y,\) and \(z\), respectively.

As for motion of the body, six degrees of freedom is assumed, and the combined motion of the translation and rotation of the body is considered. The rigid body equations of motion are written as follows:

\[
\frac{dp_B}{dt} = f_B, \tag{4}
\]

\[
\frac{dL_B}{dt} = N_B. \tag{5}
\]

Here, \(p_B\) is the momentum vector of the body, \(f_B\) is the external force vector, \(L_B\) is the angular momentum vector, and \(N_B\) is the torque vector, respectively.

### 2.2 Moving-grid Finite-volume method

To assure the geometric conservation laws, we adopt a control volume in the space-time unified domain \((x, y, z, t)\), which is four-dimensional in the case of three-dimensional flows. Now, Eq. (2) can be written in divergence form as

\[
\nabla \cdot \tilde{F} = 0 \tag{6}
\]

where

\[
\nabla = \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} & \frac{\partial}{\partial t} \end{bmatrix}^T, \quad \tilde{F} = \begin{bmatrix} E & F & G & q \end{bmatrix}^T, \\
E = E_a - E_v + E_p, \quad F = F_a - F_v + F_p, \quad G = G_a - G_v + G_p. \tag{7}
\]

The present method is based on a cell-centered finite-volume method and, thus, the flow variables are defined at the center of the cell in the \((x, y, z)\) space. The control
volume becomes a four-dimensional polyhedron in the \((x, y, z, t)\) domain, as schematically illustrated in Fig. 1.

We apply volume integration to Eq. (2) with respect to the control volume illustrated in Fig. 1. By using the Gauss theorem, we can write Eq. (6) in surface integral form as,

\[
\int_{\tilde{\Omega}} \nabla \cdot \tilde{\mathbf{F}} \, d\tilde{\Omega} = \oint_{\partial \tilde{\Omega}} \tilde{\mathbf{F}} \cdot \tilde{k} d(\partial \tilde{\Omega}) \approx \sum_{l=1}^{8} \left( \tilde{\mathbf{F}} \cdot \tilde{n} \right)_l = 0. \tag{8}
\]

Here, \(\tilde{k}\) is an outward unit vector normal to the surface, \(\partial \tilde{\Omega}\), of the octahedron control volume \(\tilde{\Omega}\), in the space-time unified four-dimensional domain, and, \(\tilde{n}_l = (\tilde{n}_x, \tilde{n}_y, \tilde{n}_z, \tilde{n}_t)_l\), \((l = 1, 2, \ldots, 8)\) denotes the surface normal vector of the control volume and its length equal to the boundary surface area in four-dimensional \((x, y, z, t)\) space. The upper and lower boundaries of the control volume \((l = 7 \text{ and } 8)\) are perpendicular to the \(t\)-axis, and, therefore, they only have an \(\tilde{n}_t\) component, and its length corresponds to the volume of the cell in the \((x, y, z)\)-space at times \(t^n\) and \(t^{n+1}\) respectively. Thus, Eq. (8) can be expressed as,

\[
q^{n+1}(\tilde{n}_t)_8 + q^n(\tilde{n}_t)_7 + \sum_{l=1}^{6} \left( \tilde{\mathbf{F}}^{n+1/2} \cdot \tilde{n} \right)_l = 0. \tag{9}
\]

2.3 Concept of Trans-mesh method

In the Trans-mesh method, a body moves freely in a stationary mesh while mesh planes are added, removed, and/or shifted in the main mesh, as illustrated in Fig. 2. The front mesh plane of the moving body is eliminated from the main mesh in order to prevent the mesh from folding because of the reduced mesh spacing that occurs due to the movement of the body. Moreover, a new mesh plane is added between the rear plane of the moving body and the main mesh in order to maintain the maximum allowable mesh spacing. At the same time, the cells existing and connecting between the main mesh and the side
Shinichi ASAO and Kenichi MATSUNO

of the moving body are skewed due to the movement of the body. Thus, reconnection or exchange of mesh lines between the moving body and the main mesh is necessary. Hence, the present method essentially includes three inevitable procedures: addition and elimination of mesh planes as well as changes in the connecting relationships of structured mesh lines.

![Figure 2](image_url): Concept of Trans-mesh method.

### 2.4 Concept of Moving Computational Domain method

The basic coordinate system of the Moving Computational Domain method is the general, fixed, stationary \((x, y, z)\) Cartesian coordinate system. The computational domain itself, including the body inside, moves in the fixed \((x, y, z)\)-space. The flow around the body is calculated as the moving boundary problem. Unknown flow variables, such as pressure \(p\), \(x\)-directional velocity \(u\) and so on, are defined at each grid cell center in the computational domain. The motion of the computational domain according to the motion of the body in the physical space is arbitrary, and thus the any kind of the motion of the body can be simulated by the Moving Computational Domain method. The flow field driven by the body is calculated in the computational domain in which the body fitted mesh system is used. Since the computational domain itself moves in the physical \((x, y, z)\) space time-dependently and thus the mesh system of the computational domain also moves in the \((x, y, z)\) space, the flow solver has to be constructed for the moving grid system. In the present Moving Computational Domain method, the Moving-Grid Finite-Volume method [3] is adopted. Only necessary and essential assumption is that the condition in front of the moving computational domain has to be known because it is necessary as a boundary condition of the flow solver. The natural assumption may be the stationary fluid condition in front of the moving computational domain.

### 2.5 Numerical method

To solve the discretization equation of equation (2) using the Trans-mesh method and the Moving Computational Domain method, we apply the SMAC method [5]. Thus, this equation can be solved in three stages. The equation to be solved at the first stage contains the unknown variables at \((n+1)\)-time step in the flux terms. Thus, the equation is iteratively solved using the LU-SGS method [6]. The equation to be solved at the
second stage is the Poisson equation about the pressure correction. This equation is iteratively solved using the Bi-CGSTAB method [7]. The incompressible fluid-rigid body interaction with collision is calculated in the first step of the SMAC method. Figure 4 shows the flowchart of the fluid-body interaction. First, we calculate the force applied to the body. This force includes the force of collision. The collision force is evaluated based on Glowinski’s method [8]. The Transmission Mesh method does not require special treatment when collisions happen. Next, the equation of motion for the body is calculated, and the mesh is moved. Then, the Navier-Stokes equations are calculated. If the physical amounts converge, we proceed to the second step of the SMAC method.

Figure 3: Conventional method (left) and Moving Computational Domain method (right).

Figure 4: Flowchart for calculation of interaction.
3 Numerical Results

3.1 Falling sphere by gravity in the infinite long bending pipe

As an application of the present methods, the movement of a sphere falling by gravity in the infinite long bending pipe is investigated, as illustrated in Fig. 5.

![Figure 5: The moving computational domain for falling sphere in a bending pipe.](image)

The diameter of the pipe is $3.0d$, where $d$ denotes the diameter of the sphere. The wavelength and amplitude of the pipe are $12d$ and $0.75d$, respectively. The moving computational domain is red domain in the pipe shown as Fig. 5. The initial position of the sphere is in the center of the moving computational domain. The ratio between body and fluid density ($\rho_b/\rho_f$) is 1.167. The initial stationary condition of pressure, velocity components in the $x$, $y$, $z$ directions are given by $p = 1.0$, $u = v = w = 0.0$. The Reynolds number is 300 based on the diameter of the sphere $d$, the terminal speed of the sphere $w_\infty$, and the kinematic viscosity. A mesh system of $52 \times 52 \times 104$ is adopted. The time step size is $\Delta t = 0.01$. Table 1 shows setup of physical quantity.

<table>
<thead>
<tr>
<th>$d$ [mm]</th>
<th>$g$ [m/s$^2$]</th>
<th>$\rho_f$ [kg/m$^3$]</th>
<th>$\mu_f$ [Ns/m$^2$]</th>
<th>$w_\infty$ [m/s]</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>9.81</td>
<td>960</td>
<td>$10 \times 10^{-3}$</td>
<td>0.220</td>
<td>300</td>
</tr>
</tbody>
</table>

As a result, Fig. 6 shows the pressure distribution and velocity vectors on central plane of $y$ as well as the sphere position at the time $t = 2.80$ and $3.01$. The sphere moves along
the wall of the pipe after the sphere collides against the wall. It is confirmed that the pressure in the front of the sphere is always higher than the pressure in the rear.

Figure 6: Overall view, pressure distribution, and velocity vectors on central plane of $y$.

Figure 7 shows the computational mesh near the sphere on central plane of $y$. We confirmed that the computational mesh holds its shape well even if the computational mesh and the sphere move in any direction.
3.2 A trajectory of a flying sphere over ground in incompressible fluid

A flying sphere over ground in incompressible fluid, as illustrated in Fig. 8 is presented.

The size of the moving computational domain is $L_x = 21.0d$, $L_y = 21.0d$, $L_z = 21.0d$, where $d$ denotes the diameter of the sphere. The initial stationary condition of pressure, velocity components in the $x$, $y$, $z$ directions are given by $p = 1.0$, $u = v = w = 0.0$. Table 2 shows setup of physical quantity. The size and mass of the sphere are the same as a soccer ball. It is assumed that the sphere is a rigid body. The stitching of the soccer ball is not considered. The Reynolds number is 3000 based on the diameter of the sphere $d$, the initial speed of the sphere, and the kinematic viscosity. A mesh system of $81 \times 81 \times 81$
is adopted. The time step size is $\Delta t = 0.01$.

<table>
<thead>
<tr>
<th>Table 2: Setup of physical quantity.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter of ball</td>
</tr>
<tr>
<td>Mass of ball</td>
</tr>
<tr>
<td>Initial velocity</td>
</tr>
<tr>
<td>Initial angle</td>
</tr>
</tbody>
</table>

As a result, Fig. 9 shows the pressure distribution and velocity vectors on central plane of $y$ as well as the sphere position at the time $t = 0.28$ s, $0.55$ s, $0.82$ s, $1.10$ s, and $1.38$ s. Figure 10 shows the trajectory of the sphere and angular velocity of the sphere where $\omega_x$, $\omega_y$, and $\omega_z$ are the angular velocity component of the $x$, $y$, and $z$ axis of rotation, respectively. The sphere is accelerated at $t = 0.22$ s. The sphere rises with a negative angular velocity of $x$. Next, pressure distribution in a rear of the sphere changes from a symmetric shape to dissymmetric shape as time goes by. The movement in the $y$ direction is not confirmed because the Reynolds number is low and a viscosity is high. The sphere begins to fall at $t = 0.8$ s and the magnitude of the angular velocity becomes small. It is confirmed that the sphere collides into the ground at $t = 1.3$ s.

4 Concluding Remarks

In this paper we introduced the applications using Trans-mesh method and the Moving Computational Domain method and described its application to a falling sphere by gravity in an infinite long bending pipe and a trajectory of a flying ball over ground in incompressible fluid. The results indicated that physically meaningful flows were obtained and the computational mesh holds its shape well even if the computational mesh and the sphere move in any direction. Therefore, we confirmed that the Trans-mesh method and the Moving Computational Domain method are useful for simulating the interaction of incompressible fluid-rigid body.

Acknowledgment

This work was supported by the Science Research Promotion Fund (2010) from the Promotion And Mutual Aid Corporation for Private Schools of Japan and the Grant-in-Aid for Scientific Research (21560175).

REFERENCES

Figure 9: Pressure distribution and velocity vectors on central plane of $y$ at $t = 0.28 \text{s}, 0.55 \text{s}, 0.82 \text{s}, 1.10 \text{s},$ and $1.38 \text{s}$.  

Figure 10: Trajectory of the sphere and angular velocity of the sphere.


WIND TURBINE SIMULATION: STRUCTURAL MECHANICS, FSI AND COMPUTATIONAL STEERING

Y. BAZILEVS*, A. KOROBNKO*, X. DENG*, J. TIPPMAANN* AND M.-C. HSU†

* Department of Structural Engineering
  University of California, San Diego
  La Jolla, CA 92093, USA

† Institute for Computational Engineering and Sciences
  The University of Texas at Austin
  201 E. 24th St. Stop C0200
  Austin, TX 78712, USA

Key words: Fluid–structure interaction, nonmatching interface discretization, isogeometric analysis, composite materials, wind turbine, wind turbine blade, ALE-VMS, rotor-tower interaction, KirchhoffLove shells, DDDAS, fiber waviness

Abstract. A fluid-structure interaction (FSI) validation study of Micon 65/13M wind turbine with Sandia CX-100 composite blades is presented. KirchhoffLove shell theory is used for blade structures, while the aerodynamics formulation is performed using a moving-domain finite-element-based ALE-VMS technique. The structural mechanics formulation is validated through the eigenfrequency analysis of the CX-100 blade. For coupling between two domains a nonmatching discretization of the fluid-structure interface is adopted. This adds flexibility and relaxes the requirements placed on geometry modeling and meshing tools employed. The simulations are done at realistic wind conditions and rotor speeds. The rotor-tower interaction that influences the aerodynamic torque is captured. The computed aerodynamic torque generated by the Micon 65/13M wind turbine compares well with that obtained from on-land field tests. We conclude by illustrating the application of the Dynamic Data-Driven Applications System (DDDAS) to investigate the fiber waviness defects embedded in the CX-100 wind turbine blade.
1 INTRODUCTION

In modern computational analysis of wind turbines, FSI simulations at full scale are important for accurate and reliable modeling, as well as blade failure prediction and design optimization. The wind turbine FSI problem presents significant computational challenges because of the curved geometry of the wind turbine blades, complex distribution of material properties, wind and rotor speeds that lead to complex, wall-bounded turbulent flows, and the presence of mechanical components in relative motion. Because of the challenges involved in the aerodynamics modeling, only a handful of researchers (see, e.g., [15, 17, 24, 34]) performed wind turbine simulations at full scale and with all the turbine components (i.e. blades, tower and nacelle) included. The fully coupled FSI simulations of full-scale wind turbine designs were performed only recently in [1] for the rotor-only case and in [18] for the full machine.

In this paper we present the structural, aerodynamic, and FSI simulations of the Sandia CX-100 blade. Most of the methods and results resented here may be found in [23]. The FSI framework developed in [2] is adopted wherein isogeometric analysis (IGA) [13, 19] is employed for structural mechanics. The aerodynamics formulation makes use of a finite-element-based ALE-VMS technique [9]. The sliding interface formulation developed in [5] employed in conjunction with the ALE-VMS technique [3] enables the simulation of the interaction between the spinning rotor and stationary tower. In addition, for improved boundary-layer accuracy, a key ingredient employed in the ALE-VMS simulation of wind-turbine aerodynamics and FSI are the weakly enforced no-slip boundary conditions [4, 7, 8] set on the moving blade surfaces.

The blade structures are modeled using the rotation-free multilayer composite Kirchhoff–Love shell [1, 21, 22]. Prior to using this blade in our FSI computation, we perform its eigen-frequency analysis and compare the results with the experimental data. This presents the first validation of the rotation-free Kirchhoff–Love IGA shell for a full-scale wind turbine blade. In addition to eigenfrequency analysis, we use the same isogeometric model of the CX-100 blade for the structural mechanics simulations in the context of the Dynamic Data-Driven Applications System (DDDAS) framework [14]. In particular, we focus on in-plane fiber waviness defects, which are typical for multi-layer composite structures, and show how to incorporate such defects, coming from sensor data, into the isogeometric model of the blade.

2 NUMERICAL METHODS AND RESULTS

To show the accuracy and reliability of the presented FSI framework, simulations of the Micon 65/13M wind turbine are carried out at realistic operational conditions, reported in [35]. The turbine is designed as a three-blade, fixed-pitch, upwind turbine with the total rotor diameter of 19.3 m and rated power of 100 kW. The hub is located at the height of 23 m, with a mounting flange positioned 0.6 m from the centerline of the low speed shaft. The wind turbine stands on a tubular steel tower, with a base diameter of 1.9 m. The drive train generator operates at 1200 rpm, while the rotor spins at a nominal speed of 55 rpm. The Micon 65/13M wind turbine was used for the Long-Term Inflow and Structural Testing (LIST) program [27] at the USDA-ARS test facility in Bushland, Texas. This project was initiated by Sandia National Lab-
oratories in 2001 to explore the use of carbon fiber in wind turbine blades. Three experimental blade prototypes, GX-100, CX-100 and TX-100, were developed specifically for this project.

2.1 Structural mechanics

For the simulation of the Micon 65/13M wind turbine, we make use of the Sandia CX-100 carbon-spar blade [12, 35] which is an actual design that has over 32 material zones with complex, nonsymmetric composite layup in each zone (see Figure 1). To model the blade, including the full composite layup, we use a rotation-free Kirchhoff–Love shell formulation [22] discretized using NURBS-based IGA [13, 19]. The Kirchhoff–Love formulation is augmented with the bending strip method proposed in [21].

We perform eigenfrequency calculations of the CX-100 blade using three quadratic NURBS meshes. The coarsest mesh has 1,846 elements, while the nest mesh has 18,611 elements. The mesh statistics are summarized in Table 1. The eigenfrequency results are compared with the experimental data from [25, 33]. We compute the case with free boundary conditions and the case when the blade is clamped at the root. For the free case the eigenfrequencies for the first and second flapwise bending modes and for the first edgewise bending mode are summarized in Table 2. The experimental eigenfrequencies are obtained for this blade at Sandia National Laboratories (SNL), Los Alamos National Laboratory (LANL) and the University of Massachusetts Lowell Structural Dynamics and Acoustics Laboratory (UML SDASL), and reported in [25].

![Figure 1: Left: Five primary sections of the CX-100 blade; Right: 32 distinct material zones of the CX-100 blade.](image_url)

<table>
<thead>
<tr>
<th>Number of Control Points</th>
<th>Number of Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh 1</td>
<td>3,469</td>
</tr>
<tr>
<td>Mesh 2</td>
<td>7,411</td>
</tr>
<tr>
<td>Mesh 3</td>
<td>25,896</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Control Points</th>
<th>Number of Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh 1</td>
<td>1,846</td>
</tr>
<tr>
<td>Mesh 2</td>
<td>4,647</td>
</tr>
<tr>
<td>Mesh 3</td>
<td>18,611</td>
</tr>
</tbody>
</table>

Table 1: NURBS blade meshes used in the eigenfrequency analysis.
Table 2: Comparison of experimentally measured and computed natural frequencies corresponding to the first few bending modes for the free case.

<table>
<thead>
<tr>
<th>Mode:</th>
<th>1st Flapwise (Hz)</th>
<th>1st Edgewise (Hz)</th>
<th>2nd Flapwise (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh 1</td>
<td>8.28</td>
<td>15.92</td>
<td>19.26</td>
</tr>
<tr>
<td>Mesh 2</td>
<td>8.22</td>
<td>15.61</td>
<td>18.21</td>
</tr>
<tr>
<td>Mesh 3</td>
<td>8.22</td>
<td>15.6</td>
<td>18.01</td>
</tr>
<tr>
<td>Experiment</td>
<td>7.6 - 8.2</td>
<td>15.7 - 18.1</td>
<td>20.2 - 21.3</td>
</tr>
</tbody>
</table>

Table 3: Comparison of experimentally measured and computed natural frequencies corresponding to the first few bending modes for the clamped case.

<table>
<thead>
<tr>
<th>Mode:</th>
<th>1st Flapwise (Hz)</th>
<th>2nd Flapwise (Hz)</th>
<th>3rd Flapwise (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh 1</td>
<td>4.33</td>
<td>11.82</td>
<td>19.69</td>
</tr>
<tr>
<td>Mesh 2</td>
<td>4.29</td>
<td>11.61</td>
<td>19.08</td>
</tr>
<tr>
<td>Mesh 3</td>
<td>4.27</td>
<td>11.54</td>
<td>18.98</td>
</tr>
<tr>
<td>Experiment</td>
<td>4.35</td>
<td>11.51</td>
<td>20.54</td>
</tr>
</tbody>
</table>

**Figure 2**: First flapwise bending mode (a) and first edgewise bending mode (b), for the free case.

Table 2 provides a range of experimental eigenfrequency values. For the clamped case, the eigenfrequencies for the first three bending modes are compared with the results of the tests performed at the National Renewable Energy Laboratory (NREL) [25]. In both cases, the computed natural frequencies are in good agreement with the experimental data (see Table 3). The medium mesh shows a good balance between the computational cost and accuracy of the results. For this reason, this mesh is chosen for the FSI computations presented in the next section. The mode shapes computed using the medium mesh are shown in Figures 2 and 3.
2.2 Aerodynamics and FSI

The wind turbine aerodynamics is governed by the Navier–Stokes equations of incompressible flows. The incompressible-flow assumption is valid for the present application because the Mach number is low. The aerodynamics formulation makes use of the FEM-based ALE-VMS approach [3] augmented with weakly enforced boundary conditions [4]. The former acts as a turbulence model, while the latter relaxes the mesh size requirements in the boundary layer without sacrificing the solution accuracy. The framework includes FSI modeling with nonmatching discretization of the interface between the fluid and structure subdomains. We take advantage of the flexibility associated with using nonmatching discretizations, combining the most appropriate discretization for each part of an FSI problem. The use of nonmatching grids also allows for adopting different geometry modeling and meshing tools for the fluid and structural mechanics subproblems. Nonmatching discretizations at the fluidstructure interface require the use of interpolation or projection of kinematic and traction data between the non-matching surface meshes (see, for example, [2, 28, 29]), and that is what we do here.

To simulate the full machine, we use a sliding-interface formulation developed in [5], and recently employed for wind-turbine simulations in [17, 18]. We note that in application of the FEM to flows with moving mechanical components, the Shear–Slip Mesh Update Method [10, 11, 30] and its more general versions [31, 32] may also be used to handle objects in relative motion.

An aerodynamic and FSI simulations of the full Micon 65/13M wind turbine are presented for a constant inflow wind speed of 10.5 m/s and fixed rotor speed of 55 rpm are prescribed. These correspond to the operating conditions reported for the field tests in [35]. The air density and viscosity are set to $1.23 \text{ kg/m}^3$ and $1.78 \times 10^{-5} \text{ kg/(m·s)}$, respectively.

The computations were carried out in a parallel computing environment. The meshes, which consist of linear triangular prisms in the boundary layers and linear tetrahedra elsewhere, are partitioned into subdomains using METIS [20], and each subdomain is assigned to a compute core. The parallel implementation of the methodology may be found in [16]. The time step is set to $3.0 \times 10^{-5} \text{ s}$ for all cases.

In Figure 4 the time history of the aerodynamic torque is plotted. As can be seen from the plot, using FSI, we capture the high frequency oscillations caused by the bending and torsional motions of the blades. In the case of the rigid blade the only high-frequency oscillations in
Figure 4: Aerodynamic torque history for the FSI and rigid-blade simulations. The experimental range for the aerodynamic torque and its average value are provided for comparison and are plotted using dashed lines.

The computed values of the aerodynamic torque are plotted together with field test results from [35]. The upper and lower dashed lines indicate the aerodynamic torque bounds, while the middle dashed line gives its average value. Both the aerodynamic and FSI results compare very
Figure 6: Relative wind speed at the 70% spanwise station for the FSI simulation at $t = 0.86$ s (left) and $t = 1.06$ s (right). The blade deflection is clearly visible.

well with the experimental data.

Figure 5 shows the flow field as the blade passes the tower. Figure 6 shows the relative wind speed at the 70% spanwise station rotated to the reference configuration to illustrate the blade deflection and complexity of boundary-layer turbulent flow.

2.3 DDDAS

DDDAS is a framework in which measurement data collected for a given physical system are used to dynamically update a computational model of that system. Using measurement data, the computational model geometry, boundary conditions, forcing, and material parameters may be updated to better represent physical reality. At the same time, the properly updated computational model is able to produce higher-fidelity outputs for the quantities of interest for which measurements are not readily available. As such, DDDAS is a framework in which measurement and simulation co-exist in a symbiotic environment. Many applications of DDDAS involve not only updating of the computational model on the basis of sensor data, but also adjustment of the model and physical system input parameters to optimize a desired outcome. For example, in [26], the authors developed and deployed a DDDAS framework for computational steering of the laser-guided surgery for prostate cancer treatment. The temperature and location of the laser were dynamically controlled to achieve maximum damage to the cancerous tissue.

In our case, we are interested in a DDDAS framework for large-scale structures exposed to aerodynamic loading [6]. One of the main features of the framework proposed in [6] is the use of IGA to model shell structures made of laminated composites. The use of a single geometry for modeling and simulation, which is one of the tenets of IGA, has several benefits for DDDAS. Here we take advantage of the parametric description of the blade geometry in order to locate the composite fiber waviness defect (detected by the sensors) on the model surface, and to impose the fiber waviness during the through-thickness homogenization process. Because fiber waviness is a local phenomenon, the computational mesh needs to be refined in order to capture the local variation in the strain (and stress) components. Schematic of the DDDAS modeling applied in the context of composite fiber waviness defects is illustrated in Figure 7. The more general DDDAS procedure may be described as a feedback loop, which
Figure 7: DDDAS loop for in-plane fiber waviness.

Figure 8: Local mesh refinement that captured in-plane shear stress variation due to the detected in-plane fiber waviness.

contains the actual structure with embedded sensors that provide feedback to the simulation and modeling/meshing modules. The concentration of the in-plane shear strain after the defect was located and the mesh was refined locally is shown in Figure 8.
3 CONCLUSIONS

We constructed a detailed structural model of the actual wind turbine blade, which has 32 material zones. Each of the material zones was characterized with its distinct composite layup. Nonsymmetric layup was used in all zones, which introduced coupling between the membrane and bending response of the Kirchhoff–Love shell. Despite the discontinuity in the material properties, almost everywhere $C^1$-continuous quadratic NURBS basis was used to discretize the shell kinematics, and the NURBS mesh lines were not aligned with the boundaries of the material zones. The material properties were approximated at the quadrature points of the NURBS mesh. With this construction, we were able to reproduce the experimentally measured eigenfrequencies of the CX-100 blade. To our knowledge, this is the first full-scale validation of the IGA-based thin-shell composite formulation.

We used the previously developed ALE-VMS formulation for modeling the turbulent flow, FSI coupling of FEM and IGA, weakly enforced boundary conditions, and a sliding interface formulation to compute the full Micon 65/13M wind turbine with the CX-100 blades mounted on its rotor. The results of the aerodynamic and FSI simulations are in a good agreement with field test data for this wind turbine. The FSI simulation captures high-frequency oscillations in the aerodynamic torque, which are caused by the blade structural response.

Finally, we illustrated the benefits of using IGA within a recently proposed DDDAS framework for in-plane fiber waviness calculations in composites.

ACKNOWLEDGEMENTS

The support of the AFOSR Award No. FA9550-12-1-0005 is gratefully acknowledged. X. Deng was supported through AFOSR Award No. FA9550-12-1-0046. We thank the Sandia National Laboratories Wind Energy Technologies program for providing us access to the data, which enabled the construction of the wind turbine blade structural model. We also thank the Texas Advanced Computing Center (TACC) at the University of Texas at Austin and the San Diego Supercomputing Center (SDSC) at the University of California, San Diego for providing HPC resources that have contributed to the research results reported in this paper.

REFERENCES


A PARALLEL, BLACK-BOX COUPLING ALGORITHM FOR FLUID-STRUCTURE INTERACTION

BENJAMIN UEKERMANN*, HANS-JOACHIM BUNGARTZ†, BERNHARD GATZHAMMER† AND MIRIAM MEHL*

*Institute for Advanced Study (IAS), Technische Universität München
Lichtenbergstraße 2a, 85748 Garching b. München, Germany
e-mail: {uekerman,mehl}@in.tum.de, web page: http://www5.in.tum.de

†Institut für Informatik, Technische Universität München
Boltzmannstraße 3, 85748 Garching b. München, Germany
e-mail: {bungartz,gatzhammer}@in.tum.de, web page: http://www5.in.tum.de

Key words: Fluid-Structure Interaction, Strong Coupling, Black-Box Coupling, Partitioned Coupling, High Performance Computing, Scientific Computing

Abstract. The simulation of multi-physics scenarios, in particular fluid-structure interaction has gained more and more importance in the last years due to increasing accuracy requirements for a large range of applications from biomedical fields to technical design problems. At the same time, this type of simulation has become feasible due to the increased computing power of modern supercomputers. Note that only the combination of a highly accurate and, thus, highly resolved, discretization with the multi-physics model yields more detailed and more realistic results than a simple single-physics simulation. However, modern computing architectures require a good scalability of simulation methods on massively parallel systems. For fluid-structure interactions, if done in a partitioned way using separate fluid and structure codes, in particular the usually applied staggered scheme executing fluid and structure solver one after the other hinders a good scalability. This is due to the in general largely different computational costs of the two solvers. In this paper, we propose two new coupling schemes for an implicit coupling of black-box fluid and structure solvers that execute the two solvers in parallel and still yield good convergence and stability even for incompressible fluids which is shown by means of numerical results for the flow through a flexible tube.

1 INTRODUCTION

In the last years, modern computational development provided a computational power that allowed to tackle complex multi-physics simulations furthering a new range of industrial and medical applications such as interactions between aircraft wings and the
surrounding turbulent flow or between blood flow and arterial walls. However, the highly parallel architecture of modern supercomputers, at the same time, poses new challenges that require the development of new numerical algorithms in order to exploit their computational power. The focus of this paper is the development of new numerical algorithms for the example of fluid-structure interactions, where a fluid flow exerts forces on a structure and the structural movement changes the flow field such that we have a bidirectional strong interaction between flow and structural mechanics.

1.1 The modular approach: partitioned black-box coupling

To achieve a high flexibility and modularity of the fluid-structure interaction simulation environment, we focus on a special type of coupling, the so-called partitioned approach. The coupling between different physics in this case takes place at the highest possible level. In fluid-structure interaction (FSI), this is done via boundary conditions at the fluid-structure interface (wet surface). The most common technique is the Dirichlet-Neumann coupling prescribing displacements and velocities as boundary conditions for the flow solver and forces for the structure solver. The counterpart of partitioned methods, monolithic schemes, in contrast, couple at a very low level. They assemble one large system of equations that is solved as a whole and use separate fluid and structure solvers only on a preconditioner level (see e.g. [1]). The disadvantages of partitioned coupling are possible stability issues (cf. e.g. [2]), but this effort is worth taking. By separating the physical regimes, we can reuse existing solver codes and, thus, benefit from the experiences which have been made in each field. Furthermore, non-linear and linear solvers can be better tuned to the particular characteristics each physical system adheres. The clearest advantage, however, of partitioned approaches is the separation of codes. With this approach, solver codes can easily be exchanged permitting the integration of existing codes in a plug-and-play manner. To maximize the flexibility in exchanging codes, we particularly aim at using black-box solvers offering the same technical interface depth as standard commercial solvers usually provide. Thus, the coupling must be formulated independently of the spatial and time discretization of each solver as this is hidden in the black-box. Also Jacobian information is in general not available. Our group developed the coupling environment \textit{preCICE}\(^1\), which implements a partitioned black-box coupling, data mapping functionality for non-matching grids, and communication routines. It allows a very flexible and easy exchange of solver codes (cf. [3]).

1.2 Efficient parallelization

Besides numerical coupling methods sketched above, efficient parallelization is the key to highly accurate simulations\(^2\). However, a load balanced parallelization is not trivial

\(^1\)http://www5.in.tum.de/wiki/index.php/PreCICE_Webpage

\(^2\)Note that only a high accuracy in the discretization and solution of multi-physics models accounts for their inherent accuracy.
due to the physical regimes' cost-asymmetry. Structure calculations typically require far less computational power and, therefore, do not scale well on a high number of processors. Thus, if both physics are solved in a staggered fashion as shown in Fig. 1 (top), i.e., one solver always precedes the other solver, the structure part is a scalability bottleneck. Monolithic schemes can overcome this problem. However, we focus on a black-box partitioned approach, where we develop new numerical algorithms for a parallel, stable, and efficient parallel fluid-structure coupling.

For compressible flow, there are explicit coupling schemes known that allow for a high parallel efficiency. This is, e.g., accomplished by a cross-staggered scheme ([4]) or by means of an interface system ([5]) all requiring only one execution of fluid and structure solver per time step resulting in a so-called explicit coupling. Incompressible flow, however, in general needs implicit coupling, i.e., an iterative coupling loop over fluid and structure solution steps during every timestep. This is due to the different nature of the added mass effect in the incompressible case inducing well-known stability issues ([2, 6, 7]). A straightforward generalization of the classical explicit parallel methods in [4] to implicit coupling leads to redundant computations, whereas the methods in [5] require discretization details in the vicinity of the wet surface and, thus, conflict with black-box solvers. Classical implicit black-box approaches follow the staggered scheme. In the simplest form, they use (Aitken-) underrelaxation (see e.g. [8]) to achieve stability. Degroote et al. [9] describe a black-box algorithm that instead solves an interface equation by means of a quasi-Newton approach (IQN), which behaves like an implicit time-integration scheme in an ODE context. The theoretical background of this method was discussed in [10]. Similar methods have been proposed in [11], [12] and [13]. The drawback of all mentioned implicit coupling algorithms is, however, that fluid and structure are still solved in a staggered way causing a bad parallel efficiency.

In this paper we present and discuss two implicit coupling algorithms that simulate structure and fluid simultaneously, leading to a high parallel efficiency. We test these methods by means of a simple one-dimensional toy problem, the flow through a flexible tube. The structure of this paper is as follows. In Sect. 2, the two new parallel, implicit coupling algorithms will be introduced. In Sect. 3, we describe the one-dimensional application example that we use to test our new methods, covering the analytical description as well as the discretization technique. Numerical results for this example are shown in Sect. 4. We close the paper with conclusions and an outlook in Sect. 5.

2 TWO PARALLEL IMPLICIT COUPLING ALGORITHMS

In the following, we introduce our new parallel implicit coupling algorithms in two steps. First, we revisit the quasi-Newton least square (QNLS) method described in [10]. Then, we use this method to solve different abstract fixed point equations leading us to the new coupling algorithms. To close the section, we discuss different convergence criteria.
2.1 Quasi-Newton least squares method

We start from a non-linear fixed point equation, find \( x \in D \subset \mathbb{R}^n \), such that
\[
H(x) = x,
\]
where \( H : D \to \mathbb{R}^n \) is sufficiently smooth and has a unique fixed point \( x^* \in D \). The Jacobian of the residual \( R(x) = H(x) - x \) at \( x^* \), i.e. \( R'(x^*) = H'(x^*) - I \), is assumed to be non-singular. We approximate the solution \( x^* \) in an iterative quasi-Newton manner, i.e., we solve an approximation of the linearized equation
\[
R'(x^k)(x - x^k) = -R(x^k)
\]
in every iteration. As we want to be able to use black-box solvers in our fluid-structure simulations, we have no access to the Jacobian in every iteration. As we want to be able to use black-box solvers in our fluid-structure simulations, we have no access to the Jacobian
\[
V^k = [\Delta R^k_0, \ldots, \Delta R^k_{k-1}]
\]
with \( \Delta R^k_i = R(x^k) - R(x^i), i = 0 \ldots k - 1 \), and
\[
W^k = [\Delta x^k_0, \ldots, \Delta x^k_{k-1}]
\]
with \( \Delta x^k_i = x^k - x^i, i = 0 \ldots k - 1 \).

To set \( x^{k+1} \), we construct \( \Delta x^k := x^{k+1} - x^k \) in the column space of \( W_k \); \( \Delta x^k = W_k \alpha \) such that (2) is solved as accurately as possible. For this purpose, we solve a least squares problem \( \alpha = \arg\min_{\beta \in \mathbb{R}^k} \| V^k \beta + R(x^k) \|_2 \), where \( \| \cdot \| \) denotes the Euklidian norm in \( \mathbb{R}^n \).

To solve this least squares problem, we compute the decomposition \( V^k = Q^k U^k \) of \( V^k \) into an orthogonal \( n \times n \) matrix \( Q^k \) and an upper triangular matrix \( U^k \in \mathbb{R}^{n \times k} \) and compute \( \alpha \) from the quadratic \( k \times k \) system
\[
\tilde{U}^k \alpha = -\left( \hat{Q}^k \right)^T R(x^k),
\]
where \( \tilde{U}^k \) denotes the first \( k \) rows of \( U^k \) and \( \hat{Q}^k \) contains the first \( k \) columns of \( Q^k \).

With \( \alpha \), we can approximate the solution of (2) as
\[
x^{k+1} - x^k := W^k \alpha
\]
since this gives
\[
-R(x^k) \approx V^k \alpha = \sum_{i=0}^{k-1} \alpha_i \Delta R_i^k = R'(x^k) \sum_{i=0}^{k-1} \alpha_i \Delta x^k_i = R'(x^k) W^k \alpha = R'(x^k)(x^{k+1} - x^k).
\]

As the quasi-Newton method requires at least one column in \( V^k \) and \( W^k \), we have to use simple underrelaxation for the first iteration: \( x^1 = x^0 + 0.1 \cdot R(x^0) \). However, with this
approach, we get a new iterate \( x^{k+1} \) that is a linear combination of \( x^0, \ldots, x^k \) and, as a consequence, we get quasi-Newton iterates in the two-dimensional space span \( \{x^0, x^1\} \).

For this reason, we modify the method a little: We replace \( W_k \) in (4) by

\[
W^k = [\Delta \tilde{x}^k_0, \ldots, \Delta \tilde{x}^k_{k-1}] \quad \text{with} \quad \Delta \tilde{x}^k_i = \tilde{x}^k_i - \tilde{x}^i, \ i = 0 \ldots k - 1.
\]

This corresponds to approximating the solution of

\[
(R \circ H^{-1})' (\tilde{x}^k) (x^{k+1} - \tilde{x}^k) = -R \circ H^{-1} (\tilde{x}^k) = -R (x^k)
\]

instead of (2). This results in a method that performs first a fixed point iteration step \( x^k \rightarrow \tilde{x}^k = H (x^k) \) and afterwards a quasi-Newton step \( \tilde{x}^k \rightarrow x^{k+1} \) approximately solving (8) with (5) and (6) with \( \tilde{x}^k \) instead of \( x^k \) and \( R \circ H^{-1} \) instead of \( R \). See Algorithm 1 for the complete algorithm.

**Algorithm 1** Quasi-Newton least squares method in pseudocode

```plaintext
initial value \( x^0 \)
\( \tilde{x}^0 = H(x^0) \) and \( R^0 = \tilde{x}^0 - x^0 \)
\( x^1 = x^0 + 0.1 \cdot R^0 \)

for \( k = 1 \ldots \) do
    \( \tilde{x}^k = H(x^k) \) and \( R^k = \tilde{x}^k - x^k \)
    \( V^k = [\Delta R^k_0, \ldots, \Delta R^k_{k-1}] \) with \( \Delta R^k_i = R^i - R^k \)
    \( W_k = [\Delta \tilde{x}^k_0, \ldots, \Delta \tilde{x}^k_{k-1}] \) with \( \Delta \tilde{x}^k_i = \tilde{x}^i - \tilde{x}^k \)
    decompose \( V^k = QU^k \)
    solve the first \( k \) lines of \( U^k \alpha = -Q^k R^k \)
    \( \Delta \tilde{x} = W \alpha \)
    \( x^{k+1} = \tilde{x}^k + \Delta \tilde{x} \)
end for
```

In a transient fluid-structure context, i.e., if we perform quasi-Newton iterations in every timestep, the initial value \( x^0 \) for the current timestep can be extrapolated. Furthermore, the reuse of iteration values \( \tilde{x}^i \) and \( R(x^i) \) from previous timesteps in order to construct \( V \) and \( W \) accelerates the convergences (compare [9] and Sect. 4).

### 2.2 Coupling variants induced by fixed point equations

With the general formulation of the QNLS method in Sect. 2.1, different coupling algorithms can be constructed by formulating the corresponding fixed point equations. For this sake, we use an abstract notation for the fluid and structure solver action on the wet surface between fluid and structure. This notation is typically used for the description of partitioned approaches (see e.g. [14]) and can be stated independently of the solver.

\footnote{At the same time, also the columns of \( V^k \) are 'almost parallel', i.e., the QU-decomposition becomes ill-conditioned.}
details. Let $F : d \mapsto f$ be the mapping of interface displacements to interface forces induced by the fluid solver and $S : f \mapsto d$ the mapping of forces to displacements induced by the structure solver, i.e., force and displacements at the wet surface are denoted by $f$ and $d$, respectively.

Figure 1 shows schematic views of different coupling algorithms together with the corresponding fixed point equations.

Figure 1: Coupling variants and their corresponding fixed point equations. (9) yields the well-known staggered interface quasi-Newton (IQN) approach (cf. [9]). (10) and (11) define our new algorithms, which we call parallel interface quasi-Newton (PIQN) and vectorial interface quasi-Newton (VIQN).

The fixed point equation (9) yields the classical interface quasi-Newton (IQN) algorithm described in [9]. This algorithm evaluates $F$ and $S$ in serial causing a bad parallel efficiency as described in Sect. 1. A good parallel performance can only be achieved if $F$ and $S$ can be evaluated in parallel as in (10) and (11).

The method in (11) (which we call vectorial interface quasi-Newton or VIQN) is closely related to the parallel explicit method in [4]. Both solvers have the usual input (displacements and velocities for the fluid solver and forces for the structure solver). After one iteration of both solvers, the output boundary values are exchanged and the next iteration is started. Doing this without the quasi-Newton least squares step after each iteration would lead to two disconnected instances of the usual staggered iterations. However, the quasi-Newton steps introduce a connection. Section 4 shows that this connection is indeed strong enough to make one iteration of VIQN comparable to one iteration of IQN.

The second parallel coupling scheme (10) (called parallel interface quasi-Newton or PIQN) uses the inverse interface structure operator $S^{-1} : d \mapsto f$. For the structure solver, this inversion yields no higher costs, standard commercial solvers allow for this inverse formulation. Thus, both solvers compute forces at the interface as an output. If the iterations are converged, these two forces have to be the same. Thus, we actually solve the equation $(S^{-1} - F)(d) = 0$ to determine the new interface displacements used as an input for the next iteration.

2.3 Convergence criterion

The choice of a particular convergence criterion to compare the different algorithms presented in the last subsection is a delicate issue. The QNLS algorithm as itself induces
the relative criterion
\[ \| \tilde{x}^k - x^k \| < tol \cdot \| x^k \|. \] (12)

For the IQN algorithm, this criterion checks only for the convergence of the displacement values where for VIQN displacement as well as force values are checked. Furthermore, for VIQN, an equal weight of both physical values is not guaranteed. Therefore, we prefer a physically motivated approach to compare the different coupling algorithms with a uniform convergence criterion:
\[ \| \tilde{d}^k - d^k \| < tol \cdot \| d^k \| \text{ and } \| \tilde{f}^k - f^k \| < tol \cdot \| f^k \|. \] (13)

3 PROBLEM DESCRIPTION

In this section, we shortly describe the flow through a flexible tube which we use for testing our new coupling algorithms. The model and the description follow [14].

3.1 Analytical description

The fluid flow in our example is incompressible and inviscid, and no gravity is assumed. Averaging over the tube in radial directions yields a one-dimensional model as displayed in Fig. 2. The conservation of momentum and mass read
\[ \frac{\partial}{\partial t} (au) + \frac{\partial}{\partial x} (au^2) + a \frac{\partial}{\partial x} p = 0 , \] (14)
\[ \frac{\partial}{\partial t} a + \frac{\partial}{\partial x} (au) = 0 , \] (15)

where \( u \) is the flow velocity in axial direction, \( p \) the kinematic pressure, and \( a \) the cross section area of the tube. We impose a time-varying inlet velocity and a non-reflecting outlet boundary condition:
\[ u_{in} = u_0 - \frac{u_0}{100} \sin^2(\pi \frac{t}{T}) \text{ and } \frac{\partial}{\partial x} u = \frac{1}{c} \frac{\partial}{\partial t} p , \] (16)

where \( c^2 := \frac{a}{\rho a} = c_{mk}^2 - \frac{p}{2} \) is the wave speed, \( c_{mk} = \sqrt{Eh/2\rho_0} \) the Moens-Korteweg wave speed and \( E \) Young’s modulus.

The elastic wall is described by a Hookean constitutive law. Since the inertia of the tube wall is neglected, the wall contains no mass. This fact in combination with the incompressibility of the fluid constitutes a severe test for coupling algorithms, because it tightens the instabilities induced by the added mass effect, which depends on the density relation between fluid and structure. The inviscid fluid exerts only stress in circumferential direction on the structure leading to a purely radial motion of the tube wall. As a consequence, the cross section area becomes an explicit function of the pressure:
\[ a = a(p) = a_0 \left( \frac{p_0 - 2c_{mk}^2}{p - 2c_{mk}^2} \right)^2 \] (17)
Figure 2: Schematic drawing of the deformed tube with flow in $x$-direction, inner radius $r = r(x)$, length $L$, and wall thickness $h$. The fluid pressure $p(x)$ acting on the inner tube walls is causing scalar circumferential stresses $\sigma_{\phi\phi}(x)$ in the tube walls which lead to a deformation in radial direction. The test scenario is based on [14].

with $a_0$ and $p_0$ being reference values. For details of the model, refer to [14]. Thus, for this scenario, the general coupling variables force $f$ and displacement $d$ are substituted by the pressure $p$ and the cross section area $a$. We simulate the scenario over a time interval $[0; T]$, a full period of the inlet velocity.

3.2 Discretization

The tube of length $L$ is subdivided into $N_S$ equidistant cells of length $\Delta x = L/N_S$ with values $u_i, p_i$ and $a_i$ assigned to the cell centers$^4$. Equation (14) and (15) are discretized with finite volumes using central discretization schemes in combination with a first-order upwind scheme for the convective term in the momentum equation. The central discretization scheme of the pressure is stabilized by a pressure stabilization with coefficient $\alpha = \frac{a_0}{u_0 + \Delta x / \Delta t}$. For time discretization of the fluid equations, we use a backward Euler scheme and subdivide $[0, T]$ into $N_T$ equidistant slots of length $\Delta t$. The discretized system yields for all spatial segments $i = 1, \ldots, N - 1$

$$\frac{\Delta x}{\Delta t} \left( u_i^{(n+1)} a_i^{(n+1)} - u_i^{(n)} a_i^{(n)} \right) + \left[ u_i u_{i+\frac{1}{2}} a_{i+\frac{1}{2}} - u_{i-\frac{1}{2}} u_{i-\frac{1}{2}} a_{i-\frac{1}{2}} \right]^{(n+1)}$$

$$+ \frac{1}{2} \left[ a_{i+\frac{1}{2}} (p_{i+1} - p_i) + a_{i-\frac{1}{2}} (p_i - p_{i-1}) \right]^{(n+1)} = 0 , \quad (18)$$

$$\frac{\Delta x}{\Delta t} \left( a_i^{(n+1)} - a_i^{(n)} \right) + \left[ u_{i+\frac{1}{2}} a_{i+\frac{1}{2}} - u_{i-\frac{1}{2}} a_{i-\frac{1}{2}} \right]^{(n+1)} - \alpha \left[ p_{i-1} - 2p_i + p_{i+1} \right]^{(n+1)} = 0 . \quad (19)$$

The subscript $i \pm 1/2$ indicates the values calculated at the cell interfaces, e.g., $u_{i-1/2} = 0.5 \cdot (u_{i-1} + u_i)$. The superscripts $(n)$ and $(n + 1)$ refer to time $t^{(n)} = n \cdot \Delta t$ and $t^{(n+1)} = (n + 1) \cdot \Delta t$, respectively. The pressure value at the inlet and the velocity value at the outlet are linearly extrapolated ($p_{in} = 2p_1 - p_2$ and $u_{out} = 2u_N - u_{N-1}$), whereas the pressure-outlet condition is discretized as

$$p_{out} = 2 \left[ c_{mk}^2 - \left( \sqrt{2} c_{mk} \frac{p_{out}^2}{2} - \frac{u_{out}^2 - u_{out}^n}{4} \right) \right] . \quad (20)$$

$^4$This means, we use the same grid for the fluid and the structure solver.
The velocity values $u_i$, the pressure values $p_i$, and the cross section area values $a_i$ are initially set to their reference values $u_0$, $p_0$, and $a_0$.

To achieve comparable results, we follow the idea of [14] and define the dimensionless structural stiffness $\kappa = (\sqrt{Eh/2\rho r_0} - p_0/2)/u_0$ and the dimensionless timestep size $\tau = u_0\Delta t/L$ as physical parameters of our scenario. We fix $N_S = 100$ and $N_T = 100$. The stability analyses in [14], as well as in [2] and [7] show that achieving stability for the FSI coupling gets more challenging with decreasing structural stiffness $\kappa$ or decreasing timestep size $\tau$.

4 NUMERICAL RESULTS

In this section, we discuss the numerical results obtained for our new parallel coupling algorithms introduced in Sect. 2 if applied to the problem from Sect. 3. Figure 3 shows results for the cross section area for different timesteps. All results presented in this paper yield the same values for $u$, $p$, and $a$ as a quasi-monolithic approach (for details see [14]) up to an $l_2$-error of $10^{-7}$.

For the performance discussion of the different algorithms, we assume that all interface calculations are negligible in terms of computational costs compared to the fluid and the structure solver calculations. This assumption holds for real applications, where, in contrast to the toy problem used in this paper, the fluid and the structure domain have a higher dimensionality than the wet surface. Results in [9] confirmed this assumption. As a consequence, the number of solver calls, hence the number of coupling iterations, is a sufficient estimator of the computational costs. We use the convergence criterion discussed in Sect. 2.3 in an $l_2$-sense with $tol = 10^{-7}$.

We test the two new algorithms PIQN and VIQN introduced in Sect. 2 against the IQN algorithm of [9] for different values of the dimensionless timestep size $\tau$ and the dimensionless structural stiffness $\kappa$. By IQN(5), e.g., we denote the IQN algorithm with
the reuse of the values from the last 5 timesteps. Table 4 shows the average numbers of resulting coupling iterations. The optimal number of reused timesteps depends on the

Table 1: Average coupling iteration numbers for the different fixed point equations and different solver parameters $\tau$ and $\kappa$. For each method, the average iteration numbers with no reuse of vectors are shown in the top half of the table. The lower half shows the results using reuse of previous vectors. Here, only the best configuration are presented, which are the reuse of 5 previous steps for IQN, respectively 8 and 5 for VIQN and PIQN.

<table>
<thead>
<tr>
<th></th>
<th>IQN</th>
<th>VIQN</th>
<th>PIQN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau \backslash \kappa$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>0.1</td>
<td>3.96</td>
<td>4.07</td>
<td>5.59</td>
</tr>
<tr>
<td>0.01</td>
<td>3.97</td>
<td>5.01</td>
<td>9.19</td>
</tr>
<tr>
<td>0.001</td>
<td>5.00</td>
<td>9.09</td>
<td>28.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IQN(5)</th>
<th>VIQN(8)</th>
<th>PIQN(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau \backslash \kappa$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>0.1</td>
<td>2.99</td>
<td>3.04</td>
<td>3.25</td>
</tr>
<tr>
<td>0.01</td>
<td>3.02</td>
<td>3.09</td>
<td>3.58</td>
</tr>
<tr>
<td>0.001</td>
<td>3.07</td>
<td>3.28</td>
<td>6.83</td>
</tr>
</tbody>
</table>

parameters, but is in general higher for VIQN than for IQN and PIQN. For our test scenario, the reuse of iterations from 8 previous timesteps results, for most parameter settings, in the smallest number of iterations for VIQN, whereas for IQN and PIQN, the reuse of 5 timesteps was optimal. The number of iterations increases, in general, with decreasing dimensionless timestep size $\tau$, and decreasing dimensionless structural stiffness $\kappa$. PIQN shows competitive results, but slightly worse than IQN and VIQN. VIQN needs, in most cases, even less coupling iterations than IQN. The time evolution of the iteration numbers clarifies this fact. Figure 4 shows the iteration numbers for the first 15 timesteps of each method for one exemplary parameter setting. In the first timesteps, IQN needs less iterations than VIQN, but this changes quickly after 3 timesteps. In principle, the fixed-point equation behind IQN is more suited for the coupling, since the force variables $f$ and the displacement variables $d$ are more interconnected compared to the VIQN fixed-point equation. VIQN, however, benefits more from the reuse of previous timesteps because force values can also be reused instead of just displacements.

A better understanding of the difference between VIQN and IQN can be obtained by looking at the absolute errors (compare Fig. 5). VIQN shows a worse convergence than IQN while VIQN(5) outperforms IQN(5). Moreover, VIQN shows smaller absolute errors already in the first iterations due to the extrapolation of the pressure values - for IQN only the displacements are extrapolated. Furthermore, a small oscillation of the VIQN residuals without the reuse of previous timesteps can be observed, due to the underrelaxation step in the first iteration yielding a shifted influence on the pressure $p$ and cross section area.
values $a$. As the reuse of previous timesteps supersedes the starting underrelaxation step, this feature disappears for VIQN($n$) with $n > 0$.

![Figure 5: Absolute errors $\|a_k - a^20\|_\infty$, $\|p_k - p^20\|_\infty$ of different coupling algorithms. On the left side, the simple coupling algorithms are shown, while on the right side, the values from the previous 5 timesteps are reused. The dimensionless timestep size for this run is $\tau = 0.001$ and the dimensionless structural stiffness is $\kappa = 10$. We used a stricter convergence criterion for this simulations to show a longer development of the errors. All residuals are plotted exemplary for timestep 30.](image)

5 CONCLUSIONS AND OUTLOOK

We have presented two new algorithms exploiting the potential of interface quasi-Newton methods for a parallel instead of a staggered coupling in fluid-structure interaction simulations. These methods are important steps towards scalable massively parallel fluid-structure simulations. The numerical results for an incompressible flow through an massless flexible tube showed the large potential of our methods. In particular, the VIQN method that has the additional advantage of not changing the boundary conditions used for fluid and structure in common staggered Dirichlet-Neumann methods, proved to be very efficient. The performance of the PIQN method is slightly worse. It turned out that for parallel coupling the reuse of information from previous timesteps is of crucial importance. The choice of an optimal number of steps to reuse is, however, still an open issue and subject of future work as well as the implementation of the method for more general two- and three-dimensional examples and supercomputing architectures.

Acknowledgement The financial support of the Institute for Advanced Study (IAS) of the Technische Universität München is thankfully acknowledged.

REFERENCES


A CONSTITUTIVE MODEL OF COUPLED THERMOELASTICITY WITH PLASTICITY

R. BARRETTA*, F. MAROTTI DE SCIARRA† AND M. SALERNO+

* University of Naples Federico II, Department of Structures for Engineering and Architecture
Via Claudio 21 80121 Naples, Italy
e-mail: rabaret@unina.it

† University of Naples Federico II, Department of Structures for Engineering and Architecture
Via Claudio 21 80121 Naples, Italy
e-mail: marotti@unina.it

+ University of Naples Federico II, Department of Structures for Engineering and Architecture
Via Forno Vecchio 36 80134 Naples, Italy
e-mail: maria.salerno@unina.it

Key words: Thermoelasticity without energy dissipation, Thermoplasticity, Constitutive model, Thermodynamic potentials.

Abstract. A consistent set of the thermodynamic functions in the framework of the Green and Naghdi (GN) coupled thermoelasticity with plasticity is addressed. The relations between these functions are provided and the constitutive relations are obtained.

1 INTRODUCTION

It is well-known that the classical linear theory of heat conduction, based on Fourier's law for the thermal flux, predicts that a thermal effect at a point of a body is felt instantly at other points of the body. Therefore in past years several alternative theories of heat conduction have been proposed. In this paper we consider the non-classical theory of thermoelasticity developed by Green and Naghdi [1,2] which incorporates the approach based on Fourier's law (referred to as type I), the theory without energy dissipation (type II) and a theory which allows finite wave propagation as well as energy dissipation (type III). Contributions on the Green and Naghdi (GN) approach can be found, among others, in Bargmann and Steinmann [3] and references therein.

An analytical treatment of coupled thermoelastic problems is complex so that the development of alternative methods of analysis turns out to be important. Accordingly variational formulations of such problems can be of great interest from a theoretical point of view and from a computational standpoint since they are the foundation to develop mixed finite elements.
In the present paper we treat the non-classical thermoelastic problem following the GN model of type II. The purpose of the contribution is to provide a constitutive model of thermoelasticity coupled with plasticity suitable to formulate variational formulations (see Romano et al. [4-6] and Marotti de Sciarra [7-9] for nonlocal problems), also for functionally graded materials [10-13], in order to consistently derive a finite element [14-15] or boundary element [16] approach and the related algorithmic procedure.

The GN coupled thermoelastic model with plasticity is formulated within the framework of generalized standard material (Halphen and Nguyen [17]).

Moreover we adopt the hypotheses of small strains, additive plasticity and rate-dependent plasticity (Bartels and Roubíček [18]), whereas we get the assumption that the elastoplastic behaviour of the material does not influence the thermoelastic constitutive properties, i.e. the thermal behaviour in all the elastic ranges. (Bertram and Krawietz [19]).

The relations between the thermodynamic functions and their alternative formulations are obtained. A consistent set of the eight thermodynamic functions has been derived into the framework of convex analysis and conjugate functions.

In the considered GN model, the thermodynamic functions depend on three state variables where the dual set of constitutive state variables are given by strain, stresses, temperature, entropy, gradient of thermal displacement and entropy flux.

Using a systematic procedure based on Legendre transforms, the thermodynamic potentials are expressed in terms of different combinations of the abovementioned state variables obtaining a set of eight alternative functions. A characteristic feature of the proposed approach is that the derivatives of the thermodynamic potentials and of their alternative forms provide different expressions of the constitutive relations which turn out to be all equivalent each others.

2 CLASSICAL THERMODYNAMIC FRAMEWORK

Let us consider the classical thermodynamic framework of the thermo-mechanical processes for an elastoplastic body. The constitutive model follows from the first principle of thermodynamics, expressed in pointwise form of the internal energy $U$:

$$U = \varepsilon : \sigma + \dot{Q} + r$$  

where $\varepsilon, \sigma$ denote the strain and the stress tensors, $Q = -\text{div} q$ is the heat supply from conduction, being $q$ the heat flux, and $r$ is the heat supply from irradiation.

The second principle of thermodynamics is assumed in form of Clausius-Duhem inequality:

$$\eta - \left( \frac{r}{\theta} - \frac{\text{div} q}{\theta} \right) \geq 0$$  

where $\eta$ is the internal entropy production rate per unit volume and $\theta$ is the absolute temperature ($\theta > 0$).

The thermodynamic eqs. (1) and (2) yield the non-negative total dissipation:
The inequality (3) is employed to derive some restrictions on the constitutive equations. The thermo-elastoplastic solid is assumed to have a saddle internal energy (convex in terms of elastic strain and internal variables and concave in terms of the elastic entropy), differentiable with respect to the arguments, as

$$U = U(e, \alpha_1, \alpha_2, \eta_e)$$  \hspace{1cm} (4)

where the plastic behaviour with hardening is introduced by a suitable set of kinematic internal variables \((\alpha_1, \alpha_2)\) that describe the kinematic and isotropic hardening. The dual static internal variables are \((\chi_1, \chi_2)\).

Moreover, following a consolidated procedure, the total strain \(\varepsilon\) is assumed to be the sum of an elastic strain \(\varepsilon_e\) and of a plastic strain \(\varepsilon_p\) (Coleman and Owen [20]).

Expanding the derivative of \(U(e, \alpha_1, \alpha_2, \eta_e)\) and substituting in (3), we can write the overall dissipation:

$$D = (\sigma - \varepsilon_e \cdot \dot{\varepsilon}) \cdot \dot{\varepsilon} + \sigma \cdot \dot{\varepsilon}_p - \partial_{\varepsilon} U \cdot \dot{\alpha}_1 - \partial_{\alpha_1} U \cdot \dot{\alpha}_2 - \partial_{\eta_e} U \cdot \dot{\eta}_e + \dot{\eta} \theta - \nabla \theta \cdot \frac{q}{\theta} \geq 0$$ \hspace{1cm} (5)

that leads to

$$\sigma = \partial_{\varepsilon} U(e, \alpha_1, \alpha_2, \eta_e)$$ \hspace{1cm} (6)

$$\chi_1 = \partial_{\alpha_1} U(e, \alpha_1, \alpha_2, \eta_e) \quad \chi_2 = \partial_{\alpha_2} U(e, \alpha_1, \alpha_2, \eta_e)$$

$$\theta = \partial_{\eta_e} U(e, \alpha_1, \alpha_2, \eta_e)$$

where \(\partial_{\varepsilon} U\) denotes the partial derivative of \(U\) with respect to \(\varepsilon\) and similarly the other ones.

Then the inequality (3) reduces to following form

$$D = \sigma \cdot \dot{\varepsilon}_p - \chi_1 \cdot \dot{\alpha}_1 - \chi_2 \cdot \dot{\alpha}_2 - \theta \cdot \dot{\eta}_e + \dot{\eta} \theta - \nabla \theta \cdot \frac{q}{\theta} \geq 0$$

Accordingly, introducing the following additive decomposition of the total entropy \(\eta\) in elastic and plastic parts, it turns out to be:

$$\eta = \eta_e + \eta_p$$ \hspace{1cm} (7)

so then it follows

$$D = \sigma \cdot \dot{\varepsilon}_p - \chi_1 \cdot \dot{\alpha}_1 - \chi_2 \cdot \dot{\alpha}_2 + \theta \cdot \dot{\eta}_p - \nabla \theta \cdot \frac{q}{\theta} \geq 0$$ \hspace{1cm} (8)

where \(\dot{\eta}_p\) is the rate plastic entropy \(\dot{\eta}_p = \dot{\eta} - \dot{\eta}_e\).

The plastic entropy is related to dissipative plastic structural changes and does not affect the internal energy (Simo and Miehe [21]).
3 NON-CLASSICAL DISSIPATIONLESS THERMOPLASTICITY

In this paper we analyse the GN thermoelastic framework of type II coupled with plastic behaviour with linear hardening.

The GN model introduces a scalar variable $\alpha$, called the thermal displacement, which is related to the temperature $\Theta$ by the relation:

$$\alpha(x, t) = \int_0^t \Theta(x, \tau) d\tau + \alpha_0$$

(9)

where $x$ is a point pertaining to the thermo-elastoplastic body defined on a regular bounded domain $\Omega$ of an Euclidean space, $\Theta = \Theta - \Theta_r$ represents the temperature variation from the uniform reference temperature $\Theta_r$ and $\alpha_0$ is the initial value of $\alpha$ at the time $t = 0$. As a consequence the time derivative of the thermal displacement field is the temperature variation, i.e. $\dot{\alpha} = \dot{\Theta}$. The thermal displacement gradient $\nabla \alpha$ is denoted by $\mathbf{g} = \nabla \alpha$.

The internal energy assumes the following form:

$$U = U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$$

(10)

Expanding the derivative of $U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$ and substituting in (3), we can write the overall dissipation:

$$D = (\sigma - \partial_{e_c} U) \cdot \dot{e}_c + \sigma \cdot \dot{\mathbf{g}} - \partial_{\alpha_1} U \cdot \dot{\alpha}_1 - \partial_{\alpha_2} U \cdot \dot{\alpha}_2 - \partial_{\eta_c} U \cdot \dot{\eta}_c - \partial_{\mathbf{g}} U \cdot \dot{\mathbf{g}} + \partial_\Theta \Theta \cdot \dot{\mathbf{g}} \geq 0$$

(11)

that leads to

$$\sigma = \partial_{e_c} U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$$

$$\chi_1 = \partial_{\alpha_1} U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$$

$$\chi_2 = \partial_{\alpha_2} U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$$

$$\Theta = \partial_{\eta_c} U(e_c, \alpha_1, \alpha_2, \eta_c, \mathbf{g})$$

Recalling that $\mathbf{g} = \nabla \dot{\alpha} = \nabla \dot{\Theta}$, introducing the entropy flux vector $\mathbf{p} = \frac{\mathbf{q}}{\Theta}$ and considering the relations (12), the total dissipation (11) reduces to this form

$$D = \sigma \cdot \dot{e}_p - \chi_1 \cdot \dot{\alpha}_1 - \chi_2 \cdot \dot{\alpha}_2 + \Theta \cdot \dot{\mathbf{h}}_p - (\partial_{\mathbf{g}} U \cdot \mathbf{p}) \cdot \nabla \Theta \geq 0$$

(13)

that leads to

$$\mathbf{p} = -\partial_{\mathbf{g}} U$$

(14)

Finally the total dissipation in the GN model turns to be:

$$D = \sigma \cdot \dot{e}_p - \chi_1 \cdot \dot{\alpha}_1 - \chi_2 \cdot \dot{\alpha}_2 + \Theta \cdot \dot{\mathbf{h}}_p \geq 0$$

(15)

being null the dissipation due to heat conduction. It is apparent that the above dissipation coincides to the classical mechanical dissipation $D_m = \sigma \cdot \dot{e}_p - \chi_1 \cdot \dot{\alpha}_1 - \chi_2 \cdot \dot{\alpha}_2$ [22] and the thermal dissipation $D_{th} = \Theta \cdot \dot{\mathbf{h}}_p$. 
4 ENERGETIC FUNCTIONS FOR COUPLED THERMOELASTICITY WITH PLASTICITY

In this paper the GN thermodynamic framework is formulated by introducing the following generalized vectors of the kinematic and dual static internal variables:

- total strain $\varepsilon = (\varepsilon, 0, 0)$,
- elastic strain $\varepsilon_e = (\varepsilon_e, \alpha_1, \alpha_2)$,
- plastic strain $\varepsilon_p = (\varepsilon_p, -\alpha_1, -\alpha_2)$,
- stress $\sigma = (\sigma, \chi_1, \chi_2)$.

The scalar product between dual quantities $\varepsilon$ and $\sigma$ has the mechanical meaning of the internal virtual work:

$$\varepsilon_n \bullet \varepsilon = \varepsilon_n \bullet \varepsilon + \chi_1 \cdot \alpha_1 + \chi_2 \cdot \alpha_2$$

(16)

Let us now derive the complete set of the energy functions for the considered GN model and the related constitutive relations in the framework provided by convex/saddle functions (Houlsby and Puzrin [23]) and, for different models, see also [24-28] and [29-31].

Classically, the analysis of a thermodynamic process starts from the Helmholtz free energy $\Phi$, which is a function of strain $\varepsilon$, temperature $\theta$ and thermal gradient vector $g$, that is $\Phi(\varepsilon, \theta, g)$ . The Helmholtz free energy $\Phi$ is assumed to be convex in the strain $\varepsilon$, and in the thermal gradient vector $g$ and concave in the temperature $\theta$ at any point $x$. The dual variables of the triplet $(\varepsilon, \theta, g)$ are the stress $\sigma$, the elastic entropy $\eta_e$ and the entropy flux vector $p$.

Now we derive the three conjugates of the free energy $\Phi$ with respect to one state variable.

A saddle function $\phi^*$ can be associated with the Helmholtz free energy $\Phi$, by considering the conjugate of $\Phi$ with respect to the strain tensor in the form:

$$\phi^*_x (\varepsilon, \theta, g) = \sup_{\varepsilon} \left\{ \varepsilon \bullet \varepsilon - \Phi(\varepsilon, \theta, g) \right\}$$

(17)

The function $\phi^*_x$ turns out to be convex in $(\varepsilon, \theta)$ and concave in $g$.

The conjugate of the function $\Phi$ with respect to $\theta$ is the concave function $\phi^*_{\theta}$ given by:

$$\phi^*_{\theta} (\varepsilon, \eta_e, g) = \inf_{\theta} \left\{ \eta_e \theta - \Phi(\varepsilon, \theta, g) \right\}$$

(18)

where $\eta_e$ is the entropy.

A saddle function $\phi^*_p$ can be associated with the Helmholtz free energy $\Phi$, by considering the conjugate of $\Phi$ with respect to the thermal gradient vector $g$, in the form:

$$\phi^*_p (\varepsilon, \theta, p) = \sup_{g} \left\{ p \cdot g - \Phi(\varepsilon, \theta, g) \right\}$$

(19)

and it is convex in $(\theta, p)$ and concave in $\varepsilon$. The variable $p$ denotes the entropy flux vector and is related to the heat flux $q$ by the relation $p = q/\theta$. 

257
Let us now derive the conjugate functions of the Helmholtz free energy with respect to the pairs of variables \((\varepsilon_e, \theta), (\theta, g)\) and \((\varepsilon_e, g)\).

The conjugate of \(\Phi\) with respect to \((\varepsilon_e, \theta)\) is the saddle function \(\varphi_{12}^*\), convex in \(\sigma\) and concave in \((\eta_e, g)\); given by:

\[
\varphi_{12}^*(\sigma, \eta_e, g) = \inf_{\varepsilon_e} \sup_{\theta} \left\{ \sigma \varepsilon_e + \eta_e \theta - \Phi(\varepsilon_e, \theta, g) \right\}
\]

(20)

The saddle function \(\varphi_{23}^*\) is the conjugate of \(\Phi\) with respect to the pair \((\theta, g)\):

\[
\varphi_{23}^*(\varepsilon_e, \eta_e, p) = \inf_{\theta} \sup_{g} \left\{ \eta_e \theta + p \star g - \Phi(\varepsilon_e, \theta, g) \right\}
\]

(21)

and results concave in \((\varepsilon_e, \eta_e)\) and convex in \(p\).

The conjugate of the Helmholtz free energy \(\Phi\) with respect to the pair \((\varepsilon_e, g)\) is the convex function \(\varphi_{13}^*\) in \((\sigma, \theta, \rho)\) given by:

\[
\varphi_{13}^*(\sigma, \theta, \rho) = \sup_{\varepsilon_e} \left\{ \sigma \varepsilon_e + \rho \star g - \Phi(\varepsilon_e, \theta, g) \right\}
\]

(22)

The saddle function \(\varphi^*\) is the conjugate of the Helmholtz free energy \(\Phi\) with respect to the triplet \((\varepsilon_e, \theta, g)\) and is defined as:

\[
\varphi^*(\sigma, \eta_e, p) = \inf_{\varepsilon_e} \sup_{\theta} \sup_{g} \left\{ \sigma \varepsilon_e + \eta_e \theta + p \star g - \Phi(\varepsilon_e, \theta, g) \right\}
\]

(23)

The conjugate function \(\varphi^*\) turns out to be convex in \((\sigma, \rho)\) and concave in \(\eta_e\).

It is worth noting that further relations connecting the above conjugate potentials can be derived following the above mentioned procedure but they are not explicitly provided for sake of conciseness.

5 CONSTITUTIVE RELATIONS

Let us now show that the constitutive relations for the GN thermoelasticity can involve, as a maximum, eight different thermodynamic potentials.

To this end the following equivalent relations in terms of the functions \(\Phi, \varphi_{13}^*, \varphi^*_{23}, \varphi_{12}^*\) and \(\varphi^*, \varphi_{23}^*, \varphi_{13}^*, \varphi_{12}^*\) are provided:

\[
\begin{align*}
(\sigma, -\eta_e, -p) &= d\Phi(\varepsilon_e, \theta, g) \\
(\varepsilon_e, \theta, g) &= d\varphi_{13}^*(\sigma, -\eta_e, -p) \\
(\varepsilon_e, \eta_e, p) &= d\varphi_{13}^*(\sigma, \theta, g) \\
(-\sigma, \theta, g) &= d\varphi_{23}^*(\varepsilon_e, -\eta_e, -p) \\
(-\sigma, \theta, p) &= d\varphi_{23}^*(\varepsilon_e, -\eta_e, g) \\
(\varepsilon_e, \eta_e, g) &= d\varphi_{12}^*(\sigma, \theta, -p) \\
(\varepsilon_e, \theta, p) &= d\varphi_{12}^*(\sigma, -\eta_e, g) \\
(-\sigma, \eta_e, g) &= d\varphi_{12}^*(\varepsilon_e, \theta, -p)
\end{align*}
\]

(24)
The relations (24) among conjugate functions are equivalent to the following Fenchel’s equalities, in terms of the functions \( \Phi, \phi^*, \phi_2, \phi_3 \) and \( \varphi^*, \varphi_{23}, \varphi_{13}, \varphi_3^* \):

\[
\begin{align*}
\Phi(\varepsilon, \theta, \mathbf{g}) + \varphi^*(\sigma, -\eta, -\mathbf{p}) &= \sigma \varepsilon \eta p - \mathbf{g} \\
-\varphi_{23}^*(\varepsilon, -\eta, -\mathbf{p}) + \varphi_2^*(\sigma, 0, \mathbf{g}) &= \sigma \varepsilon \eta p + \mathbf{g} \\
\varphi_{13}^*(\sigma, 0, -\mathbf{p}) - \varphi_2^*(\varepsilon, -\eta, \mathbf{g}) &= \sigma \varepsilon \eta p - \mathbf{g} \\
-\varphi_3^*(\varepsilon, \theta, -\mathbf{p}) - \varphi_{12}^*(\sigma, -\eta, \mathbf{g}) &= -\sigma \varepsilon \eta p - \mathbf{g}
\end{align*}
\]

(25)

On the basis of the above relations (24) eight different thermodynamic functions, with different combinations of the state variables, can be defined.

As a consequence, the differential relations (24) show that the constitutive relations for the considered GN model can be equivalently expressed in terms of such eight thermodynamic functions.

The energy definitions and the constitutive relations for the considered GN thermoelastic model without dissipation can be obtained by giving the mechanical meanings to the introduced energetic functions as reported in Table 1.

**Table 1**: Relations among the convex/concave functions and the thermodynamic energy functions.

<table>
<thead>
<tr>
<th>Helmholtz free energy:</th>
<th>( \Phi(\varepsilon, \theta, \mathbf{g}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal energy:</td>
<td>( U(\varepsilon, \eta, \mathbf{g}) = -\varphi_2^*(\varepsilon, -\eta, \mathbf{g}) )</td>
</tr>
<tr>
<td>Enthalpy:</td>
<td>( H(\sigma, \eta, \mathbf{g}) = -\varphi_{12}^*(\sigma, -\eta, \mathbf{g}) )</td>
</tr>
<tr>
<td>Gibbs free energy:</td>
<td>( G(\sigma, \theta, \mathbf{g}) = -\varphi_3^*(\sigma, \theta, \mathbf{g}) )</td>
</tr>
</tbody>
</table>

In addition, further four thermodynamic potentials can be defined with different combinations of the state variables from the above energy functions.

These thermodynamic potentials are named alternative since they differ from the GN thermodynamic potentials reported in Table 1 by the exchange of the displacement gradient \( \mathbf{g} \) with the entropy heat flux \( \mathbf{p} \).

The alternative thermodynamic potentials are reported in Table 2.

**Table 2**: Relations among the convex/concave functions and the alternative formulations of the thermodynamic energy functions.

<table>
<thead>
<tr>
<th>Alternative form of Helmholtz free energy:</th>
<th>( \hat{\Phi}(\varepsilon, \theta, \mathbf{p}) = -\varphi_3^*(\varepsilon, \theta, -\mathbf{p}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative form of the internal energy:</td>
<td>( \hat{U}(\varepsilon, \eta, \mathbf{p}) = -\varphi_{23}^*(\varepsilon, -\eta, -\mathbf{p}) )</td>
</tr>
<tr>
<td>Alternative form of the enthalpy:</td>
<td>( \hat{H}(\sigma, \eta, \mathbf{p}) = -\varphi_1^*(\sigma, -\eta, \mathbf{p}) )</td>
</tr>
<tr>
<td>Alternative form of the Gibbs free energy:</td>
<td>( \hat{G}(\sigma, \theta, \mathbf{p}) = -\varphi_1^*(\sigma, \theta, -\mathbf{p}) )</td>
</tr>
</tbody>
</table>

The constitutive relations for the considered GN model can be consistently deduced from the differential relations reported in (24) in terms of the thermodynamic energy functions and...
of their alternative formulations. In particular, the GN constitutive relations in terms of the four thermodynamic potentials are provided in Table 3.

Table 3: The thermodynamic potentials and the related constitutive relations.

<table>
<thead>
<tr>
<th>Helmholtz free energy $\Phi(\varepsilon_e, \theta, g)$</th>
<th>Internal energy $U(\varepsilon_e, \eta_e, g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = \partial_{\varepsilon_e} \Phi(\varepsilon_e, \theta, g)$</td>
<td>$\sigma = \partial_{\varepsilon_e} U(\varepsilon_e, \eta_e, g)$</td>
</tr>
<tr>
<td>$\chi_i = \partial_{\varepsilon_e} \Phi(\varepsilon_e, \theta, g)$</td>
<td>$\chi_i = \partial_{\varepsilon_e} U(\varepsilon_e, \eta_e, g)$</td>
</tr>
<tr>
<td>$\chi_2 = \partial_{\varepsilon_e} \Phi(\varepsilon_e, \theta, g)$</td>
<td>$\chi_2 = \partial_{\varepsilon_e} U(\varepsilon_e, \eta_e, g)$</td>
</tr>
<tr>
<td>$-\eta_e = \partial_{\eta_e} \Phi(\varepsilon_e, \theta, g)$</td>
<td>$\theta = \partial_{\eta_e} U(\varepsilon_e, \eta_e, g)$</td>
</tr>
<tr>
<td>$-p = \partial_{p} \Phi(\varepsilon_e, \theta, g)$</td>
<td>$-p = \partial_{p} U(\varepsilon_e, \eta_e, g)$</td>
</tr>
</tbody>
</table>

Moreover the GN constitutive relations in terms of the four alternative forms of the thermodynamic potentials are given in Table 4.

Table 4: Alternative formulations of the thermodynamic potentials and the related constitutive relations.

<table>
<thead>
<tr>
<th>Alternative form of Helmholtz free energy $\hat{\Phi}(\varepsilon_e, \theta, p)$</th>
<th>Alternative form of the internal energy $\hat{U}(\varepsilon_e, \eta_e, p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = \partial_{\varepsilon_e} \hat{\Phi}(\varepsilon_e, \theta, p)$</td>
<td>$\sigma = \partial_{\varepsilon_e} \hat{U}(\varepsilon_e, \eta_e, p)$</td>
</tr>
<tr>
<td>$\chi_i = \partial_{\varepsilon_e} \hat{\Phi}(\varepsilon_e, \theta, p)$</td>
<td>$\chi_i = \partial_{\varepsilon_e} \hat{U}(\varepsilon_e, \eta_e, p)$</td>
</tr>
<tr>
<td>$\chi_2 = \partial_{\varepsilon_e} \hat{\Phi}(\varepsilon_e, \theta, p)$</td>
<td>$\chi_2 = \partial_{\varepsilon_e} \hat{U}(\varepsilon_e, \eta_e, p)$</td>
</tr>
<tr>
<td>$-\eta_e = \partial_{\eta_e} \hat{\Phi}(\varepsilon_e, \theta, p)$</td>
<td>$\theta = \partial_{\eta_e} \hat{U}(\varepsilon_e, \eta_e, p)$</td>
</tr>
<tr>
<td>$g = \partial_{p} \hat{\Phi}(\varepsilon_e, \theta, p)$</td>
<td>$g = \partial_{p} \hat{U}(\varepsilon_e, \eta_e, p)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternative form of the enthalpy $\hat{H}(\sigma, \eta_e, p)$</th>
<th>Alternative form of Gibbs free energy $\hat{G}(\sigma, \theta, p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-\varepsilon_e = \partial_{\varepsilon_e} \hat{H}(\sigma, \eta_e, p)$</td>
<td>$-\varepsilon_e = \partial_{\varepsilon_e} \hat{G}(\sigma, \theta, p)$</td>
</tr>
<tr>
<td>$-\alpha_i = \partial_{\alpha_i} \hat{H}(\sigma, \eta_e, p)$</td>
<td>$-\alpha_i = \partial_{\alpha_i} \hat{G}(\sigma, \theta, p)$</td>
</tr>
<tr>
<td>$-\alpha_2 = \partial_{\varepsilon_e} \hat{H}(\sigma, \eta_e, p)$</td>
<td>$-\alpha_2 = \partial_{\alpha_2} \hat{G}(\sigma, \theta, p)$</td>
</tr>
<tr>
<td>$\theta = \partial_{\eta_e} \hat{H}(\sigma, \eta_e, p)$</td>
<td>$-\eta_e = \partial_{\eta_e} \hat{G}(\sigma, \theta, p)$</td>
</tr>
<tr>
<td>$g = \partial_{p} \hat{H}(\sigma, \eta_e, p)$</td>
<td>$g = \partial_{p} \hat{G}(\sigma, \theta, p)$</td>
</tr>
</tbody>
</table>

It is worth noting that the constitutive relations reported in Table 3 and 4 in terms of the
complete set of the thermodynamic functions turn out to be all equivalent each other.

The correspondences between the thermodynamic potentials provide a generalization to the present GN thermoelastic framework of the classical Legendre transform concerning the sum of the elastic and complementarity energies. In fact, assuming that the state variables \((\varepsilon, \theta, g)\) and \((\sigma, \eta, p)\) fulfil the constitutive relations reported in Tables 3 and 4, the Fenchel's equalities (24) can be rewritten in the following form:

\[
\begin{align*}
\Phi(\varepsilon, \theta, g) - \hat{H}(\sigma, \eta, p) &= \sigma * \varepsilon - \eta * \theta - p * g \\
-G(\sigma, \theta, g) + \hat{U}(\varepsilon, \eta, p) &= \sigma * \varepsilon + \eta * \theta + p * g \\
U(\varepsilon, \eta, g) - \hat{G}(\sigma, \theta, p) &= \sigma * \varepsilon + \eta * \theta - p * g \\
H(\sigma, \eta, g) + \hat{\Phi}(\varepsilon, \theta, p) &= -\sigma * \varepsilon + \eta * \theta - p * g
\end{align*}
\]  

(26)

In particular we underline the mechanical meaning of the second relation of (26):

– the difference between the alternative form of the internal energy and the Gibbs free energy provides the sum of mechanical virtual work and thermal virtual work between elastic entropy and temperature and between the thermal displacement gradient and the entropy flux vector.

5 CONCLUSION

Using a systematic procedure based on convex/concave functions and Legendre transforms, the consistent set of the thermodynamic functions in the framework of GN dissipationless thermo-elasto-plasticity is addressed. Starting from Helmholtz free energy, the thermodynamic potentials, i.e. the internal energy, the enthalpy and the Gibbs free energy, are derived. Moreover it is shown that four more thermodynamic potentials, named alternative, can be provided. The relations between these functions are provided and the constitutive relations are obtained.

REFERENCES


[31] Guarracino F., Considerations on the Numerical Analysis of Initial Post-Buckling Behaviour in Plates and Beams, Thin-Walled Structures (2007) 45: 845-848,
CONSIDERATIONS ON THE PLASTIC BUCKLING OF BI-LAYERED CYLINDRICAL SHELLS

FEDERICO GUARRACINO * AND ALASTAIR C. WALKER†

* Department of Structural Engineering (DIST)
Università degli Studi di Napoli “Federico II”
via Claudio, 21 – 80125 Napoli, Italy
e-mail: fguarrac@unina.it

† INTECSEA
WorleyParsons Group
Perth WA 6000, Australia

Key words: Bi-layered Cylindrical Shells, Local Buckling, Heat Treatment, Experimental Testing, Computing Methods.

Abstract. The mechanical behaviour of lined cylindrical shells under bending is analysed from an experimental and a theoretical standpoint. Attention is focused on the parameters that determine the resistance to shell ovalisation and liner wrinkling. This is attained by means of a simplified analytical approach which could provide guidance to safe and economic design of lined shells and the paper describes the results obtained so far during the research.

1 INTRODUCTION

Oil and gas industry is increasingly facing the need to transport large volumes of untreated and corrosive products over long distances from the well to the processing plant. In order to reduce costs, recourse is more and more being made to lined pipes, which consist of a carbon steel load bearing outer pipe, which provides the structural capacity, and a corrosion resistant alloy (CRA) liner, protecting the carbon steel outer pipe from the transported corrosive product. The liner is mechanically fitted inside the outer pipe. Currently there is inadequate information available with which to carry out analysis of the response of a lined pipe to the application of bending, axial loading and internal net pressure and the development in the manufacturing and use of CRA lined pipes has been essentially based on heuristic and empirical approaches. On the other hand very little guidance is available from design codes.

In the present work the variables that have an influence on the mechanical behaviour of lined circular shells under bending and determine the resistance to ovalisation and liner wrinkling are identified and discussed by means of a simplified analytical approach which could provide guidance to safe and economic installation of these elements.

2 THE MECHANICS OF CRA LINED PIPES

There are two forms of pipe that are internally protected by a layer of corrosion resistant
alloy (CRA): the first one is commonly called ‘clad pipe’ and the second one is known as ‘lined pipe’. The clad pipe has a layer of CRA rolled onto the carbon steel plate prior to forming the pipe. The CRA material is metallurgically intermixed with the carbon steel at the interface of the two materials and thus forms a very strong bond. On the contrary, the lined pipe has the layer of CRA material formed from a thin sheet that is bonded to the inside of the carbon steel pipe by means of mechanical processes, as shown in Figure 1.

![Figure 1: Sketch of a bi-layered pipe](image)

The requirement for CRA pipes to be used for high temperature flow lines is that the pipe should be able to sustain bending deformations exceeding the material yield strain, without degradation of the CRA material and liner. Experience has shown that CRA clad pipe can sustain very large levels of bending strain whereas, on the basis of previous tests, the CRA liner in lined pipe has ‘wrinkled’, or developed local buckles at comparatively low levels of strain. This factor has limited the use of CRA lined pipe to pipelines that have low levels of temperature or which have been prevented from bending on the seabed.

However, from a theoretical standpoint there is still a certain degree of uncertainty about the variables that influence the mechanical behaviour of lined pipe during bending and the parameters that determine the resistance to pipe ovalisation and liner wrinkling. Therefore, a better understanding of the mechanics would allow a safer and more economic design of lined pipe and help to extend the field of use of the more cost-effective lined pipes in place of the clad ones.

2.1 The manufacturing method

The methods of manufacture of CRA lined pipe from various suppliers are generally similar, even if can vary in detail. The carbon steel pipe is received in a manufactured state and the inner surface is grit blasted to obtain a smooth clean surface. The CRA material is received in the form of a cold-reduced sheet wound into a coil. The sheet is cut to width and length and then formed into a cylinder by longitudinally welding the seam. The cylinder is subjected to heating to anneal the CRA and then fitted into the bore of the pipe. This operation is facilitated by manufacturing the CRA cylinder to a diameter less than the inside diameter of the pipe. The radial gap between the CRA cylinder and the bore of the pipe is controlled by the width of the CRA sheet cut from the coil and is decided purely on the basis of ease of fitting the CRA cylinder into the pipe joint. The inclusions of the radial gap is
termed here as ‘looseness of fit’.

The liner is expanded to the inner diameter of the pipe using a tool that is powered hydraulically and then the combined pipe and liner are expanded to a specified pressure applied to the tool. Following the expansion, the hydraulic pressure is reduced to zero and the tool is moved along the inside of the liner to expand the next section of the pipe and liner. The tool is about 3 m long so that four separate expansion operations are required for the usual 12.2 m pipe joint. The degree of expansion is generally such that the maximum hoop strain imposed on the carbon steel pipe is just less than the material yield strain.

Assuming that the CRA liner and the bore of the pipe are exactly the same, Figure 2 shows the increase of strain applied by the expansion tool. The strain-stress line for the carbon steel and the CRA follow their respective material properties up to the level of the strain specified by the hydraulic pressure applied to the expanding tool, shown by the vertical green line. When the pressure is reduced the carbon steel follows the same stress-strain line, back to A, as was followed during the initial expansion. The CRA material unloads along the elastic line but with an offset value of strain, to B. There is a residual hoop strain in the CRA which represents an ‘interference fit’ which causes a pressure between the outer surface of the liner and the inside of the pipe, i.e. the interface pressure. This pressure makes the pipe diameter to dilate a little and the liner diameter to contract a little until a common strain, somewhere between A and B is attained together with a common interface pressure. This interface pressure is the basis of the bonding of the liner with the pipe.

It may be noted from Figure 2 that the level of the interface pressure is purely a matter of the expansion strain and the respective material properties of the carbon steel and the CRA. However, as noted above, in order to assemble the CRA tube into the pipe bore, it is necessary to have looseness of fit. The radial gap is generally in the range of 3 mm to 4 mm.

![Figure 2: Diagrammatic representation of the manufacturing expansion process](image)

Figure 3 demonstrates how the initial radial gap affects the manufacturing method. The liner effectively has an initial negative strain, compared to the zero strain of the bore of the pipe, at point A. The value of the negative strain is simply the radial gap divided by the radius of the CRA cylinder. The expanding tool first increases the diameter of the CRA cylinder.
until it contacts the bore of the pipe, see the dashed red line in Figure 3.

![Figure 3: Diagrammatic representation of the expansion operation for an assumed level of looseness of fit](image)

Then the combined pipe and liner is expanded by the tool up to the specified levels of strain, shown by the vertical green line. Removing the hydraulic pressure causes the liner and the pipe to reduce their respective hoop strains and an interference bond results, as in the case for no looseness of fit, although the presence of the initial radial gap can result in a lower interface pressure.

### 2.2 Local buckling of liner under bending

It has been observed during tests that if a lined pipe is subjected to bending, or combinations of axial forces and bending, the liner will develop wrinkles. An example of the form of the wrinkling, sometimes call ‘local buckling’, is shown in Figure 4.

![Figure 4: Typical form of liner wrinkling following completion of test [1]](image)
These localised deformations are typical of the patterns observed in lined pipes that have diameters ranging from 10-inch to 26-inch. The compressive strains at which such wrinkles are first observed in tests have generally been derived by means of finite non-linear modelling. In this manner the initiation of local deformations in the liner due to bending deformations, applied to the lined pipe at ambient temperature and pressure, and the development of these deformations into wrinkles are calculated, see Figure 5.

It can be pointed out that a comprehensive numerical modelling of such a phenomenon is rather complicated and requires, in principle, the involvement of plasticity, damage and coupling of different numerical techniques. However, in the past several studies have been conducted at the University of Naples on each of these topics (see refs. [2-7], [8-10] and [11-15], respectively) and a good understanding of the suitable theoretical and computational tools for the problem at hand is therefore available.

The primary loading applied to the model is a constant bending moment to replicate the conditions in the central region of bending tests. As a pipe subject to bending ovalises, the boundary conditions and load application methods must allow ovalisation to occur [16].

The wrinkling mechanism observed is as follows.

**Figure 5**: Example result from FE modelling, showing the liner wrinkling for bending strain =2.5%

During bending, ovalisation of the pipe and liner section occurs. The degree of ovalisation is related to the wall thickness of the section. This is primarily a geometric effect (independent of material properties) and the thin-walled liner, if it was unrestrained, would ovalise significantly. This results in a decrease in the contact pressure between the liner and pipe at the top and bottom sections of a pipe, furthest away from the neutral axis of bending, with an increase in contact pressure at the side. A necessary condition for wrinkling commencement is that the contact pressure reduces to zero along the compression face, although wrinkling does not necessarily occur immediately after zero pressure is attained.

It appears that the longitudinal stress levels also need to reach a critical value determined by the tangent elastic modulus from the stress strain curve. This reduces as the stress, and hence plastic strain, increases and it implies that at low strains wrinkling will not occur, even for an unsupported condition.

Very small initial imperfections are likely to have exceedingly large effects on reducing
the bending strains at which wrinkling is initiated, given the well-known unstable nature of thin-walled cylinders when subjected to axial compression.

3 A SIMPLIFIED BUCKLING MODEL FOR COUPLED LAYERS

Results from basic studies of shell buckling have established that if a thin shell, such as the CRA liner, is contained within a relatively rigid boundary, the thin shell cannot buckle and develop deformations around its surface. That previous work has been presented as the theory of one-way buckling. It is concluded from this theory that the initial deformations, and later wrinkles, can occur in the liner only when the initial interface pressure has been reduced to zero and probably even when the relative radial movements of the liner and the pipe result in a gap between them, albeit the magnitude of the gap might be very small.

Following achievement of a level of axial bending strain at which the liner detaches from the pipe wall and causes a small gap between the liner and the pipe wall surfaces, the liner is free to buckle locally. In such conditions the critical stress is given by the following equation [18]

$$\sigma_{cr} = \frac{2Et_l}{D_l\sqrt{3(1-v^2)}}$$

where $E$ is the Young’s modulus, $v$ is the Poisson’s ratio, $t_l$ is the liner thickness and $D_l$ is the liner diameter.

However, in the majority of practical cases the value given by equation (1) is greater than the CRA yield stress and so the local buckling needs to be calculated using the tangent modulus of the CRA instead of the elastic modulus [17]. In order to do so, an analytical description of the stress-strain relationship is required, such as the Ramberg-Osgood one

$$\varepsilon = \frac{\sigma}{E} + \left(\varepsilon_y - \frac{\sigma}{E}\right) \left(\frac{\sigma}{\sigma_y}\right)^\beta$$

$\sigma_y$ and $\varepsilon_y$ being the yield stress and the corresponding elongation, respectively. For most engineering cases it is $\beta \geq 5$.

Equation (2) gives the following expression of the tangent modulus

$$E_t = \left(\frac{\partial \varepsilon}{\partial \sigma}\right)^{-1} = E\left(\frac{\sigma_y}{\sigma^\beta + \beta(E\varepsilon_y - \sigma_y)\sigma^{\beta-1}}\right)$$

and Figure 6 shows the calculated tangent modulus for a typical CRA obeying to a Ramberg Osgood material description.

By virtue of equation (3), equation (1) becomes

$$\sigma_{cr} = \frac{2Et_l}{D_l\sqrt{3(1-v^2)}}$$

269
Figure 6: CRA material tangent modulus for a range of axial applied stresses

In order to analyse the detachment of the liner from the outer pipe it can be assumed that the pipe wall is a rigid surface and that all the radial deformations due to the applied axial bending strains occur only in the liner. In such a manner the ovality of the liner tends to be inhibited by the interaction of the liner with the inside of the pipe wall. However, by considering only the section of the liner at the bottom of the pipe, that is where the liner potentially could lift-off the pipe inner surface, a simplified understanding of the mechanism causing initiation of local liner deformations can be obtained. The ovalising pressure \( p \) acting on the unit of area of the liner in the direction normal to the neutral axis, see Figure 7, can be expressed as [19]

\[
p = \chi^2EI_y
\]

where \( y \) is the distance from the neutral axis and the curvature, \( \chi \), is given by \( \chi = M / EI \). \( M \) is the bending moment and \( I \) is the second moment of area of the composite section about its neutral axis.

Therefore, assuming a unit width of the liner at the bottom of the pipe, the equivalent pressure around the CRA tube that causes the liner to ovalise, for purely elastic material properties, can be described by

\[
p = \frac{2t\sigma_b\varepsilon_b}{D_i}
\]

where \( \sigma_b \) is the maximum bending stress and \( \varepsilon_b \) is the bending strain.

The pipe/liner assembly must be loaded to levels of axial bending strain greater than the CRA and steel yield strains before the maximum bending conditions at the liner would enable separation of the liner and pipe surfaces. Thus, for the post-yield condition in the liner it is

\[
p = \frac{2t\sigma_b \left[ \sigma_y + E_i (\varepsilon_b - \varepsilon_y) \right]}{D_i}
\]
Finally, assuming that a separation is necessary between the liner and the pipe wall to enable the ignition of liner wrinkling, the equivalent radial pressure caused by the bending strain can be equated to the sum of the initial manufacturing interface pressure and the internal applied pressure and equations (4) and (7) suffice for an approximate evaluation of the problem.

Figures 8 and 9 show that the presented simplified approach has the same trend as the numerical model, but underestimates the rate of growth of the amplitude of the gap between the line and the pipe inner surface. Nevertheless, the simplified method provides confirmation of the mechanics underlying the initiation of lift-off and the growth of the gap between the liner and the pipe. Figure 10 shows plots of liner wrinkling strain versus outside diameter for experimental and calculated values.
4 CONCLUSIONS

From the proposed simplified analysis for a CRA lined pipe subjected to bending loads a few conclusions can be drawn.
First of all, at maximum compressive bending strain levels that are less than the CRA yield strain, the liner remains attached to the inside surface of the pipe wall, held there by the initial interface pressure. Additionally, at maximum bending strains greater than the yield strain of the liner material, the interface pressure is progressively reduced by increasing the maximum compressive bending strains; the curvature of the lined pipe under bending applies a ‘lift-off pressure’ which counters the interface pressure. Finally, further increase in the bending strains results in separation of the liner from the pipe wall at the bottom of the pipe.

Overall, it has been shown that a reasonable evaluation of the phenomenon can be attained by means of the illustrated formulae. Further theoretical investigation is currently underway at the University of Naples in the framework of a long term pipeline assessment programme [20-26].

REFERENCES


Abstract. Non-proportionality of straining, initial texture and hardening have been investigated in a ferritic steel polycrystal within a crystal plasticity finite element framework. Two extreme forms of hardening are investigated; namely, isotropic latent-hardening and anisotropic self-hardening. Dislocation density evolutions on all independent slip systems have been calculated in order to investigate the establishment of dislocation distributions and evaluate their dependence on non-proportionality, hardening, texture and predicted ductility. The results show that non-proportionality effects are more pronounced under isotropic latent-hardening as opposed to anisotropic self-hardening especially under non-proportional uniaxial strains.

Key words: Non-proportionality, texture, hardening, dislocation distributions, forming limits, necking, strain localization

1. Introduction

This paper provides an overview on the effects of non-proportionality of strain, its link with texture, dislocation distributions and their consequent effects on predicted ductility in ferritic steel polycrystals within a crystal plasticity finite element framework. Two initial textures are incorporated within a polycrystal and investigated under two hardening forms (anisotropic self- and isotropic latent-hardening). The formation of dislocation distributions under non-proportional strain paths is evaluated and in addition, ductility is predicted using a proposed failure criterion for various non-proportional strain paths. The motivation for this study lies in further understanding the mechanisms that drive non-proportionality effects in ferritic steel.

Sheet processing is an important stage in the manufacturing of automobiles. Ideally, strains are applied to sheets within an established strain safe region commonly described by a forming limit diagram (FLD) [1]. The limit diagram represents a safety curve above which the material is considered necked a below which, safe straining is assumed. Empirical evidence suggests that subjecting metals to a non-proportional strain path enables higher levels of strain to be achieved. Conversely, premature failure may also occur when a counter-beneficial strain path is employed. Limit strains are obtained experimentally by using a hemispherical punch-stretching test on carefully marked flat metal sheets until the onset of necking is observed. However, due to associated costs, FLDs are predicted computationally by evaluating the onset of strain instabilities within the material on the basis of localization criteria.
Numerous techniques have been reported and implemented to predict limit strains within phenomenological and crystal plasticity frameworks [2-4], [5], [6] [7-10]. However, these studies have centered on predicting FLDs under proportional strains and have not investigated the effects of strain non-proportionality on ductility in the context of hardening, and texture. Similarly, studies on dislocation structures in polycrystals undertaken within phenomenological and crystal plasticity frameworks also lack information in predicting response to non-proportional strains [11-14].

In this paper, two hardening models are investigated; self- and isotropic latent-hardening models adopted from [15]. A failure criterion implemented within a crystal plasticity framework is coupled with these hardening models and a representative oligocrystal is used to investigate the response of ferritic steel to non-proportional strain paths. The development of dislocation distributions is shown and the effect of texture on predicted ductility in ferritic steel is presented.

2. Crystal plasticity framework, hardening model and dislocation distributions

The crystal plasticity framework used here is based on the kinematic decomposition of the deformation gradient into elastic ($F^e$) and plastic ($F^p$) tensors laid out by [16] and implemented into a VUMAT (see [17] for an overview).

2.1. Hardening model and dislocation distributions

On the basis that non-proportionality effects in polycrystals result from the formation of dislocation distributions that are strain path dependent, dislocation evolution on slip systems in the polycrystal is tracked within a crystal plasticity framework. Dislocations are calculated based on the hardening developed on each slip system. The hardening law adopted by [15] and others [7] and [8] has been employed in this study. Consider the slip system strength $g$, also indicative of the resistance to dislocation motion calculated such that the hardening modulus $h_0$, and fitting parameter, $m$ are determined for a particular BCC steel being investigated. Hence, the strength on each slip system is calculated using

$$g^a = h_0 \left(1+ \frac{h_0 \gamma_{sum}}{\tau_0 m} \right)^{m-1} \gamma^a \tag{1}$$

where the accumulated slip is given by

$$\gamma_{sum} = \sum_{\beta=1}^{N_{slip}} \left( \int_0^t \gamma \, dt \right) \tag{2}$$

and dislocation densities on active slip systems are updated according to

$$\rho^a = \int_0^t \rho^a \, dt. \tag{3}$$

The physical relationship between slip system strength and dislocation density is given by

$$g^a = g_0 + \beta \sqrt[\nu]{\rho^a}. \tag{4}$$

where the density of dislocations on each slip system is calculated based on an initial value of strength for a particular slip system and updated based on eqs. (1) – (4). The two forms of
hardening considered here are isotropic latent-hardening and anisotropic self-hardening. Under isotropic latent-hardening, active slip systems cause hardening on the inactive systems.

\[ g = h_0 (1 + \frac{h_0}{\tau_0 m} \gamma_{\text{sum}})^{m-1} \gamma_{\max}. \]  

(5)

A second possible form is termed self-hardening. In this case, the slip resistance only develops on active slip systems and dislocation densities are calculated based on these active systems.

\[ g^a = h_0 (1 + \frac{h_0}{\tau_0 m} \gamma_{\text{sum}})^{m-1} \gamma^a. \]  

(6)

The behavioral trend observed experimentally shows that mixed forms of hardening typically develop in polycrystals. However, implementing these two forms of hardening (anisotropic self- and isotropic latent) provide an insight into dislocation responses as well as failure prediction in ferritic steel polycrystals. Another form of hardening not investigated here is a form of latent hardening in which hardening on active slip systems result in more hardening on the inactive systems. Investigations have shown that the predicted response is similar to isotropic latent-hardening adopted here [18]. Also, the authors are aware of more advanced dislocation models to predict dislocation evolution on slip systems [19]. However, this simple phenomenological model has been adopted in this paper for investigatory purposes prior to implementing more advanced dislocation models.

**Dislocation distributions**

The investigation of dislocation distributions formed upon straining a polycrystal provides a basis for justifying the effects of non-proportionally straining BCC polycrystals. Apart from texture which also accounts for non-proportionality effects (usually in the form of higher or lower ductility), the possibility of evaluating dislocation distributions provides an insight into potential strain paths that will exhibit clear differences in terms of ductility. Since all the straining is applied in the 1-2 plane, the distributions are generated by rotating all slip directions in each grain of the polycrystal into the 1-2 plane and averaging dislocation densities, binned based on the angle $\theta$ between the rotated slip vector and the [100] direction such that

\[ \rho_0^{\text{avg}} = \frac{\rho_0^{\text{tot}}}{N_0} = \frac{1}{N_0} \sum_{\beta=1}^{N_{\text{grains}}} \sum_{a=1}^{N_{\text{slip}}} \rho_0^a. \]  

(7)

$N_{\text{grains}}$ and $N_{\text{slip}}$ represent the number of grains in the oligocrystal and the number of slip directions in each grain respectively and $\rho_0^{\text{avg}}$ denotes the average density of dislocations of orientation $\theta$ based on the summation of binned slip direction densities $\rho_0^{\text{tot}}$ and the total number of slip directions within each grain that fall within the cutoff criterion of 1 degree, $N_0$.

**2.2. Material properties**

The material properties used in the slip rule and hardening model have been determined for a ferritic steel using a representative volume element with 216 randomly
oriented grains under uniaxial loading conditions, illustrated in Fig. 1a. The yield strength and isotropic hardening properties were chosen in order for the macroscopic stress-strain response from the polycrystal shown in Fig. 1b to match experimental observations for the ferritic steel considered. Furthermore, strain rate effects have been neglected since they are likely to be negligible at room temperature for ferritic steel and the rate sensitivity parameter, \( n \), in the slip rule has been chosen appropriately for this purpose. The macro-mechanical properties for the ferritic steel being considered and the material properties used in the simulations are given in Table 1.

2.3. Body centered cubic system (BCC) slip

The twelve slip systems, \{110\}<111>, present in BCC crystals have been incorporated into an ABAQUS user material subroutine and used in all simulations. It has been shown that texture and stress response in BCC metals can be adequately predicted using the \{110\}<111> set of slip systems [20, 21]. An earlier study by [22] on dislocation evolution in Niobium crystals investigated the relative roles of the BCC slip system families and concluded that using the first 24 slip system families provided the closest agreement to experimental data. It can however be drawn from that study that the error associated with the prediction using the first 12 slip systems is moderate. Hence, due to the numerous number of simulations to be undertaken in this paper and for computational efficiency, the full 48 slip systems, which consist of \{112\}<111> and \{123\}<111> slip systems in addition to the \{110\}<111> slip systems have not been accounted for in this present work.

3. 3-D model development, localization criterion and measure of non-proportionality

This section outlines a description of the models adopted in this study. Further, the initial textures used in this paper are shown, boundary conditions are outlined, the localization criterion adopted is validated and non-proportional strain paths are illustrated.

3.1. 3-D model development

The model adopted in this paper has 216 grains, each grain consisting of \(3\times3\times3\) C3D8R elements shown in Fig. 2. In all simulations, displacement is applied to the in-plane surfaces (positive 1- and 2- surfaces), and the back surfaces are constrained to remain planar. Also, the positive 3- surface is constrained to remain planar.

3.2. Initial texture

Two textures are adopted in this paper. An initially randomly oriented texture (R-1) has been investigated as shown by the pole figures in Fig. 3a and a measured ferritic steel (FS-1) texture is also evaluated (Fig. 3b). Texture R-1, computationally generated, gives a general insight into material behavior while FS-1, shows the effects of directionality typically present in rolled sheet metals.

3.3. Localization criterion

The onset of necking is determined by evaluating local critical grains developing plastic strains in the polycrystal model based on the increments of plastic strain between \( t_n \) and \( t_{n+1} \), normalized with respect to the macro plastic strain increment within the
corresponding deformation states. This criterion is similar to the combined approach adopted by Volk and Hora [9] in which the onset of instabilities within a polycrystal is evaluated by tracking plastic strain increments during a deformation history on the basis of experimental knowledge of fracture point. The failure criterion utilized here is fully outlined and calibrated in [23].

**Validation of failure criterion under proportional strain paths**

The forming limit curve under proportional strain paths has been predicted by subjecting the polycrystal in Fig. 2a to a range of proportional strain paths. These proportional strain paths (A-E) shown in Fig. 4a were applied to the positive 1- and 2-surfaces of the polycrystal and the onset of localization along each path based on the normalized, incremental average plastic strain is indicated in Fig. 4b. The predicted limit strain curve for the Numisheet 2008 BM 1 ferritic steel [24] is compared to its experimentally obtained response [9]. As seen in Fig. 4b, a close agreement is observed and on this basis, the failure criterion will be used to predict limit strain under non-proportional strain paths.

3.4. **Non-proportional strain paths adopted**

The primary purpose of this paper is to investigate the response of ferritic steel polycrystals to non-proportional strain paths such as that shown in Fig. 5. The notation is such that proportional biaxial and uniaxial strain paths are represented by B-1 and U-1 respectively. Both proportional strain paths are encompassed by a range of non-proportional paths consisting of two stages. Consider path B-2 which represents a non-proportional biaxial strain path and consists of B-2a and B-2b that terminate at a biaxial state (B). Similarly, U-5 represents a non-proportional uniaxial strain consisting of plane strain tension followed by a plane strain compression to a final uniaxial state (U).

The positive 1- and 2-surfaces of the polycrystal in Fig. 2 have been subjected to the individual strain paths in Fig. 5 and the results are discussed next.

4. **Effects of texture, hardening and non-proportionality on limit strain in BCC polycrystals**

The coupled effects of non-proportionality, texture and hardening on predicted ductility is discussed in this section. Firstly, formed dislocation distributions based on the slip system binning methodology outlined earlier is shown schematically for non-proportional biaxial strain paths on the basis of anisotropic self-hardening. Subsequently, the distributions are shown graphically in more details for both cases of hardening- anisotropic self- and isotropic latent-hardening. And finally, the effects of non-proportionality on predicted ductility are presented on the basis of a failure criterion fully outlined in [23].

The term dislocation distributions and its link with non-proportionality is a concept used continuously throughout this paper. Since dislocation evolution is accounted for on each slip system within the crystal plasticity framework, it is possible to classify similarly oriented slip systems within a polycrystal and to schematically show the density of dislocations. On the basis of the dislocation binning methodology outlined in the previous section, similarly oriented slip systems are shown in Fig. 6 for a through section of a polycrystal subjected to proportional and non-proportional biaxial strain B-1 and B-4 respectively (illustrated in Fig. 5). Note that B-1 represents proportional biaxial strains while B-4 consists of consecutive
plane strain phases in the 1- and 2- directions to achieve a final biaxial state. A cutoff criterion of 10 degrees has been adopted in Fig. 6 and the strain-path based distribution of dislocation within the polycrystal is shown for angular ranges between 0° and 180°. It is clear that different dislocation distributions result within the polycrystal at the final biaxial state. Although, evaluating the effects of these distributions on non-proportionality behavior remains unclear, it is worth investigating since it can provide insight into particular observed responses. To provide a clearer overview on strain-path based response to non-proportional strains, the average binned dislocation density on similarly oriented slip systems throughout the polycrystal will be subsequently shown graphically.

Fig. 7 illustrates the dislocation distribution for non-proportional biaxial strain under anisotropic self- and isotropic latent-hardening. Consider Fig. 7a which shows dislocation distribution responses for texture R-1 at the final biaxial state for non-proportional strain paths B-1, B-3 and B-4, it is seen that the dislocation distribution response to the various strain paths are similar. However, in comparison with Fig. 7b, clearer differences in response are seen even at an identical biaxial state and at considerably low strains (25%). Consider the initially textured polycrystal denoted by FS-1, a similar response is observed under both forms of hardening, but a significantly higher dispersion is observed at higher angles under isotropic latent-hardening.

Fig. 8 shows the dislocation distribution responses for textures R-1 and FS-1 at a final uniaxial state for strain proportional strain path U-1 and a range of non-proportional paths U-3, U-4 and U-5 (illustrated in Fig. 5). Consider Fig. 8a and 8b in which the response of the polycrystal with initially random texture R-1 is shown for the case of anisotropic self- and isotropic latent-hardening respectively. The differences in dislocation distributions are clearly seen for the range of non-proportional paths under isotropic latent-hardening as opposed to anisotropic self-hardening. This can also be observed in Fig. 8c and 8d for ferritic steel texture FS-1 under both hardening forms respectively.

By comparing the dislocation distribution responses to biaxial and uniaxial strain paths, it is clear that the effects of non-proportionality are more visible under uniaxial strain paths and more so, under isotropic latent-hardening. The symmetry associated with biaxial strain arguably results in less pronounced differences in dislocation distributions at the final biaxial state. But, it is imperative to investigate ductility and evaluate the effects of non-proportionality. An important question that arises is whether larger differences in ductility are achievable under non-proportional uniaxial strains as against biaxial strains on the basis of the formed dislocation distributions.

The ductility (based on the average plastic strain in the polycrystal) under non-proportional strain paths are shown below for textures R-1 and FS-1 under both forms of hardening. Consider Fig. 9a and 9b which shows the predicted ductility for both initial textures under isotropic latent-hardening and anisotropic self-hardening respectively. Here, minor deviations from proportionality indicated by non-proportional paths B-3 and U-3 previously illustrated in Fig. 5 are compared with their corresponding proportional strains B-1 and U-1.

In both hardening cases, clear differences are observed under non-proportional uniaxial strain paths as opposed to biaxial strains for minor deviations from proportionality. However, consider Fig. 9a in which the predicted ductility is shown for both textures for strain paths B-1 and B-3. It can be clearly seen that the ductility achieved by ferritic steel texture FS-1 is higher than the initially random texture R-1 under proportional biaxial strains.
In addition, an increase in ductility is achieved by subjecting texture FS-1 to non-proportional strain path B-3 as against the moderate change observed in texture R-1. It is therefore plausible to argue that texture affects achievable ductility under non-proportional paths and there exists a possibility of designing a suitable starting texture in addition to a favorable strain path in order to achieve higher ductility.

5. Conclusions

The effects of straining ferritic steel polycrystals under non-proportional paths have been investigated within a crystal plasticity finite element framework incorporating a hardening law to impose isotropic latent-hardening and (anisotropic) self-hardening respectively. Systematic studies have also been carried out for polycrystal aggregates in order to investigate the effects of initial texture in combination with non-proportionality. Also, ductility under non-proportional strain paths has been predicted. The results show that:

- The differences in dislocation distributions formed under isotropic latent-hardening are more pronounced in comparison to anisotropic self-hardening for non-proportional biaxial strain paths and similarly for uniaxial strains.
- Non-proportionality, texture and the nature of the hardening affect the predicted limit strain and increases in ductility are more apparent under non-proportional uniaxial strain paths as opposed to biaxial paths (evident from the predicted dislocation distributions).

References

Table 1 Material properties of ferritic steel obtained from experiment and used to fit simulation parameters

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Simulation parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield strength 150 MPa</td>
<td>$\rho_0$ $10^{9}$m$^2$, $E$ 211GPa</td>
</tr>
<tr>
<td>Tensile strength 380 MPa</td>
<td>$n$ 50, $\nu$ 0.3</td>
</tr>
<tr>
<td>Elastic modulus 211GPa</td>
<td>$\gamma_0$ 1s$^{-1}$, $\tau_0$ 70MPa</td>
</tr>
<tr>
<td></td>
<td>$m$ 0.245, $h_0$ 0.9GPa</td>
</tr>
</tbody>
</table>

Fig. 1. Systematic determination of material properties (a) uniaxial loading on RVE (b) averaged RVE stress-strain response to determine yield strength $\sigma_\gamma$. 

Fig. 2. The oligocrystal used in the simulations (showing 216 grains). The negative 1-, 2- and 3- surfaces are always constrained such that displacements in these directions are zero. Strain paths shown in Fig. 2 are applied to the positive 1- and 2- surfaces.

Fig. 3. Pole figures showing the initial textures used in the simulations (a) R-1: random texture and (b) FS-1: ferritic steel texture.

Fig. 4. Prediction of limit strain (a) schematic of proportional strain paths used to validate the failure criterion (b) comparison of limit strains predicted using the computational model as against experimental results from Volk and Hora [9].

Fig. 5. Displacement controlled non-proportional strain paths.
Fig. 6  Schematic representation of dislocation distribution for texture R-1 subjected to non-proportional strain paths to biaxial tension state. Note that the angle ranges indicate dislocation densities on similarly oriented slip systems within the polycrystal binned at 10 degree intervals.

\[ \begin{array}{c} a \end{array} \]

\[ \begin{array}{c} b \end{array} \]
Fig. 7 Dislocation distribution for textures R-1 and FS-1 under non-proportional strain paths to biaxial tension state. (a) dislocation distribution for R-1 at biaxial state under anisotropic self-hardening, (b) dislocation distribution of R-1 at final biaxial tension state under isotropic latent hardening (c) dislocation distribution for FS-1 at biaxial state under anisotropic self-hardening, (b) dislocation distribution of FS-1 at final biaxial tension state under isotropic latent hardening.

Fig. 8 Dislocation distribution for textures R-1 and FS-1 under non-proportional strain paths to uniaxial tension state. (a) dislocation distribution for R-1 at uniaxial state under anisotropic self-hardening, (b) dislocation distribution of R-1 at final uniaxial tension state under isotropic latent hardening (c) dislocation distribution for FS-1 at uniaxial state under anisotropic self-hardening, (b) dislocation distribution of FS-1 at final uniaxial tension state under isotropic latent hardening.
Fig. 9. Comparison of predicted ductility achieved under proportional biaxial (B-1) and uniaxial (U-1) strain paths as against the non-proportional cases indicated by B-3 and U-3 respectively. (a) under isotropic latent-hardening (b) anisotropic self-hardening.
THE GEOMETRIC PARADIGM IN COMPUTATIONAL ELASTO-PLASTICITY

G. Romano, R. Barreutta and F. Marotti de Sciarra
Department of Structures for Engineering and Architecture
Via Claudio 21, 80125 Naples, Italy
e-mail: romano@unina.it, rabarret@unina, marotti@unina.it

Key words: Continuum mechanics, Covariance, Finite deformations, Elasto-Plasticity, Computing Methods

Abstract. Computational methods, for large displacements of continua in the elasto-plastic range, rely on the mathematical modeling of the nonlinear constitutive behavior. In last decades an increasing favor has been deserved to nonlinear models based on a chain decomposition of the deformation gradient. The troubles involved in a structural analysis based on this model are well-known and have not been overcome although many efforts devoted to this end. Our investigation towards a more satisfactory model starts from the new analysis of the rate elastic behavior performed in [1, 2] since the difficulties faced by previous formulations were the very motivation for the discard of rate constitutive models in elasto-plasticity [3]. The new definition of hypo-elasticity, the detection of simple integrability conditions and a new formulation of conservativeness, lead to a definition of rate elasticity suitable for an effective modeling of rate elasto-plastic constitutive behaviors [4]. The treatment is based on a geometric definition of spatial and material fields and on the statement of a geometric paradigm assessing the rules for comparison of material fields naturally provided by push-pull according to the relevant transformation. The rates involved in constitutive relations are Lie-derivatives of stress field and constitutive parameters. Geometric compatibility requires that elastic and plastic stretchings additively give the Lie-derivative of metric field. No privileged reference configuration is involved and no consequent multiplicative decomposition of deformation gradient is assumed. Computational methods are shown to be based on the pull-back of constitutive relations to a straightened out trajectory segment which plays the role of computation chamber wherein linear operations of differentiation and integration may be performed. Accordingly, finite elastic and plastic stretches are considered as purely computational tools with no physical interpretation in constitutive relations. Both 3-D and lower dimensional structural models, such as wires and membranes, may be analyzed by a direct application of the theory. The outcome is a significant improvement of physical insight and computational effectiveness with respect to previous treatments of finite elasto-plasticity.
1 INTRODUCTION

To comply with principles of Geometric Naturality and Dimensionality Independence and with the dictates of the ensuing Geometric Paradigm, enunciated in [1, 2, 4], the elasticity model will be introduced with a rate formulation by a complete rephrasing of the original hypo-elastic model proposed by Truesdell [5]. The basic distinctive feature is that stress time-rate is defined in the natural way as Lie derivative of the stress field along the motion and that the constitutive law is inverted, so to provide the definition of elastic stretching. By virtue of this new definition, a simple and complete analysis of time-independence, integrability and conservativeness may be performed and this is a major merit of the geometric approach to elasticity. The consequent brand new analysis of rate constitutive relations, lead naturally to a clear, definite and computationally effective theory of elasto-plastic constitutive behavior in the non-linear range. The geometric treatment reveals that the rate formulation of the elasticity model, defined as a time-invariant and integrable hypo-elastic model, is self-proposing as natural and physically meaningful. Moreover, compatibility with a hyper-elastic behavior and conservation of elastic energy are ensured by conservation of mass and by directly verifiable integrability conditions. The key answer given by a proper geometric formulation concerns the way material tensors should be compared and differentiated in time along the trajectory. This requires a clear distinction between spatial and material tensor fields in continuum mechanics. The topic was first pointed out in [1] with reference to hypo-elasticity and then investigated in comprehensive geometrical terms in a space-time framework [2, 4]. Once the basic tools have been made available, the natural and geometrically consistent choice in elasto-plasticity is to formulate the rate constitututive behavior in terms of Lie-derivatives of the relevant material tensors along the trajectory, by performing an additive split of the total stretching into an elastic stretching and a plastic stretching. These latter are not Lie-derivatives of material fields, unless either the elastic or the plastic stretching vanishes. The hypo-elastic relation is governed by a constitutive operator which is non-linearly dependent on the stress tensor and provides the elastic stretching as a linear function of the stressing tensor. The visco-plastic flow rule is governed by a multi-valued operator. Indeed the current values of stress and tensorial internal state parameters determine a convex set of variation of the visco-plastic stretching. The evolution law for the internal parameters may be included in the visco-plastic flow by suitably enlarging the stress and the stretching spaces, for instance according to the treatment proposed in [6]. No privileged reference configuration nor mysterious intermediate states are needed. Finite plastic and elastic strains may however be conveniently introduced in reference linear spaces as purely computational tools to perform there, on suitable tensor fibers, basic linear operations such as time differentiations or integrations and approximate numerical evaluations. This computational trick is based on the peculiar properties of Lie derivatives under push-pull transformations, but no physical interpretation is given to finite plastic and elastic strains. The conclusions of the new theory restore credit to early computational
choices of a rate description of elasto-plastic behavior, but with the decisive improvement that the rules of the game are now clearly assessed. The theory leads to a conceptually clean, methodologically definite model and to drastic simplifications of both theoretical and computational aspects of geometrically non-linearized elasto-visco-plasticity.

2 CONSTITUTIVE LAWS

Let $E$ be a four-dimensional events manifold in which a nonlinear trajectory manifold $\mathcal{T}$, possibly lower dimensional, is immersed [4]. In a theoretical framework suitable for investigating a sufficiently general class of material behaviour for engineering applications, a constitutive law may be defined in terms of a constitutive operator $H_\mathcal{T}$, as follows.

**Definition 2.1 (Constitutive laws)** A constitutive operator $H_\mathcal{T}$, in a body motion detected by an Euclid observer, is a possibly multivalued material tensor bundle morphism whose domain and codomain are Whitney products\(^1\) of material tensor bundles.

The tensor bundle morphism requirement means that material tensors in the domain and codomain of the constitutive map should have the same base point in the tangent trajectory bundle, that is, they should be evaluated at a common event (i.e. a pair particle position, time instant) in the trajectory. The covariance (or geometric) paradigm [1, 2, 4] allows to compare the constitutive laws at displaced configurations of a body. Invariance of the constitutive law is the property that material tensor fields related by the constitutive law must be still related:

- either when transformed by push according to the relevant material displacement map $\varphi_{\tau,t}$, and this is the meaning of time invariance (TI) of the constitutive law,
- or when transformed by push according to a relative motion between observers, and this is the meaning of frame invariance (FI) of the constitutive law, which is named material frame indifference (MFI) when changes of Euclid observers are considered.

These definitions modify the standard TI and FI definitions as enunciated in [8] and quoted in subsequent literature, e.g. [9, 10, 11, 12, 13, 14, 15]. The change is from an invariance property to a property of variance by push [1, 2, 4]. In standard statements equality between constitutive responses measured by different observers is imposed without taking into account that constitutive operators to be compared do not have the same domain and codomain. This improper identification is responsible for the wrong assertion that FI implies isotropy, see [8] formula 99.5.

\(^1\)The Whitney product of two tensor bundles $(N, \pi_{M,N}, M)$ and $(H, \pi_{M,H}, M)$, over the same base manifold $M$, is the linear bundle defined by [7]:

$$N \times_M H := \{ (n, h) \in N \times H | \pi_{N,M}(n) = \pi_{M,H}(h) \}.$$
Definition 2.2 (Constitutive time invariance) According to the covariance paradigm, time invariance of the constitutive operator means that, along the motion:

$$H_{\tau,t} = \varphi_{\tau,t}^\dagger H_{\tau,t},$$

for any time instants $\tau, t \in I$. Explicitly the condition writes:

$$H_{\tau,t}(\varphi_{\tau,t}^\dagger s_{\tau,t}) = (\varphi_{\tau,t}^\dagger H_{\tau,t})(\varphi_{\tau,t}^\dagger s_{\tau,t}) = \varphi_{\tau,t}^\dagger (H_{\tau,t}(s_{\tau,t})).$$

This means that time-invariant material tensor fields, fulfilling the constitutive relation at time $t \in I$, are still related by the law at time $\tau \in I$.

2.1 HYPO-ELASTICITY

The hypo-elastic constitutive law is properly formulated as a linear dependence of the elastic stretching upon the stressing, by means of a stress dependent compliance constitutive operator, as follows [1].

Definition 2.3 (Hypo-elasticity) A hypo-elastic constitutive model is expressed by the law:

$$\text{el}_T = H_T(\sigma_T) \cdot \dot{\sigma}_T,$$

where $\text{el}_T \in C^1(T; \text{Cov}(\mathbb{V}_T))$ is the covariant elastic stretching and the stressing $\dot{\sigma}_T := L_{\nu_T} \sigma_T$ is the Lie derivative of the contravariant stress tensor $\sigma_T \in C^1(T; \text{Con}(\mathbb{V}_T))$, along the motion. A purely elastic process occurs when the elastic stretching is equal to the total stretching, i.e. $\text{el}_T := \varepsilon_T = \frac{1}{2} L_{\nu_T} g_T$, where $g_T \in C^1(T, \text{Sym}(\mathbb{V}_T))$ is the material metric [2].

Lemma 2.1 (Integrability) The hypo-elastic compliance operator is integrable if and only if the constitutive operator $H_T$ and its fiber derivative (i.e. the derivative taken at a fixed event in the trajectory) are symmetric:

$$\langle d_F H_T(\sigma_T) \cdot \delta \sigma_T \cdot \delta_1 \sigma_T, \delta_2 \sigma_T, \delta_1 \sigma_T \rangle = \langle d_F H_T(\sigma_T) \cdot \delta \sigma_T \cdot \delta_2 \sigma_T, \delta_1 \sigma_T \rangle,$$

$$\langle H_T(\sigma_T) \cdot \delta_1 \sigma_T, \delta_2 \sigma_T, \delta_1 \sigma_T \rangle = \langle H_T(\sigma_T) \cdot \delta_2 \sigma_T, \delta_1 \sigma_T \rangle,$$

for all $\delta \sigma_T, \delta_1 \sigma_T, \delta_2 \sigma_T \in C^1(T, \text{Sym}^+(\mathbb{V}_T))$. The fiber-derivative $d_F$ is taken by holding the base point fixed.

The former condition ensures CAUCHY integrability, stating the existence of a stretching-valued stress potential $\Phi_T \in C^1(\text{Con}(\mathbb{V}_T); \text{Cov}(\mathbb{V}_T))$ such that $d_F \Phi_T = H_T$. Both conditions ensure GREEN integrability, stating existence of a scalar-valued stress potential $E^*_T \in C^1(\text{Con}(\mathbb{V}_T); \text{Fun}(\mathbb{V}_T))$ such that:

$$d^2_F E^*_T = d_F \Phi_T = H_T.$$

Integrability of a time-invariant hypo-elastic constitutive operator, at a given time, implies integrability at every time.
ELASTICITY AND HYPER-ELASTICITY

Definition 3.1 (Elasticity) An elastic (resp. hyper-elastic) constitutive model is a time-invariant and Cauchy (resp. Green) integrable hypo-elastic model, so that:

\[ \mathbf{e}_T = d_T \Phi_T(\mathbf{\sigma}_T) \cdot \mathbf{\sigma}_T, \quad (d_T \Phi_T(\mathbf{\sigma}_T) \cdot \mathbf{\sigma}_T), \]

with time-invariance expressed by:

\[ \Phi_{T,t} = \varphi_{T,t} \Phi_{T,t}, \quad (E_{T,t}^* = \varphi_{t,t} E_{T,t}^*). \]

Denoting by \( \rho_T \in C^1(T; \text{Fun}(V_T)) \) the scalar density field, the material mass form is defined by \( m_T : = \rho_T \mu_T \in C^1(T; \text{Vol}(V_T)) \), where \( \mu_T \) is the material volume form [2]. The next result shows that conservation of mass and Green’s integrability of the hypo-elastic operator imply conservation of the mechanical energy [4].

Proposition 3.1 (Conservativeness) The constitutive operator of a hypo-elastic material, which is hyper-elastic when expressed in terms of the Kirchhoff stress tensor, is conservative, that is:

\[ \int_I \int_{\Omega_T} \langle \mathbf{\sigma}_{T,t}, \mathbf{e}_T, \mathbf{m}_{T,t} \rangle dt = 0, \]

for any covariantly closed stress path, i.e. any path such that its pull-back to any fixed reference placement is a cycle, a property expressed by the condition:

\[ \sigma_{T,t_2} = \varphi_{t_2,t_1} \sigma_{T,t_1}, \quad I = [t_1, t_2]. \]

3.1 Computational issues

In computational algorithms of the static evolution of an elastic structure undergoing large displacements, the equilibrium process is approximated by finite step incremental solutions in time. To underline the decisive role of a referential formulation, let us briefly describe the iterative scheme leading to the solution of the elastostatic problem in a finite time step. The control process is described by a time-parametrized curve \( c \in C^1(I; \text{C}) \) in a control set \( \text{C} \). The force acting on a body at time \( \tau \in I \) is assumed to depend on the control point \( c(\tau) \) and on the displacement \( \varphi_{\tau,t} \) from a given placement \( \Omega_t \), so that we may write

\[ f(\tau) = F(c(\tau), \varphi_{\tau,t}(\Omega_t)) : \Omega_t \mapsto T^*(\varphi_{\tau,t}(\Omega_t)). \]

An iterative trial and error procedure is then formulated as follows.

1. The start point is an equilibrium solution at time \( t_1 \in I \) under a force system

\[ f(t_1) = F(c(t_1), \text{id}_{\Omega_{t_1}}) : \Omega_{t_1} \mapsto T^*(\Omega_{t_1}) \]

and a stress field \( \sigma(t_1) \) fulfilling the virtual power variational equality

\[ \langle f(t_1), \delta v \rangle = \int_{\Omega(t_1)} \langle \sigma(t_1), \varepsilon(\delta v) \rangle m. \]
2. An initial guess of the displacement $\varphi_{t_{2}, t_{1}}$ corresponding to the update of the input control point from $c(t_{1})$ to $c(t_{2})$ may be obtained by computing the force system $F(c(t_{2}), ID_{t_{1}}) : \Omega_{t_{1}} \mapsto T(\Omega_{t_{1}})$ solution of the linear elastostatic problem

$$\langle F(c(t_{2}), ID_{t_{1}}), \delta v \rangle = \int_{\Omega(t_{1})} \langle H(\sigma(t_{1})) \cdot \varepsilon(u), \varepsilon(\delta v) \rangle m.$$ 

3. Setting $\varphi_{t_{2}, t_{1}}(x) = u(x) + x$ for any $x \in \Omega_{t_{1}}$, the control algorithm provides the trial force system

$$f(t_{2}) = F(c(t_{2}), \varphi_{t_{2}, t_{1}}) : \varphi_{t_{2}, t_{1}}(\Omega(t_{1})) \mapsto T(\varphi_{t_{2}, t_{1}}(\Omega(t_{1}))).$$

The trial referential finite step elastic strain is given by

$$e_{t_{2}, t_{1}} = \varphi_{t_{2}, t_{1}} \frac{1}{2} (\varphi_{t_{2}, t_{1}}^{-1} g \Gamma - g \Gamma).$$

The updated stress in the trial placement $\Omega(t_{2})$ is evaluated by

$$\sigma(t_{2}) = \varphi_{t_{2}, t_{1}} \cdot (\sigma_{t_{2}, t_{1}} + \Phi_{t_{2}, t_{1}}^{-1} e_{t_{2}, t_{1}}),$$

and the related elastic response is provided by the virtual power principle

$$\langle r(t_{2}), \delta v \rangle = \int_{\Omega(t_{2})} \langle \sigma(t_{2}), \varepsilon(\delta v) \rangle m.$$ 

4. The residual force gap

$$f(t_{2}) - r(t_{2}) : \varphi_{t_{2}, t_{1}}(\Omega(t_{1})) \mapsto T(\varphi_{t_{2}, t_{1}}(\Omega(t_{1}))),$$

is then applied to perform a correction to the previous guess concerning the displacement $\varphi_{t_{2}, t_{1}}$ by solving the linear elastostatic problem

$$\langle f(t_{2}) - r(t_{2}), \delta v \rangle = \int_{\varphi_{t_{2}, t_{1}}(\Omega(t_{1}))} \langle H(\sigma(t_{2})) \cdot \varepsilon(u), \varepsilon(\delta v) \rangle m,$$

and then performing the replacement $\varphi_{t_{2}, t_{1}}(x) \mapsto u(x) + \varphi_{t_{2}, t_{1}}(x)$ for any $x \in \Omega_{t_{1}}$.

The procedure is iterated on the new guess until the ratio between a suitable norm of the force gap and of the trial force system becomes less than a prescribed tolerance, thus reaching the approximated fixed point of the algorithm. The next time-step is then performed starting at the solution placement $\Omega(t_{2}) = \varphi_{t_{2}, t_{1}}(\Omega(t_{1}))$ under the force system $f(t_{2}) = F(c(t_{2}), \varphi_{t_{2}, t_{1}}) : \Omega_{t_{2}} \mapsto T(\Omega_{t_{2}})$. The hyper-elastic model can be extended to nonlocal models presented in [16, 17].
4 Elastic extension of a wire

Let us consider, as a simple example, a wire of initial length $L$ under the action of an axial force increment $F$. The elastic constitutive relation writes $E l := H_{\text{Mix}}(K) \cdot \dot{K}$ with $E l$ mixed elastic stretching, $K$ is mixed Kirchhoff stress and $\dot{K} := \dot{\sigma} \circ g_T$ mixed alteration of the Kirchhoff stressing $\sigma$. The hypo-elastic constitutive operator is given by

$$H_{\text{Mix}}(K) := \frac{1}{2 \mu} I - \frac{\nu}{E} I \otimes I.$$ 

The referential mixed elastic stretching is given by the pull-back formula $E l_{\text{ref}} = T \varphi^{-1}_{\text{ref}} \circ E l \circ T \varphi_{\text{ref}}$. The force increment is equal to 1 and the initial linearized response is equal to 1.5 for a Poisson ratio $\nu = 0.00$. The initial length is equal to 1 and is doubled by the first linearized estimate. Convergence features of the algorithm are shown in fig.1 for the values $\nu = 0.00, 0.30, 0.40, 0.49$. The physically significant example is the one with Poisson ratio $\nu = 0.49$, which is appropriate for a rubber wire.\(^2\)

![Figure 1: Elastic extension of a wire.](image)

Upper diagrams represent the length while lower diagrams represent the elastic response, as the iterative algorithm described in Sect.3.1 proceeds.

5 ELASTO-PLASTICITY

Once that the hypo-elastic model has been properly formulated and the right conditions for time invariance and conservativeness have been assessed, the rate model of elasto-visco-plastic constitutive behavior, which is of primary applicable interest in NLCM, may be

---

\(^2\)Computations and graphics implemented with Wolfram Mathematica 8.
readily described by the relations:

\[
\begin{cases}
\varepsilon_T = e_T + p_T, \\
e_T = d^2_T E^*_T(\sigma_T) \cdot \dot{\sigma}_T, \\
p_T \in \partial_F F_T(\sigma_T),
\end{cases}
\]

with \( F_T \subset \text{Fun}(\mathbb{V}T) \) a fiberwise subdifferentiable convex potential [18]. These constitutive relations are in fact extensions of the classical formula introduced, with reference to visco-elasticity, by James Clerk-Maxwell in [19]. Neither the elastic stretching \( e_T \in C^1(T, \text{Sym}(\mathbb{V}T)) \) nor the plastic stretching \( p_T \in C^1(T, \text{Sym}(\mathbb{V}T)) \) are convective time derivatives of a material field. The elastic and plastic stretchings should then not be denoted by a superimposed dot, as usually made in literature. The issue will be further addressed in Sect. 6. Elasto-plasticity is modeled by assuming that the stress potential is the indicator function of the convex set of admissible stresses \( K_T \subset \text{Sym}^+(\mathbb{V}T) \), so that:

\[
\partial_F F_T(\sigma_T) = N_{K_T}(\sigma_T),
\]

where \( N_{K_T}(\sigma_T) \) is the outward normal cone to \( \sigma_T \in K_T \). The visco-plastic constitutive relation specializes then into the plastic flow rule:

\[
p_T \in N_{K_T}(\sigma_T).
\]

A nonlocal model of elastoplasticity has been analysed in [20, 21]. Moreover a local elastoplastic behavior and a nonlocal damage model in the strain space has been addressed in [22, 23].

5.1 Computational issues

By a pull-back procedure the elasto-visco-plastic constitutive relations may be formulated in terms of material tensor fields defined in a fixed placement \( \Omega_{\text{REF}} \). Setting

\[
\begin{align*}
elel_{T,t}^{\text{REF}} &= \varphi_{t,\text{REF}}^{-1} e_T, \\
ppl_{T,t}^{\text{REF}} &= \varphi_{t,\text{REF}}^{-1} p_T,
\end{align*}
\]

with \( \varphi_{t,\text{REF}} : \Omega_{\text{REF}} \mapsto \Omega_t \), we get:

\[
\begin{cases}
\partial_{t=0} \varepsilon_{T,t}^{\text{REF}} = e_{T,t}^{\text{REF}} + p_{T,t}^{\text{REF}}, \\
e_{T,t}^{\text{REF}} = \partial_{t=0} d_T E^*_T(\sigma_{T,t}^{\text{REF}}), \\
p_{T,t}^{\text{REF}} \in \partial_F F_{T,t}^{\text{REF}}(\sigma_{T,t}^{\text{REF}}).
\end{cases}
\]

In an evolution process, the computations of the pull-back of the stress fields are conveniently carried out by a discrete time integration scheme and by an iterative algorithm, for the solution at each time step of the non-linear discrete constitutive relation, on the
basis of trial estimates of the elastic stretching evaluated at a fixed placement [24, 25]. In conclusions, the constitutive relations of elasto-visco-plastic behavior, in the non-linear geometric range, differ by the ones pertaining to the linearized theory just because the Lie derivatives of the material metric tensor and of the stress tensor are approximated, in the linearized theory, by partial time derivatives at a fixed point in space. An expression in terms of partial time derivatives may be got also in the non-linear geometric range by pulling back the constitutive relations to a fixed configuration. Numerical solution algorithms are then analogous to the ones adopted in the linearized theory, but with the additional task of taking care of the fact that the unknown differential of the displacement map, from a fixed to the current configuration, is involved in the expression of the referential stress.

6 FINITE ELASTIC AND PLASTIC STRAINS

Given a plastic stretching field on the trajectory, a finite plastic strain from an initial time \( t_o \) to the current time \( t \), may be defined by integration, in the time interval \( I = [t_o, t] \), of the plastic stretching field pulled-back from the current to a fixed configuration along the displacement \( \varphi_{t,\text{REF}} \in C^1(\Omega_t; \Omega_{\text{REF}}) \):

\[
\text{pl}^{\text{REF}}_{I,t} := \int_I \varphi_{t,\text{REF}} \downarrow \text{pl}_{I,t} \, dt \in C^1(\Omega_{\text{REF}}; \text{Cov}(T\Omega_{\text{REF}})).
\]

This operation defines a tensor field in the fixed configuration and, by push forward, also a tensor field in the current configuration. By performing a time-differentiation in the fixed configuration and then a push forward from the fixed to the current configuration, the plastic stretching field is recovered:

\[
\varphi_{t,\text{REF}} \downarrow \text{pl}_{I,t} = \partial_{\tau=t} \int_{t_o}^{\tau} \varphi_{\tau,\text{REF}} \downarrow \text{pl}_{I,\tau} \, d\tau \in C^1(\Omega_{\text{REF}}; \text{Cov}(T\Omega_{\text{REF}})).
\]

Anyway this fact does not motivate a definition of the plastic stretching as convective time derivative of the plastic finite strain, because the latter was not introduced independently of the notion of plastic stretching. The same occurs for elastic stretching fields and corresponding finite elastic strain fields. In fact the usefulness of finite plastic and elastic strains in the description of mechanical behaviors is rather questionable, the only directly definable finite strain being the total strain, which is expressed in terms of material displacements and of material metric tensors by:

\[
\varepsilon^{\text{REF}}_{I,t} := \frac{1}{2} \int_I \partial_{\tau=t} \varphi_{\tau,\text{REF}} \downarrow \text{g}_{I,\tau} \, dt = \frac{1}{2} \varphi_{t_2,\text{REF}} \downarrow \text{g}_{I,t_2} - \frac{1}{2} \varphi_{t_1,\text{REF}} \downarrow \text{g}_{I,t_1}.
\]

7 CONCLUSIONS

The findings of the geometric approach compel to perform a revisitation of theoretical contributions to non-linear constitutive laws adopted in most recent geometrically non-linear formulations of elasto-plasticity, elasto-visco-plasticity, poro-elasticity,
poro-plasticity, phase transformations, growth of biological tissues, and to carry out a consequent modification of relevant computational procedures. The results restore credit to early computational choices of a rate description of elasto-plastic behavior, but with the decisive improvement that the rules of the game are now clearly assessed also in the fully nonlinear range. The theory leads to a conceptually clean, methodologically definite model and to drastic simplifications of both theoretical and computational aspects of geometrically non-linearized elasto-visco-plasticity. The geometric approach to nonlinear continuum mechanics developed in this paper, with explicit application to the theory of elasto-plastic constitutive models, is a major step in a geometrization program of continuum mechanics carried out under several points of view in [26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 1, 37, 2, 4, 38].

REFERENCES


ON THE PERFORMANCE OF DETACHED-EDDY SIMULATION IN FLUID-STRUCTURE INTERACTION PROBLEMS

S. Türk*, T. Reimann†, D.C. Sternel† and M. Schäfer†

*Graduate School of Computational Engineering
Technische Universität Darmstadt
Dolivostrasse 15, 64293 Darmstadt, Germany
e-mail: tuerk@gsc.tu-darmstadt.de, web page: http://www.graduate-school-ce.de/

†Institute of Numerical Methods in Mechanical Engineering
Technische Universität Darmstadt
Dolivostrasse 15, 64293 Darmstadt, Germany
web page: http://www.fnb.tu-darmstadt.de/

Key words: Hybrid Methods, Detached-Eddy Simulation, Fluid-Structure Interaction

Abstract. In this work a formulation for a \( \zeta \)-f DES model has been used in its DES and URANS formulation to calculate the flow over an inclining plate in order to investigate the potential of DES on moving structures. It is shown that although the model is in LES mode, the fluctuations are not developed instantaneously. As a result, the turbulent kinetic energy is underpredicted in the LES zone, compared to pure URANS. The formulation for the \( \zeta \)-f DES model has been validated for the flow over a streamwise periodic 2D hill, whereat the results are in good agreement to the available benchmark data obtained from LES.

1 INTRODUCTION

Although the computational power available has been rapidly increasing over the years, it is still not feasible to run Direct Numerical Simulations (DNS) or Large-Eddy Simulations (LES) for high Reynolds number problems at a reasonable computational cost. Thus, RANS Models are still the preferred method in a broad range of applications for practically relevant Reynolds numbers [12].

To overcome the lack of accuracy in pure RANS methods, several hybrid approaches, combining LES and RANS methods, were developed within several research groups. One of these hybrid techniques is the detached-eddy simulation (DES) first proposed by Spalart [13]. Motivated by the fact that a Large-eddy simulation of complete configurations such as an airplane or a wind-turbine is not feasible for the foreseeable future, the DES aims at
combining the favorable aspects of LES and RANS techniques, namely the application of RANS models for the calculation of attached boundary layers and the LES for resolving spatio-temporal fluctuations of large eddies.

In applications of massively separated flows, the geometry-specific structures cannot be well captured by the statistical models as those are calibrated for thin turbulent shear flows containing relatively “standard” eddies [17]. Despite the broad range of RANS models available and their capabilities to predict boundary layers and their separation well, those models fail at prediction of large separation regions behind a sphere, past a vehicle or an airfoil in deep stall. In addition no unsteady information can be gained from a RANS simulation leading to vibration or noise.

Contrary to the RANS approach, where turbulence is completely modeled, an LES would resolve the fluctuations in space and time down to small scales and only the very small scales are modeled by a subgrid-scale (SGS) model. Hence the demands on the grid spacing and the time step are very restrictive.

Especially in the context of fluid-structure interaction problems (FSI), the information on the vibration can be one major result of the whole calculation. Concerning the fact that in addition to the fluid dynamics, also the equations of motion for the structural part have to be solved per iteration, the problems become much more demanding in terms of computational cost. Thus, the use of DES in the context of FSI seems to be promising to account for the important unsteady information.

The potential of DES for turbulence modeling employed to FSI is mentioned in [17], while some experiences are described in [3]. A coupling between DES flow simulation and FEM data is reported in [8], where the fluid-structure interaction in the nozzle section of the Ariane 5 under transonic wind tunnel conditions is investigated. Another work [4] using DES for the FSI of a transonic rotor at near-stall conditions showed that with such a computational setup, the basic dynamic aeroelastic characteristics have been captured successfully.

It has been reported, that the influence of the underlying RANS model is minor on the overall result in the LES region [15]. However for more complex flow situations arising particularly from movable or deformable objects the impact of the model in the RANS region is observable, particularly if separation occurs and thus justifying the use of a more complex RANS.

Within this paper, a DES formulation for the ζ-ż model by Hanjalic [6] is derived and validated with benchmark data from the flow over a streamwise periodic 2D hill. This formulation is then used to conduct a feasibility study for the use of DES in an FSI problem. A flat plate is inclined at a constant angular velocity of \( \dot{\alpha} = \frac{10\pi}{3} \text{ rad s}^{-1} \) from 0° to 45° at a Reynolds number of \( Re = 30000 \). The simulations are carried out using DES [13] and DDES [16] on two different grids. We put focus on the dynamics with which fluctuations are generated and compare the results to pure URANS.
2 GOVERNING EQUATIONS

In the following an incompressible fluid with constant fluid properties is considered. For an incompressible Newtonian fluid, the flow has to satisfy the continuity and Navier-Stokes equations in the following form:

\[ \frac{\partial u_i}{\partial x_i} = 0, \]  
\[ \rho \frac{Du_i}{Dt} = -\frac{\partial p}{\partial x_i} + \mu \frac{\partial^2 u_i}{\partial x_j^2}. \]

where \( u_i \) is the velocity vector, \( p \) is the static pressure and \( \mu \) is the dynamic viscosity. Using the Reynolds decomposition,

\[ u(x, t) = \overline{u(x)} + u'(x, t), \]

where \( \overline{u} \) is the mean motion and \( u' \) is the fluctuating component, the RANS equations can be derived from equation (2):

\[ \frac{\partial \overline{u}_i}{\partial x_i} = 0, \]
\[ \rho \frac{D\overline{u}_i}{Dt} = -\frac{\partial \overline{p}}{\partial x_i} + \mu \frac{\partial^2 \overline{u}_i}{\partial x_j^2} - \frac{\partial u'_i u'_j}{\partial x_j}. \]

Thus, equation (5) is the ensemble average of equation (2), where an additional term arises. This additional term describes the momentum transport due to turbulence motion causing a closure problem. Therefore, RANS turbulence models are used to model this new term. A detailed description of the RANS models used in this work can be found in [5] and [6].

3 MODELLING APPROACH

In the concept of detached-eddy simulations the computational domain is split into two parts, a RANS region and an LES region. In both cases the RANS turbulence model is either used as a normal turbulence model or used as a subgrid-scale (SGS) model for the LES mode, depending on the grid, the flow configuration and the location within the computational domain.

To modify a given RANS model to be used in DES, the dissipation term in the k-equation is replaced by an expression involving the turbulent length scale defined to

\[ l_{\text{turb}} = \min (l_{\text{RANS}}, C_{\text{DES}} \Delta), \]

where in the context of DES, \( \Delta \) is specifically defined by \( \Delta = \max (\Delta_x, \Delta_y, \Delta_z) \). Thus, the length scale becomes the original RANS length scale \( (l_{\text{RANS}} < C_{\text{DES}} \Delta) \) or the length scale
for the SGS model \((l_{\text{RANS}} > C_{\text{DES}}\Delta)\). For this method to be used together with low-Re turbulence models, an additional modification is necessary to eliminate the influence of the low-Re terms in the SGS mode. According to the concept introduced by Spalart [16], the DES length scale becomes

\[ l_{\text{DES}} = C_{\text{DES}}\Delta \psi, \tag{7} \]

where \(\psi\) is 1 for high-Re models and a function or a specific constant for low-Re models. The general idea of this \(\psi\)-function is based on the assumption that at equilibrium between the production and dissipation the RANS model in SGS mode should reduce to a Smagorinsky-like model, i.e. \(\nu_t = (C\Delta)^2 S\), where \(C\) is a constant and \(S\) is the modulus of the mean rate-of-strain tensor. For further information on the derivation of the \(\psi\)-function see [16], [10]. The modified transport equation for the turbulent kinetic energy for the \(\zeta\)-f model reads as

\[ \rho \frac{Dk}{Dt} = \frac{\partial}{\partial x_j} \left[ \left( \mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + P_k - \rho \frac{k^{3/2}}{l_{\text{turb}}} \tag{8} \]

The corresponding \(\psi\)-function is derived as

\[ \psi = \sqrt{\left( \frac{C_{\epsilon_1}}{C_{\epsilon_2} C_{\mu} \zeta} \right)^{3/2}} \tag{9} \]

The \(\zeta\)-f model used in this work is also used in a DES context in [1]. However, the \(\psi\)-functions derived are different but both plausible. The constant \(C_{\text{DES}}\) is calibrated using Decay of Isotropic Turbulence (DIT) to account for the different underlying turbulence models but also for the numerical schemes that are used. For the calibration two different grids, consisting of \(32^3\) and \(64^3\) grid points are used. The field is initialized with the data from [22] and the results are also compared to the DNS data from Wray [22]. For detailed description of the procedure, see [10]. For the 3D Fourier transform a tool written by St. Petersburg Technical University (SPTU) is used. In this work \(C_{\text{DES}}\) is calibrated to

\[ C_{\text{DES}} = 0.2 \tag{10} \]

Due to some drawbacks of the original DES formulation from equation (6) which we will refer to as DES97 throughout this paper, a modified version was suggested by Spalart [16] named DDES. In order to preserve RANS mode in the boundary layer also on grids where the wall parallel spacing is smaller than the turbulent length scale, causing a switch to LES mode, a blending function is introduced. The blending function limits the LES mode to regions outside the boundary layer. For the DDES, the formulation reads:

\[ r_d = \frac{\nu_t + \nu}{\sqrt{U_{i,j}^2 U_{i,j}^2}} \tag{11} \]
where $\nu_t$ is the kinematic eddy viscosity, $\nu$ is the molecular viscosity, $U_{i,j}$ are the velocity gradients, $\kappa$ is the Kármán constant and $d$ is the wall distance. The quantity $r_d$ is then used in

$$f_d = 1 - \tanh \left( [8r_d]^3 \right)$$

(12)

in such a way that $f_d$ is one in the LES region and zero elsewhere. Therefore, the DES length scale is re-defined as:

$$l_{DDES} = l_{RANS} - f_d \max (0, d - C_{DES} \Delta \psi)$$

(13)

4 NUMERICAL CONDITIONS AND METHODS

In the present work a Detached-Eddy Simulation is performed for a flat plate changing its angle of attack from 0° to 45° at a rotational speed of

$$\dot{\alpha} = \frac{10\pi}{3} \text{ rad s}^{-1}$$

To avoid discontinuities in the angular motion, the following expression is used to describe the motion of the plate:

$$\alpha(t) = \frac{\pi}{4} \left[ \sin \left( \frac{\pi t}{T} \right) \right]^2$$

where $T = 0.15$ s is one period. Thus, for a motion from 0° to 45°, only $T/2$ is considered. The reduced pitching frequency for a non periodic motion is defined to $k_{pitch} = \Delta \alpha c/(U_b \Delta t) = 0.168$, where $U_b = 3.75 \text{ m s}^{-1}$ is the bulk velocity at the inlet. The applied boundary conditions and the computational domain are shown in figure 1, where $x, y$ and $z$ correspond to the streamwise, spanwise and wall-normal direction respectively. At top and bottom of the computational domain no-slip boundary conditions are applied as well as on the plate, located in the center. In spanwise direction, periodic boundary conditions are used. On the inlet plane, a profile of a fully developed turbulent channel flow without any perturbations is applied while on the outlet plane a zero gradient boundary condition is used. The size of the computational domain is chosen according to the future experimental setup, where the measurements are taken in a square channel of a cross section of $0.45 \times 0.45$ m$^2$ and a length of 2 m. The chord length is $c = 0.12$ m$^2$ and the Aspect Ratio is $AR = 3.67$. The thickness is 0.006 m$^2$.

In the computational setup only two chord lengths in spanwise direction are resolved in order to reduce the computational cost, which can be justified by the two-dimensional character of the experimental setup. Due to the fact that the trailing edge of the plate is moving towards the bottom wall, top and bottom walls are resolved in the simulation. In this study two different grids were used to evaluate the influence of the grid on the performance of the DES. A summary of the two grids is given in table 1. Both grids are built following the guidelines given in [14]. To ensure the grid quality in the region around the plate, where the vortex shedding and recirculation is expected, the O-Grid blocks are rotated with the plate. Thus, no re-meshing or grid generation method is needed in the
Figure 1: Sketch of computational domain

Table 1: General grid information for the simulations performed

<table>
<thead>
<tr>
<th>Total No. of CVs</th>
<th>CVs in circumferential direction of the plate</th>
<th>CVs in spanwise direction per chord length</th>
<th>Wall normal grid spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2 × 10^6</td>
<td>136</td>
<td>16</td>
<td>y_{max} &lt; 1</td>
</tr>
<tr>
<td>2.6 × 10^6</td>
<td>240</td>
<td>32</td>
<td>y_{max} &lt; 1</td>
</tr>
</tbody>
</table>

region of interest leading to a decrease in the computational cost.

We use the finite volume solver FASTEST [18] with block structured, boundary fitted grids. The convective and diffusive fluxes are approximated with second-order central difference schemes in the LES mode and with the GAMMA [7] scheme in RANS mode. The blending between CDS and GAMMA is done by a blending function introduced in [20]. For the time discretization, we use a backward differencing scheme with second order accuracy and the coupling between pressure and velocity is done with the SIMPLE algorithm. The parallelization in FASTEST is done via domain decomposition using MPI.

5 RESULTS AND DISCUSSION

5.1 Flow over streamwise periodic 2D hill

To validate the DES formulation for the ζ-f model [6], the flow over a streamwise periodic 2D hill for a Reynolds number of $Re_b \approx 10600$ is calculated and compared to benchmark data. This benchmark scenario has already been investigated for DES in several studies, for example [11] and [21]. The geometry here is the same as in [19] and the results are compared to the benchmark LES from Temmerman and Leschziner [19], serving as the standard reference data. All quantities are nondimensionalized by the hill
height $h$ and the bulk velocity at the hill crest, $U_b$. The results are averaged in time and spanwise direction. The time step was chosen so that $CFL_{\text{max}} \approx 1$ to ensure the temporal resolution, while the effect of higher $CFL$ numbers is reported in [11]. The domain is discretized by approximately $1.7 \times 10^6$ control volumes, while the wall distance of the first grid point is about $y^+ \approx 0.5$.

The velocity profiles as well as the $\overline{u'v'}$ component of the Reynolds stress tensor, shown in figure 2, are in good agreement with the benchmark LES. The separation point is located at $x/h = 0.14$ while the flow reattaches at $x/h = 4.6$, whereas the reference points in [19] are at $x/h = 0.22$ and $x/h = 4.72$ respectively. Thus, both points are slightly shifted towards the crest of the hill.

5.2 Flow over an inclining plate

The main objective of this work is the further investigation of DES methods in the context of FSI focused on the generation of fluctuations for rapidly moving structures. For the given inflow conditions, the flow field does not contain any fluctuations at the beginning, but a vortex street behind the plate at zero angle of attack is developed resulting in the formation of a small LES region, including fluctuations. For the DDES test cases,
Table 2: Overview of the simulations performed

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>DES mode</th>
<th>SGS/RANS model</th>
<th>grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddes-f</td>
<td>DDES</td>
<td>ζ-f</td>
<td>fine</td>
</tr>
<tr>
<td>des97-f</td>
<td>DES97</td>
<td>k-ζ-f</td>
<td>fine</td>
</tr>
<tr>
<td>ddes-c</td>
<td>DDES</td>
<td>ζ-f</td>
<td>coarse</td>
</tr>
<tr>
<td>rans-f</td>
<td>RANS</td>
<td>ζ-f</td>
<td>fine</td>
</tr>
</tbody>
</table>

the major part of the computational domain is treated in RANS mode. For the DES97, almost in the entire region in front and around the plate, the RANS model is operating as an SGS model but not creating any fluctuations. Consequently the turbulent viscosity, turbulent kinetic energy and the dissipation is reduced in the DES97 case. However, the velocity profile does not alter significantly downstream up to the plate, which might be attributed to the chosen boundary conditions at in and outlet. The log-layer mismatch observed in periodic channel flows is not found here.

In the upstroke phase ($0 < \alpha < 45^\circ$), all configurations in table 2 predict a similar shape for the lift coefficient, $C_L$, and the drag coefficient, $C_D$, as shown in figure 3. No significant difference between RANS and DES can be detected but the grid refinement shows some influence on the results. For all cases no outstanding fluctuations are generated and the flow has a strongly two-dimensional character. This two dimensional behavior in the upstroke phase is also observed by Martinat et al. [9], where the flow past a pitching airfoil is investigated. For the downstroke phase on the other hand, they reported a reinforced adverse pressure gradient leading to a highly three dimensional flow. In the present work, this downstroke phase is excluded and thus not reinforcing three dimensional patterns. The LES zone is initialized at the lee side and consequently resulting in a reduced modeled turbulent kinetic energy, turbulent viscosity and dissipation, compared to URANS.

Once the plate remains at a constant angle of attack of $\alpha = 45^\circ$, fluctuations start to generate and become more dominant over time but not as much as expected. A deviation in $C_L$ and $C_D$ starts evolve for $2t/T > 1.5$. Comparing the velocity field and the corresponding streamlines at $2t/T = 1.8$ in figure 4, the vortex at the trailing edge is still pinned to the plate for the ddes-f and ddes-c case, whereas the vortex is already shed in the other two configurations. For the coarser grid, only one big and flat vortex at the leading edge is resolved compared to the fine grid. For the DDES cases the LES zones is growing in time but the resolved fluctuations are not developed in a large number. However, for a pure URANS simulation resolving the spanwise direction the flow field presented in figure 5 shows the most turbulent fluctuations of all cases considered and does not converge to a steady state. Comparing the modeled turbulent kinetic energy, shown in figure 6, it becomes clear that the modeled contribution is reduced in the LES region while at the same time, fluctuations have not been developed yet as illustrated in
figure 5, leading to an underprediction of the total turbulent kinetic energy as opposed to URANS. Against expectation the DES97 case does not develop an LES Zone comparable to the DDES cases and thus treating almost the entire lee side in RANS mode, which explains the good agreement between des97-f and rans-f in terms of $C_L$ and $C_D$.

Simulating further in time, the problem then results in the case of a flat plate at high incidence which has been investigated by Breuer et al. [2], who showed the capability of DES in that specific flow situation, but is out of focus in this work. The fluctuations are then clearly detectable also in our case.
6 CONCLUSIONS

We derived a DES formulation for the $\zeta$-f model by Hanjalic [6] and validated the expressions with the benchmark data from [19] for a streamwise periodic 2D hill. The results are in good agreement with the LES data. Furthermore it is shown that the use of DES97 and DDES is not advantageous over URANS for a ramping plate without separation at the beginning. Future work will be dealing with the investigation of IDDES for the rapidly inclining plate, which is promising in terms of missing fluctuations, as the region upstream of the plate should already contain those fluctuations, arising from the WMLES branch of IDDES. Additionally the results will be compared to experiments to perform a quantitative analysis. The simulations will also be repeated for the oscillatory motion.
7 AFFILIATION

This work is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

REFERENCES


SHAPE AND STRESS SENSING
OF MULTILAYERED COMPOSITE AND SANDWICH STRUCTURES
USING AN INVERSE FINITE ELEMENT METHOD

PRISCILLA CERRACCHIO*, MARCO GHERLONE*, MARCO DI SCIUVA* AND
ALEXANDER TESSLER†

* Department of Mechanical and Aerospace Engineering, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: {priscilla.cerracchio, marco.gherlone, marco.disciuv}@polito.it, web: www.aesdo.polito.it

† Structural Mechanics and Concepts Branch, NASA Langley Research Center
Mail stop 190, Hampton, Virginia 23681-2199, U.S.A
e-mail: Alexander.tessler-1@nasa.gov, web: www.nasa.gov

Key words: Composite Structures, Sandwich Structures, Shape Sensing, Stress Sensing,
Inverse Finite Element Method, Inverse Plate Element.

Abstract. The marked increase in the use of composite and sandwich material systems in
aerospace, civil, and marine structures leads to the need for integrated Structural Health
Management systems. A key capability to enable such systems is the real-time reconstruction
of structural deformations, stresses, and failure criteria that are inferred from in-situ, discrete-
location strain measurements. This technology is commonly referred to as shape- and stress-
sensing. Presented herein is a computationally efficient shape- and stress-sensing
methodology that is ideally suited for applications to laminated composite and sandwich
structures. The new approach employs the inverse Finite Element Method (iFEM) as a general
framework and the Refined Zigzag Theory (RZT) as the underlying plate theory. A three-
node inverse plate finite element is formulated. The element formulation enables robust and
efficient modeling of plate structures instrumented with strain sensors that have arbitrary
positions. The methodology leads to a set of linear algebraic equations that are solved
efficiently for the unknown nodal displacements. These displacements are then used at the
finite element level to compute full-field strains, stresses, and failure criteria that are in turn
used to assess structural integrity. Numerical results for multilayered, highly heterogeneous
laminates demonstrate the unique capability of this new formulation for shape- and stress-
sensing.

1 INTRODUCTION

The inverse problem of shape- and stress-sensing is manifested by reconstruction of
structural displacements, strains, stresses, and failure criteria using real-time strain
measurements. Aircraft wings with embedded conformal antennas and those of morphed
capability require real-time shape sensing to provide feedback for their actuation and control
systems. For structural health monitoring, shape- and stress-sensing technologies are the enabling capabilities for assessing structural integrity and cost-efficient maintenance. For composite and sandwich structures, structural health monitoring using embedded optical-fiber networks presents an attractive technology for in-situ strain measurements that give rise to a large amount of strain data. Despite their numerous advantages, composite structures may experience such modes of failure as delamination and impact damage, and these can affect their load carrying capabilities. Thus, the monitoring of structural integrity of multilayered composite and sandwich structures is an issue of primary importance.

Various shape-sensing approaches for plates undergoing bending deformations have been explored [1–3]. Bogert et al. [1] examined a modal transformation method, which requires a large number of natural vibration modes. Using classical bending assumptions, Jones et al. [2] performed the reconstruction of plate deflections by fitting discrete measures of the bending curvatures and then integrating the regression curves. Nishio et al. [3] explored shape-sensing of thin laminated composite plates by enforcing compatibility between analytic and measured strains in a weighted least-squares sense. Both [2] and [3] employed Kirchhoff plate theory, thus restricting their methods to homogeneous or nearly homogeneous thin plates. Tessler and Spangler [4] proposed a general framework for full-field reconstruction of displacements, strains, and stresses, using arbitrary positioned strain sensors on the load-carrying structural surfaces. The methodology is based on a least-squares variational principle and accounts for the complete set of First-order Shear Deformation Theory (FSDT) modes. They include stretching, bending and shear deformations. The variational principle [4] is also well suited for finite element approximations. In [5], the authors proposed an inverse Finite Element Method (iFEM) based on $C^0$-continuous kinematic approximations. This resulted in an efficient three-node inverse shell finite element called iMIN3.

Although generally regarded as an accurate theory, FSDT may lead to somewhat inadequate predictions when applied to relatively thick composite and sandwich structures. Higher-order equivalent-single-layer theories [6] provide improved predictions for such structures, specifically for the global response quantities such as deflection and natural frequency; nevertheless, even these theories fail to predict through-the-thickness distributions of displacements, strains, and stresses with sufficient accuracy. Layer-wise theories [7] usually lead to highly accurate response predictions; however, these are obtained at the expense of computational efficiency and modeling complexity, especially for multilayered structures, since the number of unknowns depends on the number of material layers. The recently developed Refined Zigzag Theory (RZT) [8,9] is a good compromise between adequate accuracy and computational efficiency. For plate analysis, RZT has seven kinematic variables regardless of the number of material layers, just two more than FSDT. This new theory is able to model the cross-sectional distortion that is typical of multilayered composite and sandwich structures.

In this paper, the Tessler-Spangler [4,5] iFEM formulation is reformulated to include the kinematic assumptions of RZT [8,9]. The new formulation is thus intended for applications dealing with multilayered composite and sandwich structures possessing a high degree of anisotropy and heterogeneity. The variational principle is then discretized using a $C^0$-continuous three-node inverse plate finite element. Numerical results are presented for moderately thick sandwich laminates subjected to various boundary and loading conditions. Finally, superior stress-sensing capabilities of the present formulation are demonstrated for a
select set of challenging material systems.

2 KINEMATIC ASSUMPTIONS OF THE REFINED ZIGZAG THEORY FOR PLATES

Herein the kinematic assumptions of the Refined Zigzag Theory (RZT) for plates are briefly reviewed. In particular, the strain field is formally re-written in order to define the strain measures to be used in the iFEM variational formulation (see Sect. 3).

Consider a plate of thickness $2h$ made of $N$ perfectly bonded orthotropic material layers (see Figure 1(a)); the superscript $(k)$ denotes the $k$th layer. The plate is referred to a Cartesian coordinate system $(x_1, x_2, z)$ where $(x_1, x_2)$ are the in-plane coordinates and $z$ is the thickness coordinate that ranges from $-h$ to $h$, with $z = 0$ identifying the mid-plane and $z_j$ identifying the $j$th interface (see Figure 1(b)).

The displacement field of RZT for plates is \[ u_{12}^{(k)}(x_1, x_2, z) = u(x_1, x_2) + z \theta_1(x_1, x_2) + \phi_{1\alpha}^{(k)}(z) \psi_1(x_1, x_2) \]
\[ u_{21}^{(k)}(x_1, x_2, z) = v(x_1, x_2) + z \theta_2(x_1, x_2) + \phi_{2\alpha}^{(k)}(z) \psi_2(x_1, x_2) \]
\[ u_z(x_1, x_2, z) = w(x_1, x_2) \]

where $u_1^{(k)}$ and $u_2^{(k)}$ are the in-plane displacements and $u_z$ is the transverse displacement. RZT has seven kinematic variables, \( \mathbf{u} = [u, v, w, \theta_1, \theta_2, \psi_1, \psi_2]^T \). $u$, $v$, and $w$ are the uniform displacement components along the $x_1$, $x_2$, and $z$-axis respectively; $\theta_1$ and $\theta_2$ are the average rotations of the transverse normal around the positive $x_2$-axis and the negative $x_1$-axis, respectively; and $\psi_\alpha$ ($\alpha = 1, 2$) are the amplitudes of the zigzag contributions to the in-plane displacement in the $x_\alpha$-directions (see Figure 1(a)). The zigzag terms $\phi_{\alpha}^{(k)} \psi_\alpha$ ($\alpha = 1, 2$) in Eq. (1) describe the $C^0$-continuous cross-sectional distortions that are typical of multilayer laminates. The zigzag functions, $\phi_{\alpha}^{(k)}(z)$, have units of length and are piecewise linear, $C^0$-continuous functions of the thickness coordinate and of the transverse shear moduli of the laminate layers. Refer to [9] for the detailed derivation of the zigzag functions.

![Figure 1: Plate notation (a) and layer notation for a three-layer laminate (b)](image)
The in-plane components of the strain field are given as
\[
\begin{align*}
\varepsilon_{11}^{(k)} &= \varepsilon_{10} + z \kappa_{10} + \mu_{10}^{(k)} \\
\varepsilon_{22}^{(k)} &= \varepsilon_{20} + z \kappa_{20} + \mu_{20}^{(k)} \\
\gamma_{12}^{(k)} &= \gamma_{120} + z \kappa_{120} + \mu_{120}^{(k)}
\end{align*}
\]
where
\[
\begin{align*}
\epsilon(u) &= \begin{pmatrix} \varepsilon_{10} \\ \varepsilon_{20} \\ \gamma_{120} \end{pmatrix} = \begin{pmatrix} u_1 \\ v_2 \\ u_2 + v_1 \end{pmatrix}, & \kappa(u) &= \begin{pmatrix} \kappa_{10} \\ \kappa_{20} \\ \kappa_{120} \end{pmatrix} = \begin{pmatrix} \theta_{11} \\ \theta_{22} \\ \theta_{12} + \theta_{21} \end{pmatrix}
\end{align*}
\]
represent the membrane, bending and zigzag strain measures, respectively. Note that, whereas the membrane and bending strain measures are constant with respect to the thickness coordinate, the zigzag strain measures have the “zigzag”, z-dependent distributions.

The transverse shear strains are given by
\[
\begin{align*}
\gamma_{1z}^{(k)} &= \eta_1 + \left(1 + \beta_{1}^{(k)} \right) \psi_1 \\
\gamma_{2z}^{(k)} &= \eta_2 + \left(1 + \beta_{2}^{(k)} \right) \psi_2
\end{align*}
\]
where \( \beta_{\alpha}^{(k)} = \phi_{\alpha z}^{(k)} (\alpha = 1, 2) \) and
\[
\begin{align*}
g(u) &= \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 - \psi_1 \\ \gamma_2 - \psi_2 \end{pmatrix} = \begin{pmatrix} w_1 + \theta_1 - \psi_1 \\ w_2 + \theta_2 - \psi_2 \end{pmatrix}
\end{align*}
\]
with \( \eta_{\alpha} (\alpha = 1, 2) \) denoting the transverse-shear strain measures of RZT.

3 INVERSE FINITE ELEMENT METHOD BASED ON RZT

In this section, the iFEM variational formulation based on the RZT kinematics is presented. The formulation enables reconstruction of the deformed shape of composite and sandwich structures from in situ strain measurements.

The general framework is that of finite element approximations. Thus, a discretization of the structure with plate elements is introduced, in which the element kinematic variables are interpolated by a set of suitable shape functions,
\[
u(x_1, x_2) = u^e = N(x_1, x_2) q^e
\]
where \( N(x_1, x_2) \) denotes the shape functions and \( q^e \) the nodal degrees-of-freedom.
3.1 Error functional

Following the iFEM methodology [4,5], the displacement solution is obtained through the minimization of an error functional, which is defined as the least-square error between the analytic strain measures (Eqs. (3), (5)) and their measured values, known at discrete locations from in-situ strain measurements. For a single element, the error functional is given as

\[
\Phi^e(\mathbf{u}^e) = \left\| \mathbf{e}(\mathbf{u}^e) - \mathbf{e}^e \right\|^2 + \left\| \mathbf{k}(\mathbf{u}^e) - \mathbf{k}^e \right\|^2 + \left\| \mathbf{m}^{(k)}(\mathbf{u}^e) - \mathbf{m}^e \right\|^2 + \alpha \left\| \mathbf{g}(\mathbf{u}^e) \right\|^2
\]

(7)

where the superscript “e” is used to denote the measured values; for the membrane, bending and zigzag contributions, the squared norms that appear in Eq. (7) have the form

\[
\left\| \mathbf{e}(\mathbf{u}^e) - \mathbf{e}^e \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbf{e}(\mathbf{u}^e) - \mathbf{e}^e \right]_{i}^2
\]

(8)

\[
\left\| \mathbf{k}(\mathbf{u}^e) - \mathbf{k}^e \right\|^2 = \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ \mathbf{k}(\mathbf{u}^e) - \mathbf{k}^e \right]_{i}^2
\]

\[
\left\| \mathbf{m}^{(k)}(\mathbf{u}^e) - \mathbf{m}^e \right\|^2 = \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ \mathbf{m}^{(k)}(\mathbf{u}^e, z_{(j)}) - \mathbf{m}^e \right]_{i}^2
\]

where \( n \) is the number of locations \( P_i = (x_i, y_i) \), where the strain measures, \( \mathbf{e}^e, \mathbf{k}^e \) and \( \mathbf{m}^e \) are evaluated from strain-sensor measurements. The zigzag strain measures, \( \mathbf{m}^{(k)}_{(j)} \), are evaluated at the \( j \)th interface using embedded strain-sensors. The transverse shear strain measures, \( \mathbf{g}(\mathbf{u}) \), cannot be obtained experimentally. Thus, the transverse shear term in Eq. (7) is given by the \( L^2 \) norm

\[
\left\| \mathbf{g}(\mathbf{u}^e) \right\|^2 = \frac{1}{A^e} \int_{A^e} \left[ \mathbf{g}(\mathbf{u}^e) \right]_{i}^2 dA
\]

(9)

where \( A^e \) denotes the element area. In Eq. (8), \( \alpha \) is a positive valued, small (compared to unity) weighting coefficient; refer to the numerical studies in Sect. 4 in which the value of \( \alpha = 10^{-5} \) is used.

Considering Eqs. (3) and (5), while invoking the kinematic field interpolations given by Eq. (6), the strain measures can be expressed in terms of the nodal degrees-of-freedom, \( \mathbf{q}^e \), as

\[
\mathbf{e}(\mathbf{u}^e) = \mathbf{B}_e \mathbf{x}_1 \mathbf{q}^e, \quad \mathbf{k}(\mathbf{u}^e) = \mathbf{B}_k \mathbf{x}_2 \mathbf{q}^e
\]

\[
\mathbf{m}^{(k)}(\mathbf{u}^e, z) = \mathbf{B}_m \mathbf{x}_3 \mathbf{q}^e, \quad \mathbf{g}(\mathbf{u}^e) = \mathbf{B}_g \mathbf{x}_4 \mathbf{q}^e
\]

(10)

where the matrices \( \mathbf{B}_e, \mathbf{B}_k, \mathbf{B}_m \) and \( \mathbf{B}_g \) contain the derivatives of the shape functions \( \mathbf{N}(x_1, x_2) \).

Substitution of Eq. (10) into Eqs.(8) and (9), and minimization of the functional \( \Phi^e \) with respect to the unknown nodal degrees-of-freedom, \( \mathbf{q}^e \), yields the element matrix equation

\[
\mathbf{a}^e \mathbf{q}^e = \mathbf{b}^e
\]

where the matrix \( \mathbf{a}^e \) is given as
\[
a^e = \frac{1}{n} \sum_{i=1}^{n} \left[ B^e_i \left(x_1, x_2, x_3, x_4\right) \right] + \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ B^e_k \left(x_1, x_2, x_3, x_4\right) \right]
\]

\[
+ \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ B^m_i \left(x_1, x_2, z_i\right) \right] B^m_\left(x_1, x_2, z_i\right) \left[ B^e \left(x_1, x_2, x_3, x_4\right) \right]
\]

\[
+ \alpha \frac{1}{n} \sum_{i=1}^{n} \left[ B^e_i \left(x_1, x_2, x_3, x_4\right) \right]
\]

whereas the right-hand-side vector is

\[
b^e = \frac{1}{n} \sum_{i=1}^{n} \left[ B^e_i \left(x_1, x_2, x_3, x_4\right) \right] e^e + \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ B^e_k \left(x_1, x_2, x_3, x_4\right) \right] k^e
\]

\[
+ \frac{(2h)^2}{n} \sum_{i=1}^{n} \left[ B^m_i \left(x_1, x_2, z_i\right) \right] m^e
\]

Taking into account appropriate coordinate transformations, the element contributions are assembled into a global system of equations. Upon enforcement of problem-dependent displacement boundary conditions that prevent rigid-body motion, the inverse-problem equations take on the form

\[
Aq = b
\]

where \( A \) is a well-conditioned square matrix.

Equation (13) is well suited for real time applications. This is because \( A \) is inverted only once (assuming small displacements, the strain-sensor locations, \([(x_1, x_2), ±h]\) and \([(x_1, x_2), (z_i)\], remain unchanged.) On the other hand, the vector \( b \) needs to be updated at each strain-data acquisition increment. Thus, the displacement solution is efficiently computed by the vector-matrix multiplication, \( q = A^{-1}b \). The strains and stresses are readily computed for each element using strain-displacement and constitutive relations of RZT. Furthermore, strains and stresses can be used to construct failure criteria.

### 3.2 Evaluation of the experimental strain measures from strain-sensor data

The present method is especially aimed at structures with embedded strain sensors. In particular, a strain-sensor configuration that measures the in-plane strains is considered on the top and bottom surfaces and at one interface (refer to Eq. (2)). Considering that the zigzag contributions to the in-plane strains vanish on the top and bottom surfaces [9], the expressions for the experimental membrane and bending strain measures are derived by evaluating Eqs. (2) at the discrete locations \( P_i = (x_1, x_2, z) \), \( z = ±h \),

\[
\begin{align*}
\epsilon_i^e &= \begin{pmatrix} \epsilon_{11}^e \\ \epsilon_{22}^e \\ \gamma_{12}^e \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \epsilon_{11}^e \\ \epsilon_{22}^e \\ \gamma_{12}^e \end{pmatrix}, & \kappa_i^e &= \begin{pmatrix} \kappa_{10}^e \\ \kappa_{20}^e \\ \kappa_{12}^e \end{pmatrix} = \frac{1}{2} h \begin{pmatrix} \epsilon_{11}^e \\ \epsilon_{22}^e \\ \gamma_{12}^e \end{pmatrix} - \begin{pmatrix} \epsilon_{11}^e \\ \epsilon_{22}^e \\ \gamma_{12}^e \end{pmatrix}
\end{align*}
\]

where \( \{\epsilon_{11}^e, \epsilon_{22}^e, \gamma_{12}^e\} \) and \( \{\epsilon_{11}^e, \epsilon_{22}^e, \gamma_{12}^e\} \) denote the in-plane strains measured respectively at \([(x_1, x_2), +h]\) and \([(x_1, x_2), -h]\). Measuring the same strain components at the \( j \)th interface,
and again using Eq. (2), the zigzag strain measures are determined as

$$m^e_{i,j} \equiv \begin{pmatrix} \mu_{10} \\ \mu_{20} \\ \mu_{120} \end{pmatrix}_{i,j} = \begin{pmatrix} \varepsilon_{11}^j \\ \varepsilon_{22}^j \\ \gamma_{12}^j \end{pmatrix} - \begin{pmatrix} \varepsilon_{10}^j \\ \varepsilon_{20}^j \end{pmatrix} - z_{i,j} \begin{pmatrix} \kappa_{10} \\ \kappa_{20} \end{pmatrix}_{i,j}$$ \quad (15)

Using this strain-sensor configuration, the zigzag strain measures, $m^e_{i,j}$, are evaluated at the $j$th interface only.

## 4 THREE-NODE PLATE INVERSE ELEMENT BASED ON RZT

In this section, a brief description of a three-node inverse plate element is presented. The element has seven degrees-of-freedom at each node, $u_k, v_k, w_k, \theta_{d_k}, \theta_{b_k}, \psi_{s_k}, \psi_{v_k}$, where $k = 1, 2, 3$ is an index ranging over the three nodes. Considering that only first derivatives of the kinematic variables appear in the functional, Eq. (7), the shape functions are required to satisfy $C^0$-continuity. For this purpose, the so-called anisoparametric interpolations are used (e.g., refer to Tessler and Hughes [12] and Versino et al. [11]).

The in-plane displacements, bending rotations and zigzag amplitudes are interpolated using linear shape functions. The functions are defined in terms of the area-parametric coordinates $L_k$ ($k = 1, \ldots, 3$),

$$\chi(x_1, x_2) = \sum_{k=1}^{3} L_k (x_1, x_2) \chi_k, \quad \chi = (u, v, \theta, \psi)$$ \quad (16)

whereas a quadratic interpolation is used for the deflection

$$w(x_1, x_2) = \sum_{k=1}^{3} L_k w_k + \sum_{k=1}^{3} L_{1k} (\theta_{d_k} - \psi_{s_k}) + L_{2k} (\theta_{b_k} - \psi_{v_k})$$ \quad (17)

where the quadratic shape functions, $L_{1k}$ and $L_{2k}$, are given by the expressions

$$L_{1k} \equiv \frac{L_k}{2} \left( a_k L_m - a_m L_k \right) \quad L_{2k} \equiv \frac{L_k}{2} \left( b_k L_q - b_q L_k \right)$$ \quad (18)

where the subscripts are given by the cyclic permutation of $k = 1 - 3$, $l = 2, 3, 1$ and $m = 3, 1, 2$.

The anisoparametric interpolations, Eqs. (16) and (17), ensure that (a) truly thin plates can be modeled without any stiffening due to shear locking, and (b) the resulting element, herein referred to as irrZT3, has the same number of degrees-of-freedom as a standard linear-interpolation element.

## 5 NUMERICAL EXAMPLES

A symmetric three-layer sandwich plate, with carbon-epoxy face-sheets and a PVC core is analyzed, under the action of different loadings and boundary conditions (see Table 1 for the
mechanical material properties). The thickness of each face-sheet is 10% of the total laminate thickness. In the following example problems, square plates are considered having edge length \( a \) and span-to-thickness ratios, \( \rho = a/2h = 10 \). Regular mesh patterns are used having the same number of elements along the plate edges, \( n' \). Figure 2 shows the plate geometry, the coordinate system, and a \( 2 \times 2 \) discretization \( (n' = 2) \).

Two example problems are analyzed: (1) a simply supported plate subjected to a transverse bi-sinusoidal pressure \( p(x_1, x_2) = p_0 \sin(\pi x_1/a) \sin(\pi x_2/a) \), and (2) a cantilever plate subjected to a uniform transverse pressure \( p(x_1, x_2) = p_0 \). In lieu of the actual experimental strain measurements, an exact elasticity solution by Pagano [13] is used for the simply supported plate. For the cantilever plate, an analytic RZT-based direct solution is used, [9]. In both cases, the strain sensor configuration consists of three strain-sensor rosettes located at the element centroid on top and bottom surfaces and at \( z_{(1)} \) (see Figures 1 and 2).

Table 1: Mechanical properties of orthotropic (C) and isotropic (P) materials. The Young’s moduli and the shear moduli are expressed in GPa.

<table>
<thead>
<tr>
<th></th>
<th>Carbon-epoxy unidirectional composite</th>
<th>PVC core</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_{kk}^{(k)} )</td>
<td>( E_{11}^{(k)}, E_{22}^{(k)}, E_{33}^{(k)} )</td>
<td>( E, \nu )</td>
</tr>
<tr>
<td>( \nu_{ik}^{(k)} )</td>
<td>( 1.579 \times 10^5, 9.584, 9.584 )</td>
<td>( 1.040 \times 10^{-1}, 0.3 )</td>
</tr>
<tr>
<td>( G_{ij}^{(k)} )</td>
<td>( 0.32, 0.32, 0.49 )</td>
<td>( 5.930, 5.930, 3.227 )</td>
</tr>
</tbody>
</table>

Figure 2: Inverse plate finite element discretization, \( 2 \times 2 \) mesh \((n' = 2)\), and strain-sensor locations.

In Figures 3–5, the in-plane displacement, in-plane normal stress, and transverse shear stress are normalized as

\[
\bar{u}_i = u_i \times \left( 10^4 D_{11} / (p_0 a^4) \right), \quad \bar{\sigma}_{11} = \sigma_{11} \times \left( 1 / (p_0 \rho^2) \right), \quad \bar{\tau}_{2z} = \tau_{2z} \times \left( 1 / (p_0 \rho) \right)
\]

where \( D_{11} \) is the bending stiffness coefficient.

Figures 3–4 depict results for the simply supported plate where the displacements and stresses obtained using the present element formulation (labeled as iRZT3) are compared to Pagano’s exact elasticity solution. In Figure 3(b), for comparison purposes results are also depicted for the predecessor inverse element “iMIN3” [5], based on FSDT. When using \( n' = 12 \) (Figure 3(a)), the present iRZT3 model accurately predicts the maximum deflection (error within 0.27%) whereas iMIN3 leads to an underestimation by 70% (not shown). Figures 3(b), 4(a) and 4(b) show that through-the-thickness distributions of in-plane displacement, in-plane
normal stress and transverse shear stress are accurately recovered using iRZT3 whereas iMIN3 is unable to model the zigzag shape of the in-plane displacement (Figure 3(b)) typical of a sandwich-like stacking sequence.

**Figure 3**: Simply supported plate, $a/2h=10$: (a) Percent error of maximum deflection, $e_w = 100\left(\frac{w}{w_{\text{Pagano}}} - 1\right)$, vs. the number of elements along the plate edge, $n_e'$; and (b) Through-the-thickness distribution of in-plane displacement, $\bar{u}_i(0,a/2)$, for the discretization $n_e' = 12$.

**Figure 4**: Simply supported plate, $a/2h=10$: through-the-thickness distributions of (a) In-plane stress, $\sigma_{11}(a/2,a/2)$, and (b) Transverse shear stress, $\tau_{2z}(0,a/2)$, for the discretization $n_e' = 12$.

In Figure 5, results are depicted for the cantilever plate, where the iFEM predictions are
compared with the corresponding RZT analytic solutions of the direct problem. Figure 5(a) depicts rapid convergence of the maximum deflection error, whereas Figure 5(b) shows a through-the-thickness distribution of the normal stress, $\sigma_{11}(a/4, a/2)$.

Figure 5: Cantilever plate, $a/2h=10$: (a) Percent error of maximum deflection, $e_w = 100(w/w_{RZT} - 1)$, vs. the number of elements along the plate edge, $n^e$; and (b) Through-the-thickness distribution of in-plane stress, $\sigma_{11}(a/4, a/2)$, for the discretization $n^e=20$.

6 CONCLUSIONS

Real-time reconstruction of structural displacements and stresses from in-situ discrete-location strain measurements, herein referred to as shape- and stress-sensing, is an inverse problem that has important implications for monitoring of structural integrity, as well as for the actuation and control of smart structures. A new variational formulation for shape- and stress-sensing of laminated composite and sandwich plates has been presented. The approach makes use of an inverse Finite Element Method (iFEM) that was previously developed on the basis of First-order Shear Deformation Theory (FSDT). The iFEM is based on the minimization of a least-square error functional in which analytic and measured strains are enforced in the least-square sense. Within the present formulation, the kinematic assumptions of the Refined Zigzag Theory (RZT) are used and require that the measured strains be available along three surfaces through the laminate thickness (instead of only two surfaces for the FSDT-based formulation.) The RZT has previously been shown to be especially well suited for the modeling of laminated composite and sandwich plates. The error functional is discretized using $C^0$-continuous interpolations of the displacement field, yielding an efficient three-node inverse plate finite element that has seven kinematic degrees-of-freedom at each node. From the reconstructed element displacements, strains and stresses at every material point of the structure are then computed using RZT’s strain-displacement and constitutive relations. Numerical results for moderately thick laminated sandwich plates undergoing
elasto-static deformations demonstrated superior full-field predictions for the displacements, strains, and stresses. As expected, the present iFEM-RZT formulation results in more accurate predictions than those based on iFEM-FSDT formulation, and this is especially evident in highly heterogeneous and sandwich laminates.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of this research by the NASA Aviation Safety Program (Vehicle Systems Safety Technologies Project). Special thanks are extended to the National Institute of Aerospace (NIA) and NASA Langley research Center, Hampton, Virginia, for hosting Professors Marco Di Sciuva and Marco Gherlone, as well as doctoral student Priscilla Cerracchio in the fall of 2012. Priscilla Cerracchio, Marco Gherlone and Marco Di Sciuva also acknowledge the financial support from the European Commission in the framework of FP7-CLEANSKY-GRA project.

REFERENCES


MASS PRESERVING DISTRIBUTED LANGRAGE MULTIPLIER APPROACH TO IMMERSED BOUNDARY METHOD

DANIELE BOFFI∗, NICOLA CAVALLINI∗, FRANCESCA GARDINI∗ AND LUCIA GASTALDI†

∗Dipartimento di Matematica “F. Casorati”
Università degli Studi di Pavia
Via Ferrata 1, 27100 Pavia, Italy
{daniele.boffi,nicola.cavallini,francesca.gardini}@unipv.it

†DICATAM
Dipartimento di Ingegneria Civile, Architettura, Territorio, Ambiente e di Matematica
Università degli Studi di Brescia
via Branze 43, 25123 Brescia, Italy
lucia.gastaldi@ing.unibs.it

Key words: Finite Elements, Immersed Boundary Method, Fluid-Structure Interactions, Mass conservation, Fictitious Domain.

Abstract. This research is devoted to mass conservation and CFL properties of the Finite Elements Immersed Boundary Method. We first explore an enhanced higher order scheme applied to the Finite Element Immersed Boundary Method technique introduced by Boffi and Gastaldi. This technique is based on a Pointwise (PW) formulation of the kinematic condition, and higher order elements show better conservation properties than the original scheme. A further improvement with respect to the classical PW formulation is achieved introducing a fully variational Distributed Lagrange Multiplier (DLM) formulation. Numerical experiments show that DLM is not affected by any CFL condition. Furthermore the mass conservation properties of this method are extremely competitive.

1 INTRODUCTION

Several applications require the modeling of fluid structure interactions. In particular, the Immersed Boundary Method (IBM) has been developed to tackle the biomedical modeling of organic tissues. The IBM has been firstly considered by Peskin in the seventies, see [13] for a review.

Boffi and Gastaldi in [3] introduced a finite element version of the IBM which enjoys interesting properties both for practical and theoretical aspects (see [4, 5, 7, 6, 8, 11, 2]).
Interesting investigations concern, in particular, the choice of the time advancing strategy and the mass conservation of the resulting scheme. In this respect, the main contribution of the present work is twofold. From one side, we explore the use of higher order finite element for the approximation of the fluid; from the other side, we examine a different, in a sense more natural, strategy for handling the movement of the immersed structure. Higher order finite elements are constructed extending the ideas of [1] for the enhancement of continuous pressure fluid schemes. The original formulation of the IBM considers a pointwise evolution (PW) of the immersed structure; on the other hand, in this paper we consider a variational formulation of the structure evolution equation (see also [12]). This gives rise to a scheme which shares several analogies with the Fictitious Domain method with Distributed Lagrange Multiplier (see, for instance, [10]); for this reason we shall refer to this scheme as DLM approach.

It turns out that the first modification of the original IBM formulation (higher-order fluid scheme) provides significant improvements of the mass conservation and that the second modification (DLM) gives rise to a scheme which is superior in terms of time advancing procedure (no CFL condition is required for its stability). Surprisingly enough, we observed that the DLM approach enjoys much better mass conservation properties as well, thus becoming a very promising scheme for future studies.

2 PROBLEM SETTING

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be the fluid domain, and let the structure domain $B_t \subset \Omega$ be immersed into the fluid one. We consider incompressible fluid and viscoelastic structure. The key assumption of the IBM lays in the particular form of the structure stress tensor. In fact we assume the structure stress tensor to be composed by an elastic and a viscous contribution. The viscous contribution has exactly the same form as for the fluid, then the principle of virtual work applies.

More in detail, at time $t$ the structure lays on the time dependent domain $B_t$. The structure domain can be parametrized by a map $X$ on a reference domain $B \subset \mathbb{R}^m$, $m = d, d-1$. In a Lagrangian framework we set $s$ as the reference variable in $B$. A material point on the current domain $B_t$ is denoted by $x$. A typical assumption is $\partial B_t \cap \partial \Omega = \emptyset$. We consider the initial domain as the reference one $B = B_0$.

The map $X$ represents the relationship between the current and the reference domains:

$$X : B \times [0,T] \rightarrow B_t \quad \text{so that } x = X(s, t) \forall x \in B_t. \quad (1)$$

We assume $X(s, t)$ being invertible at any time, which implies that the deformation gradient $F_{\alpha i} := (\nabla_s X(s, t))_{\alpha i} = X_{\alpha i}(s, t) = \frac{\partial X_{\alpha}(s, t)}{\partial s_i}$ has rank $m$. We have that $|F| = 1$ at the initial time; thanks to the incompressibility assumption this is true also at any subsequent time. Here $|F|$ stands for the determinant of $F$ in the case $m = d$. When $m = d-1$ we set $|F| = |\partial X/\partial s|$ for $m = 1$ and $|F| = |\partial X/\partial s_1 \wedge \partial X/\partial s_2|$ for $m = 2$. 

324
The kinematics of the material particle can be defined through its velocity:

\[ u(x, t) = \frac{\partial X}{\partial t}(s, t) \text{ for } x = X(s, t). \]  

(2)

The densities of the incompressible fluid and solid phases are assumed to be piecewise constant:

\[ \rho = \begin{cases} 
\rho_f & \text{in } \Omega \setminus B_t \\
\rho_s & \text{in } B_t.
\end{cases} \]  

(3)

Detailed study of stability criteria regarding the ratio \( \rho_s/\rho_f \) can be found in [2]. We already mentioned that the key idea of IBM lays in the definition of the Cauchy stress tensor \( \sigma \) as:

\[ \sigma = \begin{cases} 
\sigma_f & \text{in } \Omega \setminus B_t \\
\sigma_f + \sigma_s & \text{in } B_t.
\end{cases} \]  

(4)

Here we have the viscous term all over the domain, and the elastic contribution is taken into account where the structure is located. This assumption is accepted in biological frameworks where the viscoelasticity of the tissues play a key role (see, e.g., [14]). The stress tensor for a viscous fluid is:

\[ \sigma_f = -pI + \mu (\nabla u + (\nabla u)^T). \]  

(5)

The first Piola–Kirchhoff stress tensor can be derived from the elastic stress tensor \( \sigma_s \) using Lagrangian variables as:

\[ P(s, t) = |F(s, t)| \sigma_s(X(s, t), t) F^{-T}(s, t). \]  

(6)

Using the principle of virtual work and equations (3)-(6) we obtain the following formulation of the problem:

**Problem 2.1** Given \( u_0 \in H^1_0(\Omega)^d \) and \( X_0 : \mathcal{B} \to \Omega \) such that \( X_0 \in W^{1,\infty}(\mathcal{B}) \), for all \( t \in [0, T] \), find \( (u(t), p(t)) \in H^1_0(\Omega)^d \times L^2(\Omega) \) and \( X(t) \in W^{1,\infty}(\mathcal{B}) \), such that

\[
\rho_f \frac{d}{dt} (u(t), v) + b(u(t), u(t), v) + a(u(t), v)
\]

\[ - (\nabla \cdot v, p(t)) = \langle d(t), v \rangle + \langle F(t), v \rangle \quad \forall v \in H^1_0(\Omega)^d \]  

(7a)

\[ \langle \nabla \cdot u(t), q \rangle = 0 \quad \forall q \in L^2(\Omega) \]  

(7b)

\[ \langle d(t), v \rangle = - (\rho_s - \rho_f) \int_{\mathcal{B}} \frac{\partial^2 X}{\partial t^2} \nabla v(X(s, t)) \, ds \quad \forall v \in H^1_0(\Omega)^d \]  

(7c)

\[ \langle F(t), v \rangle = - \int_{\mathcal{B}} \mathbb{P}(F(s, t)) : \nabla v(X(s, t)) \, ds \quad \forall v \in H^1_0(\Omega)^d \]  

(7d)

\[ \frac{\partial X}{\partial t}(s, t) = u(X(s, t), t) \quad \forall s \in \mathcal{B} \]  

(7e)

\[ u(x, 0) = u_0(x) \quad \forall x \in \Omega \]  

(7f)

\[ X(s, 0) = X_0(s) \quad \forall s \in \mathcal{B}. \]  

(7g)
We have used the following definitions for the bilinear forms

\[
\begin{align*}
    a(u, v) &= \mu(\nabla_{\text{sym}} u, \nabla_{\text{sym}} v), \\
    b(u, v, w) &= \frac{\rho_f}{2} \left( (u \cdot \nabla v, w) - (u \cdot \nabla w, v) \right),
\end{align*}
\]

where \( \nabla_{\text{sym}} u = (\nabla u + (\nabla u)^T)/2 \).

We observe that in the above formulation we have that the incompressible Navier–Stokes equations \((7a)-(7b)\) are written in variational form, while the movement of the structure governed by \((7e)\) is written pointwise, since at each point \( s \in \mathcal{B} \) we solve an ordinary differential equation with initial value given by \((7g)\). For this reason we address this formulation as PW in the rest of the paper.

We assume that the structure is composed by a hyperelastic material. These materials are characterized by a positive energy density \( W(F) \) depending only on the deformation gradient. Then the first Piola–Kirchhoff stress tensor \( P \) can be expressed in terms of the potential energy density as \( P(F(s, t))_{\alpha i} = \partial W(F(s, t))/\partial F_{\alpha i} \) where \( i = 1, \ldots, m \) and \( \alpha = 1, \ldots, d \); the elastic potential energy of the body is given by:

\[
E(X(t)) = \int_{\mathcal{B}} W(F(s, t)) ds.
\]

Assuming that the potential energy density \( W \) is convex, that \( \rho_s \geq \rho_f \), and using the relation between the energy density and the Piola–Kirchhoff above, the following energy estimate can be obtained for all \( t \in [0, T] \):

\[
\frac{\rho_f}{2} \frac{d}{dt} ||u(t)||_0^2 + \mu ||\nabla u(t)||_0^2 + \frac{\rho_s - \rho_f}{2} \frac{d}{dt} \left\| \frac{\partial X}{\partial t} \right\|_{0, \mathcal{B}}^2 + \frac{d}{dt} E(X(t)) = 0. \quad (8)
\]

In the following section we present the space-time discretization of Problem 2.1 based on the use of finite elements, and in the next one we shall introduce a variational version also of equation \((7e)\). The issue of the stability for the space-time scheme will be also addressed.

### 3 MASS PRESERVING HIGH ORDER FINITE ELEMENT SPACES

In this section we discretize Problem 2.1 using high order mass preserving stable finite elements. Consider a triangulation \( \mathcal{T}_h \) of \( \Omega \) into triangles or rectangles if \( d = 2 \), and tetrahedrons or parallelepipeds if \( d = 3 \). We denote by \( K \) any single element of \( \mathcal{T}_h \).

The structure domain \( \mathcal{B} \) is subdivided into segments, triangles or tetrahedrons for \( m = 1, 2, 3 \) respectively; we denote this subdivision by \( \mathcal{S}_h \). We shall need also the following notation: \( T_k, k = 1, \ldots, M_e \) is an element of \( \mathcal{S}_h \); \( s_j, j = 1, \ldots, M \) stands for a vertex of \( \mathcal{S}_h \), and \( \mathcal{E}_h \) is the set of the edges (or faces) \( e \) of \( \mathcal{S}_h \). Given these definitions we introduce the solution space for the structure position \( S_h \):
\[ S_h = \{ Y \in C^0(B; \Omega) : Y|_{T_k} \in P_1(T_k)^d, \ k = 1, \ldots, M_e \}, \quad (9) \]

being \( P_1(T_k) \) the space of affine polynomials on the element \( T_k \).

As far as the fluid domain is concerned, we consider two finite dimensional spaces \( V_h \subseteq H^1_0(\Omega)^d \) and \( Q_h \subseteq L^2_0(\Omega) \) for the velocity and pressure discretizations, respectively. It is well known that these two spaces need to satisfy the inf-sup condition [9]. We recall also that, due to the presence of the source term localized along the structure, the pressure might present jumps along the interface between fluid and solid. For this reason, pairs of finite element spaces enjoying the so called local mass conservation property are to be preferred. Discontinuous pressure schemes enjoy this property, however many popular Stokes elements, as e.g. Hood–Taylor and Bercovier–Pironneau elements, are based on continuous pressure. In [1] we proposed and analyzed an enhancement of such elements by adding to the pressure space piecewise constant functions. In this work we consider higher polynomial degree and higher enhancing shape functions with respect to [1]. More precisely, we consider the Hood–Taylor element of degree 2, that is the velocities are piecewise cubic polynomials while the pressures are continuous piecewise quadratic. Then we add to the pressure space discontinuous piecewise affine functions. Hence the resulting two-dimensional finite element spaces are:

\[
V_h = \{ v \in H^1_0(\Omega)^d : v|_K \in P_3(K)^d \ \forall \ K \in T_h \} \\
Q_h = \{ q \in L^2_0(\Omega) : q = q_2 + q_1, \ q_2 \in C^0(\bar{\Omega}), \ q_1|_K \in P_2(K), \ q_1|_K \in P_1(K) \ \forall \ K \in T_h \}. \quad (10)
\]

Let us introduce a subdivision of the interval \((0, T)\) in \( N \) equal parts with size \( \Delta t \), then \( t_n = n\Delta t \) and \( u^n \) stands for the value of the function \( u \) at time \( t_n \). Similar notation holds for the other functions involved in the problem. Applying a modified backward Euler scheme to equations (7a)-(7b) and (7e) and using the latter in order to discretize the source term (7c) we end up with the following time advancing scheme.

**Step 1.** Compute

\[
\langle F^{n+1}_h, v \rangle = - \sum_{e \in \mathcal{E}_h} [F_h]^n \cdot v(x^n_h(s, t)) \, dA \quad \forall v \in V_h. \quad (11)
\]

**Step 2.** Solve the Navier–Stokes equations: find \( (u^{n+1}_h, p^{n+1}_h) \in V_h \times Q_h \) such that

\[
\rho_f \left( \frac{u^{n+1}_h - u^n_h}{\Delta t} , v \right) + b(u^{n+1}_h, u^{n+1}_h, v) + a(u^{n+1}_h, v) - (\nabla \cdot v, p^{n+1}_h) = \\
- (\rho_s - \rho_f) \int_B \frac{u^{n+1}_h(x^n_h(s)) - u^n_h(x^{n-1}_h(s))}{\Delta t} \cdot v(x^n_h(s)) \, ds + \langle F^{n+1}_h, v \rangle \quad \forall v \in V_h \quad (12)
\]

\[
(\nabla \cdot u^{n+1}_h, q) = 0 \quad \forall q \in Q_h.
\]
Step 3. Advance the position of the points of the structure:

$$\frac{X_{hi}^{n+1} - X_{hi}^{n}}{\Delta t} = u_{h}^{n+1}(X_{hi}^{n}) \quad \forall i = 1, \ldots, M. \quad (13)$$

We remark here that the expression we have introduced to compute the force (7d) takes into account the fact that $X_h$ is piecewise affine so that $\mathbb{P}(\mathbb{F})$ is piecewise constant.

The space-time scheme described above satisfies an energy estimate similar to (8) provided the following CFL condition is verified

$$\mu - \kappa_{\max} C \frac{h^{m-2}}{h^{d-1}} L^m C^n \geq 0 \quad (14)$$

where $L^n := \max_{T_k \in S_h} \left\{ \max_{s_j,s_i \in V(T_k)} |X_{hj}^{n} - X_{hi}^{n}| \right\}$.

4 DISTRIBUTED LAGRANGE MULTIPLIER FORMULATION

In the PW approach we advance the structure moving each point according to equation (7e). We introduce now a different variational approach whose time-space discretization seems to enjoy better stability conditions.

First of all we write equation (7e) in weak form as follows:

$$\left\langle \mu, u(X(\cdot,t),t) - \frac{\partial X(\cdot,t)}{\partial t} \right\rangle = 0 \quad \forall \mu \in (H^1(\mathcal{B})^d)^* \quad (15)$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^1(\mathcal{B})^d$ and its dual space $(H^1(\mathcal{B})^d)^*$. The notation $(\cdot,\cdot)_{\mathcal{B}}$ stands for the $L^2$-scalar product in $L^2(\mathcal{B})$. Then introducing a Lagrange multiplier associated to the above constraint, Problem 2.1 can be reformulated as follows:

**Problem 4.1** Given $u_0 \in H^1_0(\Omega)^d$ and $X_0 \in H^1(\mathcal{B})$, find $(u(t), p(t)) \in H^1_0(\Omega)^d \times L^2_0(\Omega)$, $X(t) \in H^1(\mathcal{B})$, and $\lambda(t) \in ((H^1(\mathcal{B}))^d)^*$, such that for almost every $t \in [0,T]$ it holds

$$\rho_f \frac{d}{dt}(u(t), v) + b(u(t), u(t), v) + a(u(t), v)$$

$$- (\nabla \cdot v, p(t)) + \langle \lambda(t), v(X(\cdot,t)) \rangle = 0 \quad \forall v \in H^1_0(\Omega)^d \quad (16a)$$

$$(\nabla \cdot u(t), q) = 0 \quad \forall q \in L^2_0(\Omega) \quad (16b)$$

$$(\rho_s - \rho_f) \left( \frac{\partial^2 X}{\partial t^2}(t), Y \right)_B + (\mathbb{P}(\mathbb{F}(t)), \nabla_s Y)_B - \langle \lambda(t), Y \rangle = 0 \quad \forall Y \in (H^1(\mathcal{B}))^d \quad (16c)$$

$$\left\langle \mu, u(X(\cdot,t),t) - \frac{\partial X(\cdot,t)}{\partial t} \right\rangle = 0 \quad \forall \mu \in ((H^1(\mathcal{B}))^d)^* \quad (16d)$$

$$u(0) = u_0 \quad \text{in } \Omega, \quad X(0) = X_0 \quad \text{in } \mathcal{B}.$$

We consider $\Lambda_h = \{ \mu \in C^0(\mathcal{B};\Omega) : \mu|_{T_k} \in P_1(T_k)^d, \quad k = 1, \ldots, M_s \}$; then the space-time discretization of Problem 4.1, obtained by approximating the time derivatives in (7a), (16c) and (16d) with proper finite differences, reads:
Problem 4.2 Given $u_{0,h} \in V_h$ and $X_{0,h} \in S_h$, for $n = 1, \ldots, N$ find

$$(u^n_{h}, p^n_{h}) \in V_h \times Q_h, \quad X^n_{h} \in S_h, \quad \lambda^n_{h} \in \Lambda_h,$$

such that $u^0_{h} = u_{0,h}$, $X^0_{h} = X_{0,h}$ and

$$\rho_f \left( \frac{u^{n+1}_{h} - u^n_{h}}{\Delta t}, v \right) + b(u^{n+1}_{h}(t), u^{n+1}_{h}(t), v) + a(u^{n+1}_{h}, v)$$

$$- (\nabla \cdot v, p^{n+1}_{h}) + \langle \lambda^{n+1}_{h}, v(X^n_{h}) \rangle = 0$$

$$\forall v \in V_h$$

$$(\nabla \cdot u^{n+1}_{h}, q) = 0$$

$$\forall q \in Q_h$$

$$(\rho_s - \rho_f) \left( \frac{X^{n+1}_{h} - 2X^n_{h} + X^{n-1}_{h}}{\Delta t^2}, Y \right)_{B} + (P(X^{n+1}_{h}), \nabla \cdot Y)_{B}$$

$$- \langle \lambda^{n+1}_{h}, Y \rangle = 0$$

$$\forall Y \in S_h$$

$$\langle \mu, u^{n+1}_{h}(X^n_{h}) - \frac{X^{n+1}_{h} - X^n_{h}}{\Delta t} \rangle = 0$$

$$\forall \mu \in \Lambda_h.$$ 

We remark that in Problem 4.2 we have evaluated the nonlinear terms involving the position of the structure at the previous time. It remains only one possible nonlinear contribution which depends on the expression of $P$. For clearness we write the time advancing scheme in matrix form in the linear case $P(F) = \kappa F$. Consider time $t^n = n \Delta t$ and solve for $t^{n+1}$:

$$\begin{pmatrix}
(\frac{\rho_f}{\Delta t} M_u + A) & B^\top & 0 & G(X^n_{h})^\top \\
B & 0 & 0 & 0 \\
0 & 0 & \frac{\delta \rho}{\Delta t^2} M_X + K & -J^\top \\
G(X^n_{h}) & 0 & \frac{1}{\Delta t} J & 0
\end{pmatrix} \begin{pmatrix}
u^{n+1}_{h} \\
p^{n+1}_{h} \\
X^{n+1}_{h} \\
\lambda^{n+1}_{h}
\end{pmatrix} = \begin{pmatrix}
\frac{\rho_f}{\Delta t} M_u u^n_{h} \\
0 \\
\frac{\delta \rho}{\Delta t^2} M_X (2X^n_{h} - X^{n-1}_{h}) \\
\frac{1}{\Delta t} J X^n_{h}
\end{pmatrix}$$

with the following definitions of the involved matrices:

- $M_u$, $M_X$: fluid and structure mass matrices,
- $A$: matrix associated to the convective term and the fluid stiffness,
- $G(X)$: matrix associated to $\langle \mu, v(X) \rangle$,
- $B$: matrix associated to the divergence,
- $K$: structure stiffness matrix,
- $J$: matrix associated to $\langle \mu, Y \rangle$.

5 NUMERICAL EXPERIMENTS

In this section we provide numerical experiments illustrating the techniques depicted in this work. We first control the performances of the high order enhanced finite elements defined in (10), both in terms of approximation and local mass conservation. We start with the PW scheme described in Sect. 3. A typical test case for IBM is the simulation of an immersed co-dimension one elastic string. In this case we take $P = \kappa F$. The domain
Figure 1: Stationary solution for an elastic circle immersed into the fluid, error for the pressure. In the pictures we plot only the upper right corner of the solution.

Ω is the unit square and we consider a regular mesh obtained by dividing the unit square domain into $N \times N$ squares subdivided into two triangles, with mesh size $h_x$. The reference domain $B$ is the unit interval divided into $M$ subintervals with size $h_x$. When the initial configuration of the string is circular, we have an analytical solution, characterized by a discontinuous pressure, the interested reader can refer to [8] for the details. In Fig. 1, we represent the error distribution for the pressure in the upper right quarter of the domain. In order to visually stress the differences among the different finite element performances we represent the error on a coarse mesh obtained by dividing $\Omega$ into $16 \times 16$ squares. Due to the discontinuity of the pressure one can appreciate oscillations around the position of the immersed elastic string which can be interpreted as a sort of Gibbs phenomenon. In the case of $P_3/P_2$ finite element (the standard Hood–Taylor element of degree 2), the area involved by the Gibbs oscillations covers more than three sub-squares.
(a) Area conservation for different finite elements.

(b) $L^2$-norm for the real divergence.

Figure 2: Area conservation, and $L^2$-norm of the real divergence for the PW scheme with $\kappa = 1$, $\Delta t = 10^{-3}$, $\mu = 1$, $h_x = 1/1024$, $h_x = 1/16$. close to the immersed boundary. In the case of the enhancement with piecewise constants $P_3/(P_2+P_0)$, this area is reduced to two or three times $h_x$, and the error profile is sharper. The enhancement with discontinuous piecewise affine polynomials $P_3/(P_2 + P_1)$ provides oscillations greater than 0.1 confined in the same element where the structure is located.

In the second experiment we consider an elastic string whose initial configuration consists of an ellipse. In this case the elastic structure evolves into a circle. Due to the incompressibility constraint the area internal to the elastic boundary has to remain constant, see [4] for details. In Fig. 2 we compare the evolution of the area and of the $L^2$-norm of the divergence with respect to time. In Fig. 2(a) we see that the enhancement with piecewise constant discontinuous functions is already effective in improving the area conservation, but far more effective is the enhancement with discontinuous piecewise affine polynomials. Same comments hold true for the $L^2$-norm of the divergence presented in Fig. 2(b). These results suggest that there is a correspondence between the $L^2$-norm of the divergence and the area preservation of the method.

The next set of experiments is devoted to the comparison between PW and DLM schemes when the $P_3/(P_2 + P_1)$ element is used. In Fig. 3 we compare the mass conservation properties of the two schemes. We have used here the same parameters for the two methods and it is evident that the DLM scheme behaves better from this point of view. This fact is also confirmed in Fig. 4 where the position of the structure during a very coarse simulation is shown. The plot of the elastic membrane at subsequent simulation times, demonstrates that, even with a coarse mesh for the structure, the DLM approach can achieve good mass preservation properties, while the PW scheme needs at least one structure point per fluid element. When this last condition is not fulfilled, as in Fig. 4(a), the method fails, and the ellipse collapses instead of setting to a circle.

The next figures illustrate the behavior of PW and DLM schemes from the point of view of the stability. In Sect. 3 we have pointed out that the PW scheme is stable provided the CFL condition (14) is satisfied. Here we show that this condition is no longer required
for the DLM scheme. Fig. 5 reports the energy estimate as a function of time for fixed mesh parameters $h_x = 1/64$ and $h_y = 1/128$ and different values of $\Delta t$. We see that the energy of the PW scheme blows up when $\Delta t$ is greater than $10^{-2}$ according with (14) while in the DLM scheme it remains constant. Same remark holds also for Fig. 5 where we have fixed $\Delta t = 10^{-2}$ and $h_x = 1/64$ and keep varying $h_y$.

6 CONCLUSIONS

Higher order enhanced finite elements enjoy area and approximation properties that are up to ten times better in terms of area conservation, and three times better in terms of Gibbs phenomenon controlling (with respect to standard low order elements). Moreover, enhancing with $P_1$ pressures is much more effective than with $P_0$. This means that not only it is important to capture the discontinuity in the pressure, but it is also important the accuracy of the functions capturing the pressure discontinuity.

On the other hand, first results related to the DLM formulation are extremely encour-
aging both in terms of CFL stability and in terms of mass preservation. In fact these results suggest that the performances are worth the implementation and computational price of introducing the Lagrange multiplier for the kinematic condition. These preliminary results will undergo a more extensive and detailed exploration.

REFERENCES


DAMAGE EVALUATION OF RC MEMBERS UNDER HIGH TEMPERATURE BY USING RBSM-TRUSS NETWORK MODEL

T. UNNO*, H. NAKAMURA*, Y. YAMAMOTO†, M. KUNIEDA*, N. UEDA*

* Department of Civil Engineering, Nagoya University
  Furo-cho, Chikusa-ku, Nagoya, Japan
  e-mail: unno.takahiro@d.mbox.nagoya-u.ac.jp
  web page: http://concrete-lab.civil.nagoya-u.ac.jp/homepage/e/

†Department of Civil Environmental Engineering, National Defense Academy, Japan
  1-10-20, Hashirimizu, Yokosuka-shi, Kanagawa-Pref, Japan

Key words: RBSM, truss network model, explosion, rebar, thermal expansion

Abstract. The explosive behavior of RC members due to thermal stress and vapor pressure under high temperature was investigated numerically using a developed method. RBSM-TRUSS Network model was adopted as an analytical method, in which 3-dimensional Rigid Body Spring Method (RBSM) based on dynamic equation of motion as structural analysis and truss network model as mass transfer analysis were integrated. Moreover, the effect of rebar such as heat conduction and thermal expansion was modeled. Two RC members were simulated and it was shown that rebar with small cover influence damage accumulation and explosive behaviour.

1 INTRODUCTION

Behavior of concrete structure under high temperature are time and heat dependent combination problem of material, structure and mass transfer such as temperature, moisture and vapour. The fireproof performance related to explosion spalling has been evaluated by experimental method [1]. On the other hand, several analytical methods are proposed to evaluate the performance [2]. An advantage of analytical methods is to evaluate the effect of several factors on fireproof with low cost and short time. Therefore, it is desirable to enhance the performance of analytical methods in order to solve mechanical behavior from cracking to explosive behavior, considering realistic structural conditions such as arrangement of rebar, member shape and boundary condition.

A numerical method to simulate explosion spalling behaviour of concrete subjected to high temperature has been developed by authors [3]. The method is that 3-dimensional Rigid Body Spring Method (RBSM) based on dynamic equation of motion as structural analysis and truss network model as mass transfer analysis were integrated. The method can simulate structural behaviour due to thermal stress and vapour pressure considering the changes of distributions of temperature and vapour pressure obtained from truss network model. However, the applicability was limited to concrete only and it could not simulate RC members considering the effect of rebar.
In this study, in order to simulate RC members under high temperature, the rebar is modeled by beam element as structural member and is modeled by a part of truss network as heat conduction field. Then the effect on damage accumulation and explosive behaviour of the rebar is simulated.

2 ANALYTICAL METHOD

2.1 3-Dimensional RBSM

3-dimensional RBSM is used as structural analysis [4], which represents a continuum material as an assemblage of rigid particle elements interconnected by zero-size springs along their boundaries. Each rigid particle has three translation and three rotational degrees of freedom defined at nuclei. The interface between two particles consists of several springs as shown in Figure 1. That is, a boundary surface is divided by triangles with the center of gravity point and vertices of the surface, and individual one normal and two tangential springs are set at the integral point.

Since cracks initiate and propagate along interparticle boundaries, the crack pattern is strongly affected by mesh design. Therefore, random geometry using Voronoi diagrams is applied to partition of the specimen onto an assemblage of rigid particle.[5] Random geometry does not represent any structural feature within the concrete material, but rather is used to reduce mesh bias on potential crack directions.

The dynamic analysis is performed by solving the equation of motion expressed as Equation (1) in order to simulate the explosion behavior.

\[
[M] \ddot{y} + [K]y = f
\]

where \([M]\) is mass matrix, \([K]\) is stiffness matrix. \(\{\ddot{y}\}, \{y\}\) and \(\{f\}\) are acceleration vector, displacement vector and load vector, respectively.

![Rigid Body Spring Model](image1.png)

2.2 Truss network model

3-dimensional truss network model integrated with RBSM is applied to both heat conduction analysis and vapor pressure analysis. Truss elements are arranged between nuclei and the intermediate points of particle boundary, and by the truss element on the particle

![Truss network model](image2.png)
boundary as shown in Figure 2(a). [6] The truss which is connected between each particle is assumed to have the corresponding to the area of particle boundary surface which provide mass transport in bulk concrete. On the other hand, the area of the truss on particle boundary is decided by crack width which provide mass transport through crack. Then, a simplified 1-dimensional diffusion equation using truss element is employed to carry the potential flow such as the heat and the vapour pressure, in which the conductivity is changed dependent on the location such as in bulk concrete or crack.

2.3 Concrete material models [6]

The tensile behavior of concrete up to the tensile strength is modeled by using a linear elastic, and after cracking, according to 1/4 model to which a softening slope is changed, stress was reduced with the tensile strength, \( f_t \), the tensile fracture energy, \( G_{ft} \), and the distance between nuclei, \( h \) as shown in Figure 3. The behavior of concrete under compression is modeled by using a parabolic curve up to compressive strength \( f_c' \). The slope of a linear softening branch is defined by considering the compressive fracture energy, \( G_{fc} \).

Tangential springs represent the shear transfer mechanism of uncracked and cracked concrete. The shear strength is assumed to the Mohr-Coulomb type criterion with tension and compression caps. After shear stress reaches the yields strength, the stress moves on the yields surface. The shear transfer capacity at crack interfaces depends on crack opening. Thus, the shear stress is calculated by the function of crack width, shear stiffness and shear strain.

2.4 Modeling of rebar

Each reinforcement element is represented by a series of regular beam elements. The beam nodes are attached to the concrete particles without regard to the concrete mesh design through zero-size link element as shown in Figure 5 [7]. The stress strain relationship of reinforcement is used a bi-linear model. The bond stress-slip relationship is introduced into the spring parallel to the reinforcement of linked element.

In the conventional truss network model, the mass transfer between concrete elements and through crack was considered, and the heat conduction of rebar was not taken into account. In this paper, the heat conduction of the rebar is expressed by superposing truss network on the beam element of rebar. The effect of rebar on both structure and heat conduction can be considered by simple one dimensional element, in which information of temperature obtained
from truss network is delivered to beam element at common element nodes.

2.5 Effect thermal expansion of rebar

The thermal expansion strain is defined by the following equation in both longitudinal and radius direction of rebar.

\[ \Delta \varepsilon = \alpha \Delta T \]  

(2)

where \( \varepsilon \) is the thermal expansion strain, \( \alpha \) is thermal expansion coefficient of re-bar, and \( \Delta T \) is temperature increment. In this study, thermal expansive coefficient of rebar is set to \( 1.0 \times 10^5 \) as same as concrete.

In the axis direction, the thermal expansion strain is considered by internal strain of beam element. In the radius direction, the expansion strain work as compelling action to concrete. The action is considered as compelling displacement radiately. The compelling displacement is induced by the following equation,

\[ U = R \alpha \Delta T \]  

(3)

where U is compelling displacement, R is the radius of rebar.

The compelling displacement is translated to the displacement in the direction of normal springs at adjacent points as shown in Figure 6 and the internal strain is introduced in the springs, which is calculated by dividing the displacement depending on the thermal expansion by the normal spring length.

3 OUTLINE OF ANALYSIS

Figure 7 shows the analytical flow in the method. The method combined with the heat conduction analysis, the vapor pressure transfer analysis and the structural analysis. First, the heat conduction analysis is performed by truss network model, in which the temperature dependent parameters of specific heat \( c(T) \) and heat conductivity \( \lambda(T) \) [8] are defined by the temperature in previous of step. Then, the temperature distribution obtained from the heat conduction analysis is used to estimate the saturated vapor pressure \( p^* \) and the thermal expansive strain \( \Delta \varepsilon \).

In the vapor pressure transfer analysis, the production of vapor pressure is calculated by
considering the estimated saturated vapor pressure and relative humidity obtained in previous step. The distribution of vapor pressure is calculated by solving moisture transfer equation. The estimated vapor pressure is applied to structural analysis and the relative humidity is renewed for next step.

The structural analysis is performed by considering the thermal stress and the vapor pressure. The thermal stress of concrete is considered by the initial strain problem using the thermal expansion strain. The thermal stress of rebar is considered by the initial strain problem in longitudinal direction and compelling displacement in radius direction. The vapor pressure is considered by initial stress problem, which is applied to normal spring in RBSM. The dynamic analysis is performed by solving the equation of motion in order to simulate the explosion spalling behavior. The equation of motion is solved by Newmark’s $\beta$ method implicitly. The value of $\beta$ is 0.25. The structural analysis provides the cracking and the explosion spalling behavior. Moreover, the effect of cracks is considered in the vapor pressure transfer analysis, in which larger parameters are set.

![Analytical Flow Diagram](image)

Figure 7: Analytical Flow
4 ANALYSIS

4.1 Analytical Model 1

RC member which is arranged single rebar with 5mm cover thickness is modeled as shown in Figure 8. The specimen is divided by Voronoi particle with 10mm average element size and the bottom and the side surfaces are assumed fixed condition as a boundary condition. In order to evaluate the effect of the rebar on thermal conduction to the non-heating area, the heating side is set only to one third on the upper surface of the specimen. Heating is given according to the RABT curve as shown in Figure 9. Moreover, in order to check the influence of the rebar the result is compared with the case of no rebar. The coefficient of heat conduction of rebar and concrete, material properties of concrete and the parameter of vapor pressure transfer analysis are shown in Table 1, Table 2 and Table 3, respectively.

![Figure 8 Analytical model](image)

<table>
<thead>
<tr>
<th>Table1: Heat transfer coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Thermal expansion coefficient (1/C)</td>
</tr>
<tr>
<td>Heat transfer coefficient (W/m²·K)</td>
</tr>
</tbody>
</table>

Table2: Material properties

<table>
<thead>
<tr>
<th>Table2: Material properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressive strength (MPa)</td>
</tr>
<tr>
<td>Tensile strength (MPa)</td>
</tr>
<tr>
<td>Young modulus (GPa)</td>
</tr>
</tbody>
</table>

Table3: Parameters of vapor pressure transfer

<table>
<thead>
<tr>
<th>Table3: Parameters of vapor pressure transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moisture conductivity (g/m²·h·mmHg)</td>
</tr>
<tr>
<td>Moisture capacity (g/m²·mmHg)</td>
</tr>
<tr>
<td>Young modulus (g/m²·h·mmHg)</td>
</tr>
</tbody>
</table>

The temperature distribution in middle cross section Y-Y’, deformation behavior and cracking behavior at 40minutes from heating are shown in Figure 10. In the results of cracking behavior, green shows the stress of the normal spring reached tensile strength, yellow shows the stress which fall to 1/4 of tensile strength, and red shows stress became 0 in the stress strain relationship of Figure 3 (a). In the results of temperature distribution, it is understood the heat conduct to non-heating area through the rebar. The conducted heat through the rebar affects the deformation and cracking behavior. Damages of no rebar
 specimen with are smaller than that of specimen with rebar. Damage of no rebar specimen concentrates on heating area. On the other hand, damages of specimen with a rebar propagate to non-heating area and cracks on the rebar grow to non-heating area along the rebar. These results suggest that thermal conduction of rebars wide on the damage area.

**4.2 Analytical Model 2**

The effect of the cover thickness of rebar is evaluated using RC slab. The analytical model and the boundary condition are shown in Figure 11. The analysis is performed by 1/4 model (200×50×200 mm) considering the symmetry of the heating condition. The specimen is divided by Voronoi particles, and the mean element size is about 3mm near the heating area. Outer surfaces are fixed as the structural boundary conditions. In order to evaluate the influence of the depth from the surface of rebar, Depth of the rebar is valued at 5mm and 10mm. Moreover, no rebar specimen is analyzed to compare the results of specimen with rebars. Six rebars are arranged as shown in Figure 11. Heating was given according to the RABT curve as shown in Figure 9. The coefficient of heat conduction of rebar and concrete, material properties of concrete and the parameter of vapor pressure transfer analysis are shown in Table 4, Table 5 and Table 6, respectively.

<table>
<thead>
<tr>
<th></th>
<th>concrete</th>
<th>rebar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal expansion coefficient (1/℃)</td>
<td>$1.0 \times 10^5$</td>
<td>$1.0 \times 10^5$</td>
</tr>
<tr>
<td>Heat transfer coefficient (W/m²·K)</td>
<td>22.4</td>
<td>89.2</td>
</tr>
</tbody>
</table>

**Table 4: Heat transfer coefficient**
Figure 12 shows the deformation behavior at 10 minutes and final stage that explosion spalling behavior occur. At 10 minutes, it is observed that surface crack occurrence and the damage accumulate earlier when rebar is arranged with small cover thickness. The explosion spalling occur in all cases. The behaviors are almost same, but the occurrence time is different. Figure 13 shows temperature distribution at 10 minutes on surface and in cross section. In all cases difference is much small. Figure 14 shows cracking behavior at 10 minutes in the area of 5mm depth from surface, 10mm depth from 5mm depth and 15mm depth from 10mm depth. In the no rebar specimen, a few cracks occur in surface area. In the case of 5mm cover thickness, many cracks occur to 10mm depth. The rebar affect crack behavior obviously. In the case of 10mm cover thickness, although the cracks in surface area are not so many, many cracks occur about 10mm depth near the rebars. This is effect of thermal expansion of rebar. Moreover, the cracks under 10mm depth are larger than the case of 5mm depth due to the effect of thermal expansion of rebar.

5 CONCLUSIONS

(1)The analytical method to simulate the effect of rebar under high temperature was developed. In the method, rebar is modeled by beam element as structural element and by truss element as heat conduction field.

(2) The compelling displacement due to thermal expansion of rebar was modelled. It was shown that the compelling displacement affect on the damage accumulation and spalling.
behave.

(3) The spalling behavior of RC members was simulated. The spalling occur earlier for existing of rebars and smaller thickness.

**ACKNOWLEDGEMENT**

This study was supported by JSPS KAKENHI Grant Number 24360171

**REFERENCES**


![Temperature distribution](image1)

![Cracking behavior](image2)


MOISTURE TRANSPORT IN CONCRETE DURING FIRE: AN NMR STUDY

L. PEL*, G.H.A. VAN DER HEIJDEN* AND S.J. JASPERS*

* Department of Applied Physics, Transport in Permeable Media, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: l.pel@tue.nl, www.phys.tue.nl/nfcmr

Key words: heat transfer, porous media, multiphase flow, phase change, spalling, NMR, concrete.

Abstract. The question how concrete responds to a fire is one of the main questions in fire safety. During a fire, a building material can suddenly be heated up to temperatures well above 1000 °C. At temperatures above 100 °C, water inside the pores will start to boil. Simultaneously, in concrete, but also for example in gypsum, chemically bound water will be released by dehydration of the porous matrix. If the concrete has a low permeability, the vapour pressure inside will increase which can give rise to a sudden (explosive) failure of a material. Numerous heat and mass transfer models have been used to predict the moisture transport and its consequences on the strength and permeability of the concrete. However, these models are only of use if they can be validated. For model validation, quantitative measurements of the evolution of moisture, temperature, and possibly pressure distributions in time are needed. For this purpose, we have developed an NMR setup to measure the moisture transport in heated building materials. This setup makes used of the 1.5 T magnet of a medical MRI scanner, while dedicated gradient coils are used to image the moisture profiles in 1D. The sample is heated with the help of four 100 W halogen lamps, capable of generating a heat flux of 12 kW/m². The measured combined moisture content and temperature profiles give a unique insight in the moisture transport and dehydration kinetics inside concrete during fire. These measurements give the first quantitative proof for the build-up of a moisture peak due to the vapour pressure build-up. While many moisture transport models predicted the existence of such a peak, these predictions had never been validated with moisture profile measurements during fire tests.

1 INTRODUCTION

The moisture content and transport in concrete, together with the vapour permeability, is presumed one of the most influential factors for fire spalling. It is generally assumed that below a certain moisture content of approximately 5 vol.% no spalling will occur [1, 2]. Numerous heat and mass transfer models have been used to predict the moisture transport and its consequences on the strength and permeability of the concrete [3-8]. However, these models are only of use if they can be validated. For model validation, quantitative
measurements of the evolution of moisture, temperature, and possibly pressure distributions in time are needed.

A few (indirect) observations of the moisture content in concrete have been reported in literature. Water has been observed bleeding from the cool side of a concrete wall, indicating that the surface layer is saturated [9]. Jansson et al. report a method in which a concrete cube is heated from one side. By splitting the heated cube in half, a layer with a relative higher saturation is observed by a discolouration of the concrete [10]. In this way, the sample is destroyed by the act of measurement itself. It will be extremely labourious to obtain a time evolution of the moisture content. Each time increment would also be from a different concrete sample. Experimentally determined moisture contents with a spatial resolution of about 20 cm were presented in paper by Ichikawa et al. [11]. However, the obtained resolution is far too low to validate the model results.

With a dedicated NMR fire spalling setup it is possible to measure both the moisture and temperature distributions non-destructively. This setup has been extensively tested in previous experiments on benchmark materials such as fired-clay brick, calcium silicate brick, and gypsum [16,18]. For the measurements on concrete it is necessary to have a validated measurement technique since the moisture content of concrete in equilibrium with, e.g., 50% RH, is typically in the range of \(10^{-3} \text{ m}^3\text{m}^{-3}\). Although the low moisture content is challenging the capabilities of the NMR setup, it was possible to measure the moisture transport in concrete. We will first discuss the NMR, followed by a discussion of the measurements.

2 NMR SETUP

2.1 NMR basics

Almost all nuclei have a magnetic dipole moment, resulting from their spin-angular momentum. (One can think of a nucleus as a charged sphere spinning around its axis, which corresponds to a current loop, generating a magnetic moment). In a Nuclear Magnetic Resonance (NMR) experiment the magnetic moments of the nuclei are manipulated by suitably chosen electromagnetic radio frequency (RF) fields. The frequency of the resonance conditions is given by:

\[
f_l = \frac{\gamma}{2\pi} B_0
\]

where \(f_l\) is so-called Larmor frequency [Hz], \(B_0\) [T] the externally applied static magnetic field and \(\gamma\) is the gyromagnetic ratio which is dependent on the type of nucleus (for \(1\text{H} \gamma/2\pi = 42.58 \text{ MHz T}^{-1}\)). Because of this condition the method can be made sensitive to only hydrogen and therefore to water, in contrast to the attenuation methods. When a known magnetic field gradient is applied, the constant magnetic field \(B_0\) in the resonance condition (Eq 1) has to be replaced by the spatially varying magnetic field \(B\):

\[
B(x) = B_0 + Gx
\]

where \(G\) [T m\(^{-1}\)] (see Fig. 1) is the magnetic field gradient and \(x\) is the position of the precessing magnetic moment. The resonance condition is then spatially dependent. Therefore
the moisture content at different positions $x$ in the sample can be measured by varying the resonance frequency $f$ without moving the sample. In a pulsed NMR experiment the orientation of the moments of the spins are manipulated by short RF pulses at the resonance frequency. The amplitude of the resulting signal emitted by the nuclear spins, the so-called ‘Hahn spin-echo’ signal is proportional to the number of nuclei taking part in the experiment. The spin-echo signal also gives information about the rate at which this magnetic excitation of the spins decays. The system will return to its magnetic equilibrium by two mechanisms: interactions between the nuclei themselves, causing the so-called spin-spin relaxation, and interactions between the nuclei and their environment, causing the so-called spin-lattice relaxation. Assuming that both mechanisms give rise to a single exponential relaxation and that spin lattice relaxation is much slower then the spin-spin relaxation, the magnitude of the NMR spin-echo signal is given by:

$$S = \rho \exp \left(-\frac{t_E}{T_2}\right) \left[1 - \exp \left(-\frac{t_R}{T_1}\right)\right]$$  

(3)

In this equation $\rho$ is the proton density, $T_1$ the spin-lattice relaxation time, $t_R$ the repetition time of the spin-echo experiment, $T_2$ the spin-spin or transverse relaxation time and $t_E$ the so-called spin-echo time. Obviously small $T_2$ values lead to a decrease of the spin-echo signal, whereas, on the other hand, small $T_1$ values are preferred, as this parameter limits the repetition time (usually $t_R \approx 4T_1$) and hence the rate at which the moisture profiles can be scanned. Water in a porous building material has typically a relaxation time $T_1$ in the order of $0.1 - 0.5$ s.

**Figure 1**: Schematic diagram of the NMR setup. The setup has a cylindrical symmetry. A whole-body 1.5 T MRI scanner provides the main magnetic field. An anti-Helmholtz coil configuration provides a constant magnetic field gradient $G$. A bird-cage RF coil with a diameter of 140mm is used for both sending the RF pulses and receiving the NMR signal. An array of four 100W halogen lamps is used to heat the sample. The sample has a diameter of 80 mm and a length of 100 mm. It is thermally insulated and positioned inside the bird-cage coil.
In porous materials $T_2$ will be decreased strongly with respect to that of 'pure' water, due to surface relaxation at the pore walls. In many common porous building materials, like fired-clay brick or concrete, usually large amounts of paramagnetic ions (e.g., Fe) are present. This complicates the NMR measurements by two effects. First, due to the large susceptibility of the porous material, the transverse relaxation time $T_2$ will be drastically decreased. Secondly, the variations in local magnetic susceptibility will broaden the resonance line and thereby limit the spatial resolution. The relaxation can give information about the distribution of pore sizes in a porous material [12].

For the experiments described in this study a specially designed NMR scanner was used. This instrument was especially designed for quantitative measurements of moisture in porous materials with short transverse relaxation ($T_2$) times (unlike standard Magnetic Resonance Imaging (MRI), which is generally used in a qualitative way). The machine makes use of the magnet of a whole body MRI scanner (Gyroscan, Philips) which operates at a main field of 1.5 Tesla corresponding to a frequency of 63.9 MHz. The setup is placed within the scanner and a schematic diagram is given in Fig 1. Two Helmholtz coils provide the magnetic field gradient $G$ in the direction of $B_0$. The magnitude of the gradient is 86.5 mT m$^{-1}$, providing a spatial resolution of the order of 4 mm.

A home built birdcage coil with a diameter of 140 mm is used for applying the RF pulses, and receiving the NMR signal from the sample. A birdcage coil is used because it generates a homogeneous $B_1$-field perpendicular to the sample in order to generate spin-echoes. Therefore, the coil can be placed parallel to the main magnetic field providing optimal use of the available space inside the bore. The sample is heated with 4 halogen lamps. The reflectors of these lamps are gold plated to ensure maximum reflection of the infra-red radiation. The sample is placed inside the birdcage coil and is thermally insulated using mineral wool in order to create a 1D experiment.

2.2 Signal calibration

In order to obtain a quantitative moisture content, the NMR signal must be calibrated against the moisture content of the concrete. For this purpose, the NMR signal was measured with the same settings as during the experiments. A sample was initially vacuum saturated, after which it was dried slowly to keep the moisture homogeneously distributed throughout the sample. Mass and signal were both recorded throughout the drying process. The NMR signal is shown as a function of the normalised sample mass in Fig. 2. For mass ratios between 1 and 0.94 a linear relation between the NMR signal and the moisture content is obtained. This indicates that the NMR signal can be directly related to the amount of free moisture present in the concrete. After the sample was dried at ambient temperatures a further decrease in moisture content (chemically bound) was obtained by heating the sample up to a certain constant temperature above 100 °C.

After the sample had reached a constant mass, the sample was slowly cooled down to room temperature under a 0% RH atmosphere. Thereby preventing any increase in moisture content or rehydration of the concrete. Then, at room temperature the NMR signal and mass were both measured. The results are shown in the inset of Fig. 2. A decrease in both signal and mass is observed for temperatures from 105 °C to 130 °C. For the drying temperatures of 130
°C to 250 °C the mass of the sample decreases indicating that chemically bound moisture is lost from the concrete. However, the NMR signal remains constant. For temperatures above 300 °C again both the signal as well as the moisture content decrease, until all signal is lost at a temperature between 500 and 800 °C. From this drying experiment we can conclude that the NMR signal is linearly related to the free moisture content. Furthermore, the NMR setup is capable of measuring a signal from the chemically bound moisture content. Unlike in, e.g., gypsum there is no linear relation between the chemically bound moisture content and the NMR signal. In fact, between 120 and 170 °C no decrease in NMR signal is observed, although the mass of the sample decreases. At a temperature of 800 °C the concrete sample is completely dehydrated, resulting in zero NMR signal. In the future we will try to include the chemically bound moisture, but for now we will solely concentrate on the free moisture content.

![Figure 2](image)

**Figure 2**: NMR signal as a function of normalised sample mass during a static isothermal calibration. The signal varies linearly with the amount of free moisture (solid line, m/m₀ > 0.94). The calibration of the non-evaporable moisture content (hydrated water) is shown in the inset. The temperatures indicate the temperature up to which the sample was heated.

### 3 EXPERIMENTAL RESULTS

#### 3.1 Sample preparation

The cylindrical concrete sample used in this study were drilled from larger cast blocks of 400x100x100 mm and have a diameter of 80 mm, and a length of 100 mm. The concrete has a strength class of C40, with a water cement ratio of 0.5. The details of the mix are given in Table 1. As can be seen, the largest aggregate size has been reduced to 8 mm to ensure a more representative sample volume. The concrete blocks were stored under water for three months before the samples were drilled. After drilling, the sample was first dried, after which it was equilibrated at 97 % RH until a constant mass was reached. The sample was one year old at
the time of the experiment.
Before the experiment, the samples were pressed in a PTFE holder to seal all sides except for the heated surface (transversal plane of the cylinder). In this way the moisture transport is limited to one dimension. The flow of heat is limited to one dimension by insulating the sample using mineral wool (4 cm). Numerical simulations indicate that the radial heat flow is limited to 10% of the longitudinal heat flow.

### Table 1: Concrete (C40) mix design

<table>
<thead>
<tr>
<th>Constituent</th>
<th>Amount [kgm⁻³]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEM I 32.5 R</td>
<td>350</td>
</tr>
<tr>
<td>water</td>
<td>175</td>
</tr>
<tr>
<td>W/C ratio</td>
<td>0.5</td>
</tr>
<tr>
<td>sand (0.125-0.250)</td>
<td>127</td>
</tr>
<tr>
<td>sand (0.250-0.500)</td>
<td>217</td>
</tr>
<tr>
<td>sand (0.500-1)</td>
<td>217</td>
</tr>
<tr>
<td>sand (1-2)</td>
<td>253</td>
</tr>
<tr>
<td>sand (2-4)</td>
<td>380</td>
</tr>
<tr>
<td>ravel (4-8mm)</td>
<td>614</td>
</tr>
<tr>
<td>Total</td>
<td>2330</td>
</tr>
</tbody>
</table>

3.2 Measured profiles

At the start of the experiment the concrete sample was first equilibrated at 97% RH, corresponding to a free moisture content of 0.07 m³m⁻³. The moisture and temperature profiles are shown in Fig. 3. The moisture profiles are normalised with respect to the first moisture profiles (bold line). As the surface (0 mm) is heated the temperature increases after 8.5 minutes to above 100 °C. The free moisture at the surface will boil and we can observe a boiling front developing at the surface. As the heating continues, the boiling front moves further into the material. To the left of the boiling front a zero free moisture content is measured. To provide a complete overview, we have chosen to include the signal originating from the chemically bound moisture. Therefore, the zero free moisture content level is indicated by θ=0. However, it is possible for the signal to decrease below θ=0, corresponding to loss of chemically bound moisture. The measured temperature profiles indicate that the boiling and dehydration processes have no significant influence on the heat transport. In previous heating experiments on fired-clay brick and calcium silicate brick a clear inflection in the temperature profiles was observed either caused by boiling and/or dehydration [16]. In this experiment, the moisture content of the concrete is relatively low and hence both the heat capacity, and thermal conductivity of the sample are dominated by the concrete heat capacity and conductivity. As the boiling front moves further into the sample we can observe a significant increase in moisture content to the right of the boiling front. However, as the boiling front progresses the moisture content does not increase above approximately 0.11 m³m⁻³. This moisture content corresponds to the porosity of the concrete sample (indicated by the dashed line). The moisture content in this region cannot increase any further, because the material is saturated. To our knowledge this is the first direct and quantitative proof of so-called moisture clogging in concrete.
An explanation for the increase in moisture content can be found in the local vapour pressure at the boiling front. In Fig. 3b the temperature at the boiling front is indicated. The temperature at the front increases from approximately 160 °C to 195 °C. The saturated vapour
pressures corresponding to these temperatures are 0.7 and 1.4 MPa respectively, compared to 0.1 MPa atmospheric pressure. The increase above 100 °C (atmospheric boiling point) indicates that there is a local increased vapour pressure. The vapour pressure is assumed to be equal to the saturated vapour pressure. As a result of these high vapour pressures the vapour released at the boiling front is advected both towards the heated surface as well as the back of the sample. A temperature gradient is present across the sample. As a result the saturated vapour pressure immediately to the right of the boiling front is lower. The vapour advected to the back of the sample, will therefore condensate, resulting in an increase of the free moisture content in the region to the right of the boiling front. The vapour advected towards the surface is able to exit the sample.

**Figure 4**: The first preliminary results of combined measurements of the moisture profile as measured by NMR and the temperature and pressure as measured by an embedded pressure Gauge. Indicated are the temperatures at the position of the pressure gauge (40 mm) as measured by the thermocouple in the pressure gauge and the moisture content at the position of the gauge. Note that the pressure scale is chosen so that 0.1 MPa corresponds to the saturated vapor pressure at 100 °C (or: 0.1 MPa is atmospheric pressure).

### 3.3 Pressure measurement

To provide additional information we have also performed measurements in which we for the first time combined the NMR measurements of the moisture profile, temperature profiles with pressure measurements using a gauge embedded in the concrete. Here we have used an embedded CSTB pressure gauge as first proposed by Kalifa et al. This gauge consists of a sintered metal plate in a cup which and was positioned at the main axis of the cylinder at a distance of 40 mm from the heated surface. In figure 4 some preliminary results are shown, i.e., the temperature and moisture content at the position of the Gauge and also the pressure as measured by the Gauge. As can be seen these preliminary measurements indicate the indeed the maximum vapour pressure is reached near the boiling front, whereas the maximum is close to the saturated vapour pressure.
4 CONCLUSIONS

This study shows that NMR is a very powerful tool in providing accurate moisture and temperature profiles. It is possible to measure non-destructively the moisture distributions with NMR inside building materials during heating. The moisture profiles can be measured with a resolution of 4-5 mm, while the sample is heated with a radiative heat flux of approximately 12 kW m\(^{-2}\). The measurements give the first quantitative proof for the buildup of a moisture peak. While all moisture transport models predict the existence of such a peak, these predictions were never validated with moisture profile measurements. Moreover, the development of a saturated region has been shown in the experiment.

5 ACKNOWLEDGEMENTS

Part of this research was supported by the Dutch Technology Foundation (STW), and TNO Built Environment and Geosciences. For their help in performing the pressure measurements we would also like to thank F. Pereira, P. Pimienta and H. Carré of the CSTB Centre Scientifique et Technique du Bâtiment in France and SIAME, University of Pau and Pays de l'Adour in France.

REFERENCES


A MULTI-SCALE REFINED ZIGZAG THEORY FOR MULTILAYERED COMPOSITE AND SANDWICH PLATES WITH IMPROVED TRANSVERSE SHEAR STRESSES

LUIGI IURLARO*, MARCO GHERLONE*, MARCO DI SCIUVA* AND ALEXANDER TESSLER†

* Department of Mechanical and Aerospace Engineering, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: {luigi.iurlaro, marco.gherlone, marco.disciuva}@polito.it, web: http://www.aesdo.polito.it

† Structural Mechanics and Concepts Branch, NASA Langley Research Center
Mail Stop 190, Hampton, Virginia, 23681-2199, U.S.A.
email: alexander.tessler-1@nasa.gov - web: http://www.nasa.gov

Key words: Multi-scale plate theory, Reissner’s mixed variational theorem, refined zigzag theory, multilayered composite plate, sandwich plate, transverse shear stresses.

Abstract. The Refined Zigzag Theory (RZT) enables accurate predictions of the in-plane displacements, strains, and stresses. The transverse shear stresses obtained from constitutive equations are layer-wise constant. Although these transverse shear stresses are generally accurate in the average, layer-wise sense, they are nevertheless discontinuous at layer interfaces, and thus they violate the requisite interlaminar continuity of transverse stresses. Recently, Tessler applied Reissner’s mixed variational theorem and RZT kinematic assumptions to derive an accurate and efficient shear-deformation theory for homogeneous, laminated composite, and sandwich beams, called RZT(m), where “m” stands for “mixed”. Herein, the RZT(m) for beams is extended to plate analysis, where two alternative assumptions for the transverse shear stresses field are examined: the first follows Tessler’s formulation, whereas the second is based on Murakami’s polynomial approach. Results for elasto-static simply supported and cantilever plates demonstrate that Tessler’s formulation results in a powerful and efficient structural theory that is well-suited for the analysis of multilayered composite and sandwich panels.

1 INTRODUCTION

Over the past two decades, composite materials have been increasingly used in civil, automotive, and aerospace applications. Numerous structural theories have been explored for the analysis of multilayered composite and sandwich structures. Since transverse shear deformations govern damage mechanisms that contribute to delamination initiation and propagation, efforts to develop accurate predictions of transverse shear strains and stresses have been extensive [1].
In the framework of displacement-based lamination theories, a distinction is usually made between Equivalent Single Layer (ESL) and Layer-Wise (LW) theories [2]. The former assume a coarse approximation for the displacement components, thus reducing a multilayered plate to a single-layer plate whose behavior is governed by some average constitutive properties. In contrast to ESL theories, LW theories employ kinematic assumptions for every layer, hence the number of kinematic variables increases with the number of layers. Although the LW theories are sufficiently accurate, they become computationally intensive for application to laminates that have a large number of layers.

Generally, the transverse shear stresses obtained from the constitutive equations, using either ESL and ZZ theories, suffer from the lack of interlaminar continuity and often exhibit rather inaccurate magnitudes across the laminate thickness. A post-processing analysis, which usually results in improved transverse shear stresses, is integration of equilibrium equations of elasticity theory. Since these computations involve second-order partial derivatives of kinematic variables, their application in finite element analysis is often associated with significant errors. Alternatively, mixed-field variational principles have been used [3-5], wherein the displacement and transverse shear stress assumptions are made independently.

A viable compromise between adequate accuracy and computational efficiency is offered by the so-called ZigZag (ZZ) theories, wherein the assumed kinematic field is a superposition of the coarse and fine distributions through the thickness. The fine distribution, also known as zigzag, is commonly piecewise linear, and it enables the inplane-displacement partial derivative taken with respect to the thickness coordinate to be discontinuous at the layer interfaces. The ZZ theories are computationally efficient because they have a fixed number of variables regardless of the number of layers. These theories also achieve global response predictions comparable to those of the LW theories [6,7]. Recently, Tessler et al. [8,9] developed the so-called Refined Zigzag Theory (RZT), wherein the displacement field of First-order Shear Deformation Theory (FSDT) is enriched by the addition of suitable zigzag functions that have $C^0$-continuous distributions through the thickness. Each of the in-plane zigzag displacements is controlled by a single amplitude variable. Thus, the FSDT five-variables plate kinematics is increased to seven variables. Efficient, $C^0$-continuous, RZT-based finite elements have also been developed for beams, plates, and shells [10-12].

In this paper, a mixed-field formulation is undertaken in the framework of RZT plate kinematic assumptions and Reissner’s mixed variational theorem [13]. Two alternative strategies for the approximation of transverse shear stresses are examined. The first strategy adopts Tessler’s methodology in [14] and is denoted as $RZT_1^{(m)}$. The transverse shear stresses are derived using three-dimensional equilibrium equations of elasticity theory. The second strategy, $RZT_2^{(m)}$, is based on Murakami’s [3] polynomial approach. The two approaches are assessed by way of elasto-static problems for simply supported and cantilever plates subjected to bi-sinusoidal and uniformly distributed loading. Analytic solutions are obtained and compared with three-dimensional exact elasticity solutions, high-fidelity finite element solutions, and RZT (displacement theory) solutions. The comparisons indicate that $RZT_1^{(m)}$ is consistently more accurate than $RZT_2^{(m)}$, hence it can be used effectively in the analysis of multilayered composite and sandwich plates.
2 RZT\textsuperscript{(m)} ASSUMPTIONS

Consider a laminated plate of uniform thickness $2h$ made by $N$ perfectly bonded orthotropic layers. The plate is referred to the orthogonal Cartesian coordinate system $(x, y, z)$, where the plate midplane $\Omega$ is placed on the $x$-plane, whereas the through-the-thickness coordinate $z$ ranges from $-h$ to $h$. Throughout this paper, Greek subscripts take the values 1, 2.

Reissner’s mixed variational theorem [13] permits independent assumptions to be used for displacements and transverse shear stresses. This theorem enforces, as a constraint, the compatibility between the transverse shear strains resulting from the strain-displacement relations and those that are assumed a priori, i.e.,

$$\int_{\Omega} \left[ \delta \varepsilon^T \gamma - \delta \varepsilon^T (\gamma - \gamma_p) \right] dz d\Omega = \delta W_e$$

where $\delta$ is the variational operator and $W_e$ represents the work of the external loads. The strain vector $\varepsilon^T \equiv \{\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12}, \gamma_{1z}, \gamma_{2z}\}$ contains the in-plane and transverse-shear, $\gamma^T \equiv \{\gamma_{1z}, \gamma_{2z}\}$, strains obtained by means of the linear strain-displacement relations. The stress vector $\sigma^T \equiv \{\sigma_{11}, \sigma_{22}, \tau_{12}, \tau_{a1z}, \tau_{a2z}\}$ contains the in-plane stresses obtained by means of Hooke’s law and the transverse-shear stresses, $\tau^T \equiv \{\tau_{a1z}, \tau_{a2z}\}$, that are assumed independently. The vector $\gamma^T_a \equiv \{\gamma_{a1z}, \gamma_{a2z}\}$ is obtained from the assumed stresses, $\tau_a$, using Hooke’s law.

In what follows, the assumed displacement and transverse-shear stress fields are described in detail.

2.1 Displacement field

The kinematic assumptions of RZT are adopted, having demonstrated superior predictive capabilities to model both the global and local response quantities [8-12,15]. The displacement components are defined in the Cartesian coordinate system as [9]

$$\begin{cases}
  u^T_k(x, z) = u_a(x) + z\theta(x) + \phi^{(k)}_a(z)\psi_a(x) \\
  u_t(x, z) = w(x)
\end{cases}$$

In Eq. (2), and in the following, the superscript $(k)$ indicates quantities corresponding to the $k$th lamina, $u_a$ is a uniform displacement component along the $x_a$-axis, $\theta_a$ is the average bending rotation of the transverse normal about $x_\beta$-axis ($\alpha \neq \beta$), and $w$ is the transverse deflection. The kinematic variable $\psi_a$ represents the amplitude of the zigzag rotation, whereas $\phi^{(k)}_a$, derived in [9], is a piecewise linear zigzag function that is independent of the state of deformation. The RZT kinematic variables include those of FSDT, $u_a$, $w$, and $\theta_a$, as well as the two additional variables called the zigzag rotations, $\psi_a$. 
RZT is characterized by a multi-scale kinematic description since the displacement field is given by the superposition of the coarse and fine contributions: the former, modeling the kinematics with the resolution on the scale of the entire plate thickness, corresponds to the FSDT, whereas the latter models the mechanical behavior with the resolution on the scale of the $k$th material layer.

The strains are obtained using the linear strain-displacement relations

$$2\varepsilon_{ij}^{(k)} = u_{\alpha,\beta} + u_{\beta,\alpha} + z(\theta_{\alpha,\beta} + \theta_{\beta,\alpha}) + \phi_{\alpha}^{(k)}\psi_{\alpha,\beta} + \phi_{\beta}^{(k)}\psi_{\beta,\alpha}; \quad \gamma_{\alpha\beta}^{(k)} = \gamma_{\alpha} + \beta_{\alpha}^{(k)}\psi_{\alpha}$$

with $\gamma_{\alpha} \equiv w_{\alpha} + \theta_{\alpha}$ and $\beta_{\alpha}^{(k)} \equiv \phi_{\alpha}^{(k)}$. Hooke’s constitutive relations are then invoked to compute the stresses

$$\sigma_{ij}^{(k)} = C_{ijkl}^{(k)}\varepsilon_{kl}^{(k)}; \quad \tau_{\alpha\beta}^{(k)} = Q_{\alpha\beta}^{(k)}\psi$$

where $C_{ijkl}^{(k)}$ and $Q_{\alpha\beta}^{(k)}$ are the transformed elastic stiffness coefficients referred to the $(x, z)$ coordinate system and relative to the plane-stress condition that assumes that transverse normal stress is negligibly small in relation to the in-plane stresses.

### 2.2 Transverse shear stresses

In this study, two thickness-wise distributions of the transverse shear stresses that are continuous along the layer interfaces and satisfy traction conditions at the top and bottom plate surfaces are considered: (i) a distribution derived by way of three-dimensional elasticity equilibrium equations, and (ii) an assumed polynomial distribution. In both cases, the transverse shear stresses can be expressed as

$$\tau_{\alpha} = Z_{\alpha}(z)f_{\alpha}(x) + Z_{\alpha}(z)n_{\alpha}(x)$$

where the matrices $Z_{\alpha}$, $Z_{n}$ are dependent on the thickness coordinate, and the vector $f_{\alpha}$ is a stress function of the in-plane coordinates; $n_{\alpha} = \{\overline{P}_{1}^{(b)}, \overline{P}_{1}^{(i)}, \overline{P}_{2}^{(i)}, \overline{P}_{3}^{(i)}\}$ is a vector containing prescribed surface tractions that act along the $x_{\alpha}$-direction; $\overline{P}_{\alpha}^{(i)}$ are prescribed on the top surface, $S_{t}$, and $\overline{P}_{\alpha}^{(b)}$ on the bottom surface, $S_{b}$.

#### 2.2.1 Integrated transverse shear stresses (version $RZ{T_{1}}^{(m)}$)

Three-dimensional equilibrium equations of elasticity are commonly used in an attempt to derive improved, layer interface-continuous transverse shear stresses. Auricchio and Sacco [4] used an equilibrium-integration approach to derive transverse shear stresses for the FSDT-based plate analysis. It was recognized that due to a large number of independent stress parameters, the mixed-field variational formulation tends to fit the constitutive-based stresses very closely, yielding only insignificant improvements for either the equilibrium- or polynomial-based stresses. Moreover, an ad hoc function had to be added to the integrated transverse shear stresses in order to satisfy the traction-free boundary conditions on the top bounding surface, whereas the bottom zero-traction condition was enforced a priori.
Recently, Tessler [14] presented a mixed-field formulation for RZT beams, which derives the transverse shear stress from the two-dimensional elasticity equilibrium equations. A key step in the formulation is that the transverse shear stress is made to satisfy exactly the first (axial) equilibrium equation, and hence it satisfies a priori the top and bottom traction conditions of arbitrary distributions, including the special cases of zero-traction conditions. The derived stress is also fully continuous along layer interfaces. The problem is reduced to replacing two second-order derivatives of the kinematic variables with two unknown stress functions that are determined using Reissner’s mixed-field theorem.

Herein, Tessler’s [14] methodology is used to derive an RZT mixed-field formulation for plate analysis. By neglecting the body forces, the first two equilibrium equations of elasticity theory may be written as

$$\tau_{\alpha z} = -\left(\sigma_{\alpha v} + \tau_{\alpha z}\right), \quad \beta \neq \alpha$$

Integrating with respect to the \(z\)-coordinate and enforcing the traction conditions at the bottom plate surface \((z=-h)\) yields

$$\tau_{\alpha z} = -\tau^{(0)}_{\alpha} - \int_{-h}^{h} \left(\sigma_{\alpha v} + \tau_{\alpha z}\right) dz, \quad \beta \neq \alpha$$

Introducing Eq. (4) into Eq. (7), and after some straightforward manipulations, the transverse shear stresses involving eighteen second-order partial derivatives of \(u_\alpha, \theta_\alpha\) and \(\psi_\alpha\) are obtained. To circumvent the over fitting deficiency encountered in [4], a simple strategy is pursued herein that leads to the functional simplicity associated with the mixed-field RZT beam formulation [14]. To simplify Eq. (7), cylindrical bending is considered for each of the \(x_\alpha\) directions, resulting only in the second partial derivatives of \(u_\alpha, \theta_\alpha\) and \(\psi_\alpha\) (with respect to \(x_\alpha\) in the \((x_\alpha, z)\)-plane) in the expressions for the transverse shear stresses. Thus, the integrated shear stresses become

$$\tau_{\alpha i} = -\tau^{(0)}_{\alpha} - \int_{-h}^{h} \left(\sigma_{\alpha v} + \tau_{\alpha z}\right) dz, \quad \beta \neq \alpha$$

To fulfill the full shear traction equilibrium on the top plate face, the first two RZT equilibrium equations, describing the in-plane equilibrium, (see [9]), are written for cylindrical bending case as

$$\int_{-h}^{h} C^{(i)}_{\alpha} dz u_{\alpha, aa} + \int_{-h}^{h} z C^{(i)}_{\alpha} dz \theta_{\alpha, aa} + \int_{-h}^{h} \phi^{(i)}_{\alpha} C^{(i)}_{\alpha} dz \psi_{\alpha, aa} + \bar{p}_{\alpha} = 0$$

where \(\bar{p}_{\alpha} = \tau^{(0)}_{\alpha} + \tau^{(b)}_{\alpha}\). Solving Eq. (9) for \(u_{\alpha, aa}\) gives
Substituting Eq. (10) into Eq. (8) results in the transverse shear stresses that have only two second-order partial derivatives, $\theta_{a,ax}$ and $\psi_{a,ax}$, per stress component, and the thickness-distribution functions that satisfy all traction equilibrium conditions exactly, including the layer-interface equilibrium conditions. The final form of the transverse shear stresses is given by Eq. (5) in which the surface tractions are grouped in $n$, and the second-order partial derivatives are replaced by independent functions of $x$ and are grouped in $f$. The corresponding version of the theory is herein denoted by $RZT^{(m)}$. 

**2.2.2 Polynomial approximation of shear stresses (version $RZT^{(m)}_{2}$)**

The use of polynomial approximation for the transverse shear stresses in a mixed-field formulation appears for the first time in [3], and since has been adopted by many investigators. For each material layer, a polynomial thickness distribution is assumed and is expressed as

$$
\tau^{(k)}_{ax} = \tau^{(k)}_{ax} F_{b}^{(k)}(z) + \tau^{(k)}_{ax} F_{t}^{(k)}(z) + T_{a}^{(k)} F_{m}^{(k)}(z)
$$

(11)

where $\tau^{(k)}_{ax}$ and $\tau^{(k)}_{ax}$ are the values of transverse shear stress $\tau^{(k)}_{ax}$ at the bottom and top interface of the $k$th layer, respectively; $T_{a}^{(k)}$ stands for the average shear stress in the $k$th layer of thickness $2h^{(k)}$

$$
T_{a}^{(k)} = \frac{1}{2h^{(k)}} \int_{z_{b}^{(k)}}^{z_{t}^{(k)}} \tau^{(k)}_{ax} dz
$$

(12)

Moreover

$$
F_{b}^{(k)} = \frac{3}{4} \left( z_{m}^{(k)} \right)^{2} - \frac{1}{2} z_{m}^{(k)} - 1; F_{t}^{(k)} = \frac{3}{4} \left( z_{m}^{(k)} \right)^{2} + \frac{1}{2} z_{m}^{(k)} - 1; F_{m}^{(k)} = \frac{3}{4} \left( 1 - \left( z_{m}^{(k)} \right)^{2} \right)
$$

(13)

with $z_{m}^{(k)} = \left( z_{m}^{(k)} + 1 \right) \in [-1;1]$, $z_{m}^{(k)}$ representing the coordinate of the $k$th midplane.

The mixed-field formulation which is based on the RZT kinematic assumptions, Eq. (2), and the assumed transverse shear stresses, Eq. (11), is herein denoted as $RZT^{(m)}_{2}$. 

**3 RZT^{(m)} GOVERNING EQUATIONS**

The RZT^{(m)} plate equilibrium equations are derived from Reissner’s mixed variational theorem, wherein the body forces are neglected. The plate is subjected to a transversely distributed pressure $\overline{q}$, acting on the midplane $\Omega$, to tangential surface tractions $\overline{p}_{a}^{(l)}$ and
\( \bar{p}_\alpha^{(b)} \), acting along the \( x_\alpha \)-direction and applied at the top, \( S_t \), and bottom, \( S_b \), plate surfaces, respectively, and to the tractions \( \{ \bar{T}_{\alpha}, \bar{T}_z \} \) applied on a part of the midplane boundary.

The left-hand side of Eq. (1) has two contributions: the first term is the variation of the strain energy, the second term enforces compatibility between the \( \gamma \) and \( \gamma_\alpha \) strain fields. These two terms are decoupled since the variational operator is applied to independent variables; hence, the compatibility term may be treated separately, that is

\[
\int_{\Omega} \int_{-b}^{b} \delta \tau_\alpha \left( \gamma - \gamma_\alpha \right) dz d\Omega = 0 \tag{14}
\]

Regardless of the type of approximation, the assumed transverse shear stresses have the general form given by Eq. (5). Substituting Eq. (5) into Eq. (14) and performing the variational operation and through-the-thickness integration yields the following expression for the transverse shear stresses

\[
\tau_\alpha = Z_g(z)g + Z_\psi(z)\psi + Z_e(z)\eta_v
\tag{15}
\]

where \( g^T \equiv \{ \gamma_1, \gamma_2 \} \), \( \psi^T \equiv \{ \psi_1, \psi_2 \} \), and where \( Z_g(z) \), \( Z_\psi(z) \) and \( Z_e(z) \) are matrices of functions of the \( z \)-coordinate.

Performing the integration by parts on the remaining part of Eq. (1)

\[
\int_{\Omega} \int_{-b}^{b} \left( \delta \varepsilon^T \sigma \right) dz d\Omega = \delta W_v
\tag{16}
\]

gives rise to the equilibrium equations of RZT\(^{(m)}\)

\[
N_{\alpha\beta} + \bar{p}_\alpha = 0; \quad Q_{\alpha\alpha} + \bar{q} = 0; \quad M_{\alpha\beta} - Q_\alpha + \bar{m}_\alpha = 0; \quad M^e_{\alpha\beta} - Q^e_\alpha = 0
\tag{17}
\]

and a set of variationally consistent boundary conditions that are identical to those of RZT (refer to [9]). In Eq. (17), \( \bar{p}_\alpha = \bar{p}_\alpha^{(i)} + \bar{p}_\alpha^{(b)} \), \( \bar{m}_\alpha = h(\bar{p}_\alpha^{(i)} - \bar{p}_\alpha^{(b)}) \); \( N_{\alpha\beta} \), \( M_{\alpha\beta} \), and \( M^e_{\alpha\beta} \) are the same stress resultants as those defined in RZT [9]. For this new theory, the transverse shear stress resultants are given by

\[
\begin{align*}
\{Q\} &= \int_{-b}^{b} \left\{ r^{(i)}_{\alpha z} \right\} dz = A_1 g + B_1 \psi + E_1 \eta_v \\
\{Q^e\} &= \int_{-b}^{b} \left\{ r^{(b)}_{\alpha z} \right\} dz = C_1 g + D_1 \psi + F_1 \eta_v
\end{align*}
\tag{18}
\]

where \( A_1, B_1, C_1, D_1, E_1, \) and \( F_1 \) are the resulting shear-stiffness coefficient matrices.
4 NUMERICAL RESULTS

To assess the accuracy of both $RZT_{1}^{(m)}$ and $RZT_{2}^{(m)}$, exact and approximate solutions are developed for the bending of rectangular plates defined on the domain $x \in [0,a], \ y \in [0,b], \ z \in [-h,h]$: (1) A plate that is simply supported along the four edges and subjected to a bi-sinusoidal transverse pressure, and (2) A cantilever plate subjected to a uniform transverse pressure. For problem (1), an exact solution is readily obtained using trigonometric expansions for the kinematic variables. For problem (2), an approximate solution is obtained using the Rayleigh-Ritz method, where the kinematic variables are approximated using the Gram-Schmidt polynomials. For details on these solutions, refer to [9].

The mechanical material properties and the stacking sequences are summarized in Tables 1 and 2.

### Table 1: Mechanical properties of orthotropic materials

<table>
<thead>
<tr>
<th>Material</th>
<th>$E_{1}^{(e)} / E_{2}^{(e)} / E_{3}^{(e)}$ (GPa)</th>
<th>$G_{12}^{(e)} / G_{13}^{(e)} / G_{23}^{(e)}$ (GPa)</th>
<th>$v_{12}^{(e)} / v_{13}^{(e)} / v_{23}^{(e)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>157.9/9.584/9.584</td>
<td>5.93/5.93/3.227</td>
<td>0.32/0.32/0.49</td>
</tr>
<tr>
<td>F</td>
<td>50/10/10</td>
<td>5/5/5</td>
<td>0.25/0.25/0.25</td>
</tr>
<tr>
<td>N</td>
<td>$10^{-5}/75.85 \times 10^{-3}$</td>
<td>$22.5 \times 10^{-3}/22.5 \times 10^{-3}$</td>
<td>0.01/0.01/0.01</td>
</tr>
<tr>
<td>Q</td>
<td>525/21/21</td>
<td>10/5/10</td>
<td>0.25/0.25/0.25</td>
</tr>
</tbody>
</table>

### Table 2: Laminate stacking sequences (layers sequence is in the positive $z$-direction)

<table>
<thead>
<tr>
<th>Laminate</th>
<th>Normalized lamina thickness, $h/l/h$</th>
<th>Lamina materials</th>
<th>Lamina orientation (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>(0.5/0.5)</td>
<td>(A/A)</td>
<td>(0/90)</td>
</tr>
<tr>
<td>L2</td>
<td>(0.25/0.25/0.25/0.25)</td>
<td>(A/A/A/A)</td>
<td>(0/90/0/90)</td>
</tr>
<tr>
<td>L3</td>
<td>(0.1/0.1/0.25)</td>
<td>(F/N/Q)</td>
<td>(0/0/0)</td>
</tr>
</tbody>
</table>

**Problem 1.** A simply supported rectangular plate ($b=3a$) subjected to a bi-sinusoidal transverse pressure, $q(x_{1},x_{2}) = q_{0}\sin(\pi x_{1}/a)\sin(\pi x_{2}/b)$.

For comparison purposes, both the three-dimensional elasticity [16] and exact RZT [9] solutions are used. Two cross-ply laminates (L1 and L2) are considered, and the non-dimensional maximum deflections corresponding to various values of the span-to-thickness ratio, $a/2h$, are compared. Results in Table 3 demonstrate that the use of a mixed-field formulation can lead to slight enhancements of the deflection predictions when compared with the displacement-based RZT, since the improvements in the transverse shear stresses contribute directly to the transverse-shear stiffness. In addition, it is evident that $RZT_{1}^{(m)}$ yields consistently more accurate results than the $RZT_{2}^{(m)}$ formulation. Similar observations in relation to the RZT beam analysis were made by Gherlone [15], where it was also pointed out that the mixed-field zigzag formulations retain the same level of accuracy for the in-plane response quantities as those obtained using the displacement-based formulations.
Subsequently, to focus on the transverse shear stresses, their through-the-thickness distributions are examined in detail.

Table 3: Problem 1, Laminates L1, L2: Normalized maximum (central) deflection, 
\[ \bar{w} = \left(10^2 D_i / q_a a^4\right) w(a/2, b/2) \].

<table>
<thead>
<tr>
<th>Laminate</th>
<th>(a/2h)</th>
<th>3D Elasticity</th>
<th>RZT</th>
<th>(RZT_{1}^{(m)})</th>
<th>(RZT_{2}^{(m)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>8</td>
<td>2.546</td>
<td>2.512</td>
<td>2.547</td>
<td>2.545</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2.449</td>
<td>2.427</td>
<td>2.449</td>
<td>2.448</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.319</td>
<td>2.314</td>
<td>2.319</td>
<td>2.319</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>2.283</td>
<td>2.282</td>
<td>2.283</td>
<td>2.283</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>2.278</td>
<td>2.277</td>
<td>2.278</td>
<td>2.278</td>
</tr>
<tr>
<td>L2</td>
<td>8</td>
<td>1.505</td>
<td>1.472</td>
<td>1.507</td>
<td>1.487</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.375</td>
<td>1.353</td>
<td>1.375</td>
<td>1.362</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.198</td>
<td>1.193</td>
<td>1.198</td>
<td>1.195</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.148</td>
<td>1.147</td>
<td>1.148</td>
<td>1.148</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.141</td>
<td>1.141</td>
<td>1.141</td>
<td>1.141</td>
</tr>
</tbody>
</table>

Figure 1 shows a comparison of the through-the-thickness distributions of normalized transverse shear stresses for laminate L1 (\(a/2h = 8\)). The \(RZT_{1}^{(m)}\) solution is highly accurate as evidenced by the comparison with the reference solution, labeled as 3D Elasticity. Moreover, the RZT and RZT-Integrated solutions refer to the transverse shear stresses based on, respectively, the constitutive equations and the integration of three-dimensional equilibrium equations. The integrated shear stresses are also very accurate. However, in contrast to the \(RZT_{1}^{(m)}\) solution, the integrated stresses are obtained by means of a post-processing scheme that makes use of second-order partial derivatives of the kinematic variables. On the other hand, the \(RZT_{2}^{(m)}\) stresses are significantly less accurate because they follow closely the piecewise-constant distributions of the RZT stresses that are derived from constitutive relations. This behavior is due to the fact that the number of stress variables inherent in the polynomial scheme is proportional to the number of layers. In contrast, the \(RZT_{1}^{(m)}\) formulation uses only two stress variables for each of the transverse-shear stress components, and these variables are independent of the number of layers (see details in Section 2.2.)

**Problem 2. A cantilever square plate subjected to a uniform transverse pressure**, 
\[ q(x_1, x_2) = q_0 \].

For this problem, the exact solution does not exist. To assess the accuracy of \(RZT_{1}^{(m)}\), a high-fidelity FEM (MSC/MD-NASTRAN® [17]) solution is used. The model is regularly discretized using linear-strain solid elements, HEXA8. There are sixty-five elements along each span direction, five elements through the thickness of the bottom face, eight elements along the top face and fifteen elements through the core thickness.
In Figure 2, five through-the-thickness distributions of the transverse shear stress are compared, with the $RZT_1^{(m)}$ distribution demonstrating the closest correlation with the three-dimensional FEM solution, that is comparable in accuracy only to the RZT-Integrated solution.

![Figure 1: Problem 1, Laminate L1: Through-the-thickness distribution of normalized transverse shear stresses $\bar{\tau}_{zz}(0,b/2,z)$.](image1)

![Figure 2: Problem 2, Laminate L3: Through-the-thickness distribution of normalized transverse shear stresses $\bar{\tau}_{zz}(a/2,0,z)$.](image2)
5 CONCLUSIONS

In this paper, a multi-scale refined zigzag theory for the analysis of multilayered composite and sandwich plates has been presented. The new theory, called RZT(m) (Refined Zigzag Theory, Mixed-formulation), is developed following the RZT(m) formulation for beam analysis recently proposed by Tessler. The approach is based on Reissner’s mixed variational theorem and on the kinematic assumptions of the Refined Zigzag Theory (RZT). Two alternative formulations for the transverse-shear stresses were examined. The first, RZT1(m), follows the RZT(m) beam formulation by Tessler, which is based on a closed-form integration procedure of elasticity-theory equilibrium equations. The second formulation, RZT2(m), incorporates Murakami’s polynomial approach.

Results for simply supported and cantilevered rectangular laminated composite and sandwich plates in bending have been examined using the RZT1(m), RZT2(m) and the original RZT (displacement-based) formulation. For comparison purposes, corresponding results were obtained using three-dimensional elasticity and high-fidelity FEM solutions. Both RZT(m) formulations have demonstrated improved modeling of the transverse shear stiffness and stresses, achieving slightly more accurate deflection predictions when compared with those of RZT. Of the two formulations, however, RZT1(m) produced consistently superior results. In contrast, the polynomial-based formulation, RZT2(m), has shown to exhibit serious deficiencies: the assumed transverse shear stresses tend to approximate closely those of their constitutive counterparts, leading to unsatisfactory results.

In this paper it has been clearly demonstrated that a mixed-field formulation, based on Reissner’s variational theorem and coupled with the refined zigzag theory for plates, may offer substantial improvements in predicting the transverse-shear stresses and stiffness. The accuracy of a mixed-field formulation depends on the choice of the assumed transverse shear stresses. The transverse shear stresses obtained by integration of three-dimensional equilibrium equations produced an improved theory, RZT1(m), that offers superior predictions of the in-plane behavior, typical of RZT, along with the enhanced predictions for the transverse shear stresses and stiffness. This new formulation is thus perfectly suited for the analysis of multilayered composite and sandwich structures. The RZT1(m) formulation may also be used for developing simple and efficient C0-continuous plate and shell finite elements.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of this research by the NASA Aviation Safety Program (Vehicle Systems Safety Technologies Project). Special thanks are extended to the National Institute of Aerospace (NIA) and NASA Langley research Center, Hampton, Virginia, for hosting Professors Marco Di Sciuva and Marco Gherlone in the fall of 2012. Luigi Iurlaro, Marco Gherlone and Marco Di Sciuva also acknowledge the financial support from the European Commission in the framework of FP7-CLEANSKY-GRA project.
REFERENCES


EFFECTIVE MODIFICATIONS TO DIFFERENTIAL EVOLUTION OPTIMIZATION ALGORITHM

ERIC J. INCLAN* AND GEORGE S. DULIKRAVICH*

*Florida International University (FIU)
Multidisciplinary Analysis, Inverse Design, Robust Optimization and Control (MAIDROC) Lab.
Department of Mechanical & Materials Eng., 10555 West Flagler St., Miami, Florida 33174, USA
e-mail: eincl001@fiu.edu; dulikrav@fiu.edu; web page: http://maidroc.fiu.edu

Key words: Optimization, Differential Evolution, Single Objective

Summary. Many inverse problems utilize optimization algorithms to perform minimization of least squares norms in an accurate, reliable and computationally efficient manner. Due to the cost of evaluating real-world objective functions, optimization algorithms must be both fast and robust. Differential Evolution (DE) algorithm is known for its robustness, but not its speed. This paper proposes four simple modifications to DE and compares their performance to Particle Swarm (PS) algorithm using a subset of the Schittkowski & Hock test cases. With these techniques, DE is observed to converge to the global minimum up to three times faster than PS, while maintaining robustness in some cases, and generally performs better than the original forms of DE.

1 INTRODUCTION

Researchers have utilized optimization algorithms when solving inverse problems and performing inverse shape design because a variety of inverse problems can be formulated as optimization problems [1]. In these situations, the effectiveness of the optimization algorithm will dictate the quality and computational cost of the solution. Over the past ten years two optimization algorithms, namely DE and PS, have enjoyed widespread use due to their outstanding performance in a variety of applications. As with any stochastic method, there is always room for improvement but the difficulty often lies in justifying the increased complexity of the new subroutine. This paper proposes four simple modifications to DE that improve its speed, robustness, or both. Since PS is generally considered to be the faster of the two [12], it will be used as the benchmark method.

2 OPTIMIZATION ALGORITHMS

2.1 Differential Evolution (DE)

Differential Evolution (DE) utilizes an equation (referred to here as the “update equation”) in order to generate a new solution, and replaces an existing solution with the new one if the new solution is superior.

The standard DE algorithm can be described as follows.

1) Create initial population of candidate solutions.
2) Evaluate objective function(s) for each solution.
3) Begin main loop:
   a) Copy original population to temporary population.
   b) For each solution in the temporary population,
      i. Create a new solution:
         1. Randomly select one dimension, \( j \), of solution and apply update equation.
         2. For each dimension (excluding \( j \)) of the current solution,
            a. If \( R < CR \), apply update equation, otherwise, leave unchanged.
      ii. Evaluate objective function(s) for new solution.
      iii. Compare new solution to corresponding solution from original population
      iv. If new solution is superior to original solution, replace the original with the
          new solution.
   4) End main loop once population converges or maximum number of iterations is
      reached.

In the above algorithm, \( CR \) is a user-defined scalar \( \epsilon [0,1] \) known as the “crossover rate,” and \( R \) is a uniformly distributed, random number \( \epsilon [0,1] \).

There are many forms of DE currently in use, several of which vary only by the update equation used. Three particularly successful forms are the so-called rand/1/bin, and best/2/bin proposed in [10] as well as Donor3 proposed in [2]. Their respective update equations are as follows,

\[
Y_k = X_{r_{1,k}} + F(X_{r_{2,k}} - X_{r_{3,k}}) \quad (1)
\]

\[
Y_k = X_{best,k} + F(X_{r_{1,k}} + X_{r_{2,k}} - X_{r_{3,k}} - X_{r_{4,k}}) \quad (2)
\]

\[
Y_k = \frac{\lambda_1 X_{r_{1,k}} + \lambda_2 X_{r_{2,k}} + \lambda_3 X_{r_{3,k}}}{\lambda_1 + \lambda_2 + \lambda_3} + F(X_{r_{2,k}} - X_{r_{3,k}}) \quad (3)
\]

where \( Y \) is the resulting coordinate in the \( k^{th} \) dimension of the new solution, and \( F \) is a weighting factor (a user-defined scalar \( \epsilon [0,2] \)). The variable \( X \) denotes a coordinate from an existing solution vector, and the subscript \( r \) indicates that it was randomly selected. Therefore, the component \( Y \) in Eq. 1, for rand/1/bin, is a linear combination of components from three distinct, randomly selected solutions, while Eq. 2, for best/2/bin is the linear combination of four distinct, randomly selected solutions and the global best solution. The Donor3 method utilizes a weighted average of three components, where \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are uniformly distributed, random numbers \( \epsilon [0,1] \).

2.2 Particle Swarm (PS)

Published within two years of DE, Particle Swarm (PS) has become very popular due to its simplicity and speed. It is based on the social behavior of various species [4] and, like DE, uses an update equation to generate new solutions.

The basic PS algorithm is given below.

1) Create initial population of solutions.
2) Evaluate objective function(s) for each solution.
3) Store copy of initial population to serve as “individual best” vectors and store the “global best.”
4) Begin main loop:
   a) For each solution in the population,
      i. Apply update equation.
      ii. Evaluate objective function(s) using new solution.
      iii. Replace “individual best” with new solution if new solution is superior.
   b) Replace “global best” if best new solution is superior to previous “global best.”
5) End main loop once population converges or maximum number of iterations is reached.

The update equations are,

\[
\hat{X}^{g+1} = \hat{X}^{g} + \hat{V}^{g} \quad (4)
\]

\[
\hat{V}^{g} = \alpha \hat{V}^{g-1} + \beta R_{1}(\hat{X}_{\text{best},i} - \hat{X}^{g}) + \beta R_{2}(\hat{X}_{\text{best},G} - \hat{X}^{g}) \quad (5)
\]

where \( \hat{V} \) is the so-called “velocity vector,” \( \alpha \) and \( \beta \) are user defined scalars, \( g \) is the iteration number, and \( R_{1} \) and \( R_{2} \) are uniformly distributed random numbers \( \epsilon [0,1] \). The vector \( \hat{X}_{\text{best},i} \) is the vector corresponding to the best value ever held by the \( i \)th solution, and \( \hat{X}_{\text{best},G} \) is the best solution ever found (also known as the “global best”). The first term on the right of Eq. 5 is the “inertia,” which is effectively a scalar multiple of the velocity from the previous iteration. Since the velocity magnitude can grow each iteration, it is possible for the inertia to become so large that PS oscillates erratically or diverges altogether.

To help mitigate this risk the following condition was imposed (written in C-style pseudo-code),

\[
\text{while } (V_k > \text{domain}/3) \\
V_k = V_k / 2 \quad (6)
\]

where \( V_k \) is the \( k \)th coordinate of the velocity vector. The condition above causes the velocity component to be halved until its magnitude is less than one-third the width of the search space for that coordinate. This restriction is applied before the velocity is added to the individual vector. A second restriction was also added to the velocity vector, as follows (again in pseudo-code)

\[
\text{while } (X_{i,k} + V_{i,k} \text{ outside domain}) \\
V_{i,k} = V_{i,k} \times 0.9 \quad (7)
\]

This restriction causes the entire velocity vector to be scaled down further if it would cause the new solution to move outside of the search domain.

3 MODIFICATIONS

Over the years many modifications to DE have been proposed. Some of these modifications relate to the update equation [6], while others involve merging DE with some
other technique [11,3]. This paper proposes four modifications that draw more performance out of DE while minimizing the changes made to the implementation of DE.

3.1 Randomly Varying Parameters

It is widely known that the values of F and CR greatly impact the performance of DE. Some authors have suggested values based on their experience [10], and others have used meta-optimization to determine what values to use [7]. However, no single value provides optimal performance for all problems, and many values are only good for specific problems. To date, two authors have discussed randomly varying one parameter, F, between a fixed range [11, 5]. It will be shown that randomly varying both F and CR will dramatically improve the convergence speed and robustness of the DE methods in many test functions.

3.2 Special Vectors

Since DE involves the weighted average of a collection of vectors, it may be beneficial to include two special vectors in the population after each iteration. This paper proposes including an average vector, and a weighted average vector. The average vector is simply the average of all vectors in the population. Similarly, the weighted average vector is scaled based on the superiority of each vector (the better the solution the larger the weight). In this paper, each solution was non-uniquely ranked in reverse order (the best solution had a rank equal to the population number, and the worst solution had a rank equal to one), and the weights were simply the value of the rank. In order to include these two vectors, the population was increased by two regardless of the problem dimensionality. At the end of each generation, the two worst solutions from the population were replaced by the average vector and weighted average vector in order to make them available for selection during the following update step of DE. This is expected to increase the convergence speed of DE whenever the population converges within a convex region, which it often does at some point in most optimization problems.

3.3 Sorted Comparisons

The standard practice in DE is to compare a given new solution to a corresponding existing solution, but the selection of solutions to compare is purely arbitrary. Due to the computational expense of evaluating the objective function for each new solution, sorting the mutant population may help maximize the benefit of each function evaluation. In this paper, two sorting methods are proposed: “Most” and “Best”. The Most method sorts the new solutions so that the maximum possible number of replacements is achieved. While conducting this research, the Most method did significantly improve DE in the majority of test cases, therefore, those results will be withheld. The Best method sorts the existing population from best to worst, and the new solution population from worst to best. This way, only the best solutions of each population are preserved after each comparison. This method will be shown to greatly increase convergence speed but reduce robustness, causing the algorithm to converge to local minima more frequently.
3.4 Dynamic Random Generation Rate

Sometimes, when an algorithm is too greedy, it is possible to increase the algorithm’s robustness by including a probability of randomly generating a new solution. DE and PS do not typically include this. However, since the Best method may be too greedy, such a change is warranted. To ensure that the Random Generation Rate (RGR) is used only when the algorithm converges, the RGR will vary dynamically based on the population’s convergence rate. This rate will be measured by storing the average objective function values of the population for the five most recent iterations.

![Figure 1: Example of RGR data and slopes](image)

Using those values, a least-squares-fit linear regression will be drawn through the five points (shown in Fig. 1). Since one might want to consider the effects of concavity on the RGR, a second line will be fit through the last three points in order to implicitly capture this information (albeit imperfectly). Finally, a weighted average of the two slopes will be taken. This final value will be the slope used to calculate the RGR. The RGR itself will be calculated using the binary sigmoid equation,

\[ \mu = \frac{2\mu_{\text{max}}}{1 + e^{-\sigma m}} \quad (8) \]

where \( \mu \) is the RGR, \( \mu_{\text{max}} \) is the maximum rate set to 0.1, \( \sigma \) is a user-defined parameter set to 1.5, and \( m \) is the slope, which is assumed to be nonpositive since DE converges monotonically. Note that in order to make this method valid for objective function values of any order of magnitude, the slopes are first translated so that the most recent value is always zero, and then scaled using a simple division operation. Additionally, since this mutation method will cause DE to no longer decrease monotonically, the random generation rate is set to zero whenever the slope of either linear regression is positive.

Of all the methods proposed, this is by far the most complicated. Nevertheless, it will be shown that it can improve the overall performance of DE when used in conjunction with the Best method.
4 RESULTS

In order to evaluate the impact of these modifications, the algorithms were applied to a subset of the Schittkowski and Hock test cases [8][9], listed in Table 1. Constraints were enforced using a penalty function.

<table>
<thead>
<tr>
<th>Table 1: Dimension and Population of Various Test Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

Each test problem was optimized 50 times, for 200 generations using the population sizes in Table 1. The algorithm used to generate these population values has been witheld because it is still undergoing research. The unmodified versions of DE have a weighting factor $F = 0.8$ and crossover rate $CR = 0.9$. The randomly varying versions are defined as follows: rand/1/bin uses $F \in [0.4,0.8]$, $CR \in [0.7,0.9]$, best/2/bin uses $F \in [0.2,0.8]$, $CR \in [0.6,1.0]$, and Donor3 uses $F \in [0.4,0.8]$, $CR \in [0.4,0.8]$. In PS, the user-defined scalars $\alpha$, and $\beta$ are set to 0.5 and 2, respectively. To save space in the following graphs and tables, rand/1/bin is denoted STD, best/2/bin is denoted BST, and Donor3 is denoted DN3. The letters following the acronym correspond to the subsections of Section 3 (A = randomly varying parameters, B = special vectors, C = sorted comparisons, and D = dynamical RGR). Note that the special vectors were only used with best/2/bin.

Table 2 shows the aggregate accuracy of each method (note PS is shown in bold next to rand/1/bin). The first row for each method, labeled “Mean,” lists the percentage of test problems for which a given method obtained the lowest average minimum. In some test problems, multiple methods converged to the global minimum, therefore, the sum of these rows exceeds 100%. The second row is the percentage of test problems for which a given method obtained the lowest standard deviation of minima. As indicated in Table 2, BST-A had the best performance (mean and standard deviation) of any method for the 48 test problems selected. The third row lists the percentage of test problems for which a given method found the best single result, while the fourth row relates to the worst single result. The use of randomly varying parameters dramatically improves the performance of best/2/bin. The other modifications also improve its performance but usually at the cost of robustness.
Nevertheless, all modified versions of best/2/bin outperform PS. Donor3’s robustness generally worsens, while the STD-A,C method rivals PS in robustness and accuracy.

Table 2: Comparison of Method Accuracy

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>8.3%</td>
<td>8.3%</td>
<td>22.9%</td>
<td>12.5%</td>
<td>22.9%</td>
<td>12.5%</td>
<td>4.2%</td>
<td>6.3%</td>
<td>0.0%</td>
<td>8.3%</td>
<td>50.0%</td>
<td>33.3%</td>
<td>35.4%</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>8.3%</td>
<td>8.3%</td>
<td>20.8%</td>
<td>16.7%</td>
<td>22.9%</td>
<td>12.5%</td>
<td>4.2%</td>
<td>6.3%</td>
<td>0.0%</td>
<td>8.3%</td>
<td>47.9%</td>
<td>33.3%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Best</td>
<td>16.7%</td>
<td>25.0%</td>
<td>31.3%</td>
<td>29.2%</td>
<td>41.7%</td>
<td>18.8%</td>
<td>27.1%</td>
<td>29.2%</td>
<td>22.9%</td>
<td>10.4%</td>
<td>56.3%</td>
<td>70.8%</td>
<td>58.3%</td>
</tr>
<tr>
<td>Worst</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>10.4%</td>
<td>0.0%</td>
<td>2.1%</td>
<td>39.6%</td>
<td>12.5%</td>
<td>31.3%</td>
<td>0.0%</td>
<td>4.2%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3 shows the aggregate speed of each method. Here, the “Mean” lists the percentage of test problems for which a given method obtained the highest average rate of convergence (including premature convergence to local minima). The second row is the percentage of test problems for which a given method obtained the lowest standard deviation of convergence rates. The third row lists the percentage of test problems for which a given method converged the fastest, while the fourth row relates to the slowest convergence. The convergence rate was based on the best individual of the population for each generation of the optimization run. The convergence rate is the slope of the secant line, subtracting the final objective function value of the global best from the initial objective function of the global best and dividing by the number of generations.

Table 3: Comparison of Method Convergence Rates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>12.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>10.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>8.3%</td>
<td>35.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>27.1%</td>
<td>14.6%</td>
<td>4.2%</td>
<td>0.0%</td>
<td>4.2%</td>
<td>6.3%</td>
<td>4.2%</td>
<td>0.0%</td>
<td>2.1%</td>
<td>27.1%</td>
<td>6.3%</td>
<td>2.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Best</td>
<td>2.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>10.4%</td>
<td>0.0%</td>
<td>2.1%</td>
<td>12.5%</td>
<td>0.0%</td>
<td>4.2%</td>
<td>0.0%</td>
<td>18.8%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Worst</td>
<td>4.2%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>10.4%</td>
<td>2.1%</td>
<td>2.1%</td>
<td>33.3%</td>
<td>16.7%</td>
<td>27.1%</td>
<td>0.0%</td>
<td>2.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

From Table 3 we see that, when applied to best/2/bin, the true benefit of the B, C, and D modifications is speed. In nearly half of the test problems BST-A,B,C,D had the single highest speed of any method, and in one third of problems had the highest average convergence rate of any method. One of the most striking examples of improvement in best/2/bin is test problem (TP) 201. Each convergence history, like those in Fig. 2, contains the average convergence history for a particular method. As shown in Fig. 2, the unmodified versions of rand/1/bin and best/2/bin are slower than PS, while Donor3 is slightly faster. The fully modified best/2/bin (BST- A,B,C,D in Fig. 2b) is roughly four times faster than the
unmodified version, and roughly three times faster than PS. The STD-A,C,D (Fig. 2a) is the fastest of the rand/1/bin versions, and is nearly 35% faster than PS. The DN3-A version of Donor3 (Fig. 2c) is slightly faster than the original version. While the other modifications are faster, they also cause Donor3 to converge to local minima as indicated by the curve’s rapid initial descent followed by a region of near-zero slope. The test cases TP 5, 8, and 88 show similar behavior.

For some test cases like TP 45, or TP 4 shown in Fig. 3, PS dominates all variations of DE. Here, PS is over four times faster than any DE variation, and much more robust. However, in several cases the proposed modifications of DE can make methods that normally perform worse than PS perform much better. As shown in Fig. 4, each modification causes best/2/bin to outperform PS in speed and robustness, whereas the unmodified version could not. While randomly varying parameters increases its speed, adding sorted comparisons dramatically improves its robustness. The special vectors provide an additional speed boost while slightly decreasing robustness. The rand/1/bin method shows similar improvements to a lesser degree. However, STD-A,C,D shows slower convergence than STD-A,C (Fig. 4a). In best/2/bin, the
**Figure 3:** PS Convergence History for TP 4

**Figure 4:** Convergence Histories for TP 271
BST-A,C,D version is just as fast, but less robust than BST-A,C (Fig. 4b). In Donor3, only randomly varying parameters increases speed and, in this case, preserves robustness. The other variations increase speed, but reduce robustness. Similar behavior can be observed in TP 110, 249, 266, 281, 283 and 288.

In cases where the objective function is convex, using special vectors can provide an advantage early on as shown in Fig. 5. The DN3-A (Fig. 5b) version converges faster than any best/2/bin variation, but the use of special vectors compensates for that disadvantage. Similar behavior is observed in TP 291, 292, 293, 302, 384, 391 and 394. Note that the objective function need not be strictly convex in order for the special vectors to improve performance.

As with any optimization algorithm, these modifications are not guaranteed to improve DE’s performance for every function. Sometimes, although improvements to speed can be observed, DE can still be trapped in local minima. Fig. 6 shows improvements to each DE method’s speed, but not robustness. Again best/2/bin (Fig. 6b) can outpace PS, and rand/1/bin (Fig. 6a) can match PS. Furthermore, these modifications do not always improve behavior.
when applied sequentially (i.e., A, B, C will not always be superior to A, B, which will not always be superior to A).

![Figure 6: Convergence Histories for TP 1](image)

### 5 CONCLUSIONS

The single most effective modification to rand/1/bin and best/2/bin is randomly varying parameters based on the ranges provided in this paper. The use of additional modifications has been shown to greatly improve the performance of best/2/bin, making it faster and more robust than particle swarm algorithm in many cases. The effects of these modifications on Donor3 are mixed.

### ACKNOWLEDGEMENTS

This work was partially funded by the US Air Force Office of Scientific Research under grant FA9550-12-1-0440 monitored by Dr. Ali Sayir. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the US Air Force Office of...
REFERENCES


A COUPLED THERMO-ELECTRICAL PROBLEM FOR THE OPTIMIZATION OF RADIO FREQUENCY ABLATION TREATMENTS

ALESSANDRO FORMISANO, FABRIZIO FERRAIOLI AND RAFFAELE MARTONE

Dipartimento di Ingegneria Industriale e dell’Informazione
Seconda Università DEGLI Studi di Napoli
Via Roma 29, Aversa (CE), Italy
Email: Raffaele.Martone@unina2.it

Key words: Coupled Problems, Multiphysics Problems, Biomedical applications, Numerical electromagnetic computation, Optimization methods

Abstract. Clinical treatment of some forms of epathocarcinoma makes use, under certain circumstances, of thermal ablation induced by radio frequency currents injected into the tumor volume. The success of the operation relies upon the capability of totally ablating the tumor volume simultaneously preventing healthy tissue damage and skin burns due to the ground pad heating. Such a treatment is liable to possible improvements if a mathematical model is available, and a suitable numerical optimization procedure is adopted. The problem falls in the class of strongly coupled problems; as matter of fact the thermal diffusion is linked to the electrical current conduction through the main heat source which is the Ohmic power density while, in addition, the current conduction is linked to the thermal dynamics being the electrical resistivity a function of the thermal map itself.

In this paper, a FEM based numerical model for radio frequency ablation of liver tumors is considered and its optimization is addressed with optimal planning, imposing non linear partial differential equation of the model as a constraint in a space state; in addition, the coupling aspects are highlighted and discussed.

1 INTRODUCTION

Radio Frequency Ablation (RFA) is a treatment for some forms of liver tumors, based on the local deposition of Ohmic power by injection of radiofrequency currents though suitably designed probes inserted into patient's abdomen [1, 2]. A typical RFA session requires the laparoscopical or percutaneous insertion of a probe into the ill tissue, the positioning of the probe tip into the tumor, the application of a single or multiple return electrode, usually on the back of the patient [3, 4], and, finally, the injection of RF current through a specific power supply system. Note that the probe tip is a rather sophisticated device, eventually endowed with a cooling water flow and metallic retractable prongs designed to operate in electric contact with the tumor [5].

The Ohmic power deposition produces a heat source whose intensity depends on the current density and tissues conductivity. The Ohmic power density combines with other heat sources, such as metabolic heat generation, and with heat sinks, such as blood perfusion, to determine the temperature evolution in the tumor and adjacent regions. The RFA process is designed to determine a net temperature increase inside the tumor high enough to lead tumor cells to necrosis avoiding tissue carbonization. in addition, the ideal goal of the treatment is to achieve the necrosis exclusively in the tumor volume in as few sessions as possible [5, 2].
The mathematical modeling requires several equations falling in different technical fields: (i) a thermal non-linear, dynamic diffusion equation; (ii) a stationary ac current equation; (iii) a non-linear, integral equation to describe the cellular necrosis. Further equations can be required to represent additional elements as the electrical circuit powering the process. Finally, to look for a favorable solution, an additional non-linear optimization function is used and its minimum is sought. Due to the interaction among the equations, the problem has to be faced by means of a well-suited coupled problem approaches.

The time evolution of the temperature map and the consequent tissues damage index depend on a number of parameters, many of them either only partially known (such as thermal and electrical tissues conductivities), or not under the control of the surgeon (such as the patient’s physiological reaction to the heat deposition). A number of papers were published about possible improvements to the basic protocol [2, 3, 4, 5]. In particular, non-linear programming methods have been introduced as a valuable tool to approach the problem; it has also been highlighted that the use of multiple return pads can be used to suitably control the shape of the isothermal surfaces in order to prevent possible damage to healthy tissue [6, 7, 8], or that the use of multi-prong probes could improve the heat deposition in the tumor, while preserving adjacent tissues from excessive damages.

This paper proposes a novel and more general formulation for the optimization of the current waveforms in the RFA electrodes, based on optimal planning theory [9, 10]. The advantages achievable in terms of the effectiveness are assessed by using a suitable numerical model of the human body specifically developed for this application.

In sect. II the problem formulation is presented and the coupling aspects highlighted; in Sect. III a numerical example is presented to demonstrate the effectiveness of the approach; in addition, a comparison of the optimal planning of the classical monopolar heating and of an innovative multipolar structure is discussed; in Sect. IV an example of waveforms optimization is presented; finally, in Sect. V some conclusions are drawn.

2 MATHEMATICAL MODEL

The forward operator describing RFA dynamics is first briefly presented in section 2.1, and then some details about its Finite Element (FE) discretized version are given in section 2.2. The optimization strategy, the treatment of the differential constraints, required to treat the model in the context of the optimal planning algorithms [9], as well as the choice of degrees of freedom (DOF), are discussed in section 2.3.

2.1 Forward Problem: Electro-thermal dynamics of RFA

The mathematical model describing the temperature map evolution during RFA treatments (forward problem) falls in the coupled electro-thermal problem class, including two suitably coupled electromagnetic and thermal sub-models.

In the electromagnetic sub-problem, the Ohmic power density distribution is calculated, by assuming that the electrical properties of the tissue are known as well as the values of RF currents applied through inner needle-shaped electrodes and one or more outer return pads.

\[
\begin{align*}
\nabla \cdot (\sigma \nabla \varphi(r,t) + \varepsilon_0 \nabla \varphi(r,t)) &= 0 \quad \text{in } \Omega \\
\int_{\Gamma_{\text{ref,prong}}} j_k(r,t) \cdot dS &= I_k(t) \quad k \in \{1,2,\ldots,N_k\} \\
\varphi(r,t) &= 0 \quad \text{on } \partial \Omega_{\text{ref,prong}} \\
(\sigma \nabla \varphi(r,t) + \varepsilon_0 \nabla \varphi(r,t)) \cdot \mathbf{n} &= 0 \quad \text{on } \partial \Omega_{\text{skin}}
\end{align*}
\]
where $\Omega$ is the analysis domain (typically, the patient's torso), $\mathbf{r}$ is the position vector in $\Omega$, $\sigma$ is the electrical conductivity, $\varepsilon$ the electrical permittivity, $\phi$ is the electric scalar potential.

In this paper a multiple-prongs probe, with electrically independent elements, is assumed [7], and one of the prongs is supposed at system ground potential; different currents are then imposed in the remaining prongs as well as in the external (multiple) return pads, obtaining a flexible "multipolar" probe system. Consequently, $\partial \Omega_{\text{ref}, \text{prong}}$ is the surface of the prong where the reference (electric scalar) potential is applied, $\partial \Omega_{\text{elec}(k)}$, $k=1,2 \ldots N_E$, are patches modelling both the external and internal electrodes, and $\partial \Omega_{\text{skin}}$ is the patient's skin not covered by electrodes.

The electromagnetic sub-problem is coupled to the thermal one through the resistivity ($\rho=1/\sigma$), because of its temperature-dependence:

$$\rho \left( \mathbf{r}, f, T \right) = \rho_0 \left( \mathbf{r}, f \right) \left[ 1 + \beta \left( T - T_{\text{ref}} \right) \right]$$

(2)

where $T_{\text{ref}}$ is a reference temperature. In general the conductivity depends also on the working frequency.

The tissues temperature evolution during RFA is described by the Pennes's bioheat model [11]

$$\rho \left( \mathbf{r}, f, t \right) = \rho_0 \left( \mathbf{r}, f \right) \left[ 1 + \beta \left( T - T_{\text{ref}} \right) \right]$$

where

$$\begin{align*}
mc \frac{\partial T \left( \mathbf{r}, t \right)}{\partial t} &= \nabla \cdot \mathbf{k} \nabla T \left( \mathbf{r}, t \right) + p_0 \left( \mathbf{r}, t \right) - p_\infty \left( \mathbf{r}, t \right) \quad \text{in } \Omega \\
k \frac{\partial T \left( \mathbf{r}, t \right)}{\partial n} &= h_{\text{conv}} \left( T \left( \mathbf{r}, t \right) - T_\infty \right) \quad \text{on } \partial \Omega - \partial \Omega_{\text{probe}} \\
T \left( \mathbf{r}, t \right) &= T_{\text{cooling}} \quad \text{on } \partial \Omega_{\text{probe}} \\
T \left( \mathbf{r}, 0^+ \right) &= T_0 \quad \text{in } \Omega
\end{align*}$$

(3)

where $p_\infty \left( \mathbf{r}, t \right)$ is the Ohmic power density (in this case metabolic heat sources are neglected), $p_\infty = h_{\text{conv}} \left[ T \left( \mathbf{r}, t \right) - T_{\text{blood}} \right]$ is the tissue perfusion, $T_{\text{cooling}}$ is the temperature of the internal (cooled) probe, and suitable natural convection boundary conditions are imposed on the remaining part of domain boundary. In addition, $m$ is the tissue mass density, $c$ the tissue specific heat, $k$ the thermal conductivity, $h_{\text{conv}} = m_{\text{blood}} c_{\text{blood}} w_{\text{blood}}$ the convective heat transfer coefficient accounting for the blood perfusion ($m_{\text{blood}}$ is the blood mass density, $c_{\text{blood}}$ the blood specific heat, $w_{\text{blood}}$ the blood perfusion coefficient), $T_\infty$ is the temperature of blood, $h_{\text{conv}}$ is the convective coefficient, $T_\infty$ is the bulk temperature, and $T_0$ is the temperature map at $t = 0^-$.

In a coupled problems the complete set of equations in the mathematical model should be simultaneously faced at all the time steps and in all space points relevant for the analysis. As a consequence, the computing burden required for high accuracy evaluations, could be not compatible with many of the present computer performance.

Fortunately sometimes the actual dependence of some coefficient on the exchange variables (e.g. temperature from thermal analysis to determine conductivity in electromagnetic analysis, or Ohmic power density from electromagnetic to thermal analyses) is as weak to be neglected within the margins of the computation accuracy; in addition, sometime the time scale of the coupled equations can be so different to allow freezing, within the accuracy bar, a phenomenon while the others reach their steady state.

Therefore, in the solution of coupled dynamical system a preliminary analysis of the origins of the coupling (and their consistency) together with the comparison of the typical time scales, is strongly recommended because of the beneficial effects in the reduction of the computational burden.

In the case at hand it is easy to find a number of important simplifications:

S1. The time scale of thermal diffusion is in the order of hundreds of s, much longer then the electromagnetic one, in the order of fraction of µs.

S2. Except for conductivity and perfusion coefficients, all the dependences of the model coefficients
Then two consequences follow:

CS1. the current can be considered in AC steady state; then (i) the current can be studied in the complex phasor domain; (ii) in the optimization procedure, the currents can be represented by just the simple complex representation; (iii) the transient electrical effects can be neglected in the thermal dynamic studies;

CS2. the coupling between thermal and electrical models can be assumed concentrated in the impact of the temperature on conductivity coefficient;

CS3. the circuit sources are assumed “current controlled” and then not depending on the resistances.

According to the fist simplification, the electromagnetic equations (1) can be written in the complex phasor domain

\[
\begin{aligned}
\nabla \cdot (\gamma \nabla \phi(r)) &= 0 \quad \text{in } \Omega \\
\int_{\Gamma_{\text{elec}}} j_k(r) \cdot dS &= I_k \quad k \in \{1, 2, \ldots, N_E\} \\
\tilde{\varphi} &= 0 \quad \text{on } \partial \Omega_{\text{ref. prong}} \\
\frac{\partial \tilde{\varphi}}{\partial n} &= 0 \quad \text{on } \partial \Omega_{\text{skin}}
\end{aligned}
\]

(4)

where \(\gamma(r, f, T) = \sigma(r, f, T) + j2\pi f \varepsilon(r, f)\) is the complex electrical admittivity profile at working frequency \(f\), \(\tilde{\phi}\), \(\tilde{j}_k\), \(\tilde{I}_k\) are the (phasor) electric scalar potential, current density and electrode currents, respectively.

It should be noticed that (4) is able to provide the thermal Ohmic power \(p_0\) as

\[
p_0(r, t) = \text{Real} \left\{ \frac{\mathbf{E}(r, \omega) \cdot \mathbf{J}(r, \omega)}{2} \right\} \quad \text{(where } \mathbf{E}(r, \omega) \text{ is the phasor electric field map and } \mathbf{J}(r, \omega) \text{ denotes the complex conjugate)}
\]

or, equivalently, \(p_0 = E^2 / (2 \sigma)\).

According to (1), the input variables are the complex current patterns \(\hat{I}\) applied to the internal and external electrodes (the DOF of the optimization problem), and the initial thermal map \(T_0(r)\). The time evolution of the model can be carried out by means of several approaches, including Runge-Kutta, predictor-corrector and so on. Here, a fixed time step algorithm has been used because it is able to provide a fast resolution of the optimization problem while guaranteeing the required accuracy.

The output of the model are the thermal map \(T(r, t)\) in the patient’s torso during RFA treatment. For the sake of simplicity, in the inverse problem, the forward operator will be synthetically indicated as \(\mathcal{F}(\hat{I}, T_0)\), where

\[
\mathcal{F}: \hat{I}, T_0 \rightarrow T(r, t)
\]

2.2 Finite Elements Model of the Torso

The 3D solution domain \(\Omega\) here used for the following numerical simulations is shown in Fig.1. It represents a rather simplified human torso, with a RFA probe inserted into the liver, and with multiple return pads. The main simplifying assumption is that only three different tissues were considered,
denoted in the following by “liver”, “torso” and “tumor”, characterized by different values of electromagnetic and thermal properties.

![Figure 1: (a) 3D Model of the torso Ω and (b) FE mesh used in the calculations, including the detail of the target region of the RFA treatment and of the internal probe prongs](image)

FE resolution of coupled problem (1)-(3) provides a discrete version $F$ of the operator $\mathcal{S}$ transforming the input couple $(\hat{I}, T_0)$ into the sequence of nodal temperatures array $T_n(t_k)$ with $k = 1, 2, \ldots, N_s$, where $N_s$ is the number of time steps.

### 2.3 Optimal Planning of RFA Treatment

Optimal planning of RFA treatment introduces a schedule in the heating of the various regions, compatible with the possible system dynamics, by modifying current amplitudes within values safe for human treatments (denoted from now on as Feasible Current Patterns, FCP). If a “desired” time varying temperature map $T^{\text{des}}(r,t)$ is assigned, the optimal planning problem can be formulated as that of finding the sequence of current pattern $\hat{I}_{\text{min}}$ such that the following error functional is minimized:

$$
\mathcal{E}(\hat{I}, T_0) = \left\| W(r,t) \left( T^{\text{des}}(r,t) - \mathcal{S}(\hat{I}, T_0) \right) \right\|^2_{\ell^2} + \alpha(t) \left\| \hat{A} \hat{I} \right\|^2_{\ell^2} \quad \forall t \geq 0
$$

(5)

where $W$ is a weighting function, $\alpha$ is regularization parameter and $\hat{A}$ is current pattern preconditioning matrix. The minimization of error functional (5) is subject to equations (1)-(3) as equality constraints and, as an additional constraint, $\hat{I}$ is taken as bounded by the set of FCP.

The minimization of (5) is intended for every time $t$. Thanks to the FE discretization of the forward operator, the problem (5) is expressed in a discretized version. If limiting to a single step in the thermal transient analysis, (5) can be written in a time and space discretized version:

$$
E \left( \hat{I}_{k}, T_0 \right) = \left\| \left( T^{\text{des}}_k - F_k \left( \hat{I}_{k}, T_0 \right) \right) \right\|^2_{\ell^2} + \alpha_k \left\| \hat{A} \hat{I}_{k} \right\|^2_{\ell^2} \quad \forall k \in \{0, \ldots, N\}
$$

(6)

where $k$ denotes the time step, $F_k \left( \hat{I}_{k}, T_0 \right)$ is the array of calculated temperature in the centroid of the elements mesh, $T^{\text{des}}_k$ is the corresponding "desired" value, $W_k$ is a weighing array and $\alpha_k$ is is the regularization parameter and $\hat{A}^T$ denotes transposition of an array. Notice that the weighting function/array accounts for the "importance" of the requirements in each subregions of the solution domain $\Omega$.

This approach represents a generalization of what presented in [6,7,8]: the introduction of the new independent prong currents provides improved effectiveness and flexibility to ablation process.
3. NUMERICAL EXAMPLE

In this section, the advantages of the two innovative aspects of RFA planning are demonstrated in the numerical case described in Sect. 1.2. First, independent current sources for the probe prongs (and for the return pads) are used to achieve a better temperature control in the tumor region, and then current waveforms in prongs and pads are optimized to improve spatial uniformity Ohmic power deposition.

The following geometrical data have been assumed. The tumor region \( \Omega_T \), Fig. 1, is a sphere of radius 15 mm, split into two “target” regions for RFA: \( \Omega_{T1} \) (radius 7 mm) and \( \Omega_{T2} \) (internal radius 7 mm, external radius 15 mm). Probe prongs have been modeled as cylindrical cavities (radius 1 mm, length 10 mm, centroids lying on a circle with 7.5 mm radius, equally spaced) encircled in Fig. 1b.

The FE mesh includes 33282 1\(^{\text{st}}\) order tetrahedral elements and 5740 nodes (see Fig. 1).

The first point is assessed by applying a single current between a single return pad (number 1 in Fig. 1(a)) and the whole set of prongs of the internal probe, assumed as equipotential (Single Current Optimization, SCO). Prong 0 (see Fig. 1(b)) is then grounded, while the remaining ones are supposed to be supplied each with a different current and zero current in the return pads has been imposed (Multiple Current Optimization, MCO). A time interval of 100 s has been considered for a first part of a possible RFA process, and subdivided in five subinterval 20 s each long, to actuate the step by step optimization process. The desired uniform temperature sequences \( T^{dR1} = [37.6, 38.2, 38.8, 39.4, 40] \) °C and \( T^{dR2} = [37.4, 37.8, 38.2, 38.6, 39] \) °C are assumed in (6) for the five time steps in the two controlled region, respectively.

The current to be optimized are sinusoidal functions characterized by the same frequency of \( f = 450 \) kHz and amplitudes and phases as degrees of freedom. For the sake of simplicity all the phases of external currents \( I_k \) are supposed vanishing and just the amplitudes have been included in the optimization search. In both cases, SCO and MCO, a deterministic approach has been applied (golden search and parabolic interpolation in SCO and pattern search in MCO); a starting value of 0.2 A for each unknown has been assumed at the first time step while the optimum found in the previous, has been fixed for starting in the following steps.

In addition, an absolute FCP constraint of 0.6 A has been imposed of the current in SCO, while the constraint for each currents in MCO has been fixed at 0.3 A. Finally, tissue properties and perfusion data used in the simulation are reported in Table I.

**Table I**: Main tissue properties and perfusion data used in the simulations

<table>
<thead>
<tr>
<th></th>
<th>( \sigma ) [S/m]</th>
<th>( \varepsilon_r )</th>
<th>( k ) [W/m °C]</th>
<th>( M ) [Kg/m(^3)]</th>
<th>( c ) [J/Kg°C]</th>
<th>( W_{bi} ) [kg/(sm(^3))]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>0.33</td>
<td>2700</td>
<td>0.51</td>
<td>1060</td>
<td>3600</td>
<td>16.7</td>
</tr>
<tr>
<td>Torso</td>
<td>0.25</td>
<td>2700</td>
<td>0.51</td>
<td>1060</td>
<td>3500</td>
<td>0.6</td>
</tr>
</tbody>
</table>

The results of the optimization are summarized in Fig. 2. In order to evaluate the effectiveness of the process, the uniformity of the temperature within the control domain should be considered; then the percentage of the volume in TR\(_1\) and TR\(_2\) characterized by a temperatures within the \( \pm 5\% \) of the desired values have been computed.

The results (Fig. 3) clearly show the effectiveness of the multiple independent prong currents in the ablation process.
Figure 2: Sequences of optimal current patterns for MCO (a) and for multipolar control of the internal probe SCO (b)

Figure 3: Comparison of percentage of volume where the temperature is close to the scheduled for TR1 (a) and for TR2 (b) respectively

This encouraging results is fully confirmed by the temperature map (see Fig. 4 for the profile at 100 s) in the both the domains TR1 and TR2 showing an occurrence of joining hot spots, especially in TR1, when a MCO is performed.

Figure 4: Local temperature maps in the case of equipotential internal prongs (a) and in the case of multipole control of the prongs (b) at t=100 s.

Finally, the marginal role of the external pads has been assessed. The effectiveness of the system with four internal prongs with and without an additional external pad have been compared: the difference is very small and the current in the external pad is, in any case, within the limit of a few percentages of the internal currents.
4. CONCLUSIONS

In this paper the effectiveness of prongs as independent electrodes in ablation procedures for clinical treatment of liver tumors has been discussed.

Of course, further careful analysis are required to assess the actual feasibility of the method. Particular attention should be devoted to the maximum differential currents that can be imposed on prongs and on maximum frequencies to prevent the capacitive couplings among prongs.

Unfortunately the important uncertainties affecting living tissue parameters and geometrical information of the organs, prevent to obtain results so reliable and accurate to be directly applied in a clinical RFA protocol. For this reason, the outcome of the simulations must be taken as a suggestion than as a prescription. However their value remain very important in looking for the most effective ablation treatments.

Some interesting global properties coming out of the study here performed are finally listed:

- the marked superiority of the multipolar control of the probe prongs in terms of ablation capability emerges, when compared to monopolar prongs excitation and use of external return pads. This is in full agreement with the previous theoretical and experimental investigations [1];
- the optimal way to concentrate Ohmic power density in the target regions is characterized by an almost symmetric system of AC currents, in phase opposition according with the space distribution, in such a way to exhibit a "strong" interaction among opposite prongs.

REFERENCES

INTELLIGENT COMPUTING IN OPTIMIZATION OF COUPLED PROBLEMS

T. BURCZYŃSKI *,†, A. DŁUGOSZ †

* Institute of Computational Mechanics and Engineering
Faculty of Mechanical Engineering
Silesian University of Technology
ul. Konarskiego 18a, 44-100 Gliwice, Poland
e-mail: tadeusz.burczynski@polsl.pl; adam.dlugosz@polsl.pl

† Institute of Computer Science
Faculty of Physics, Mathematics and Computer Science
Cracow University of Technology
ul. Warszawska 24, 31-155 Cracow, Poland

Key words: multiobjective optimization, evolutionary algorithm, multiphysics, actuators, MEMS, finite element method.

Abstract. This paper proposes a multiobjective optimization approach for designing selected coupled-field models. The optimization method is based on evolutionary algorithm (EA). Proposed technique overtakes one of the most popular multiobjective evolutionary algorithm NSGAI[5,6] on several benchmark and engineering problems. Coupling between electrical, thermal and mechanical fields is considered. Finite element method (FEM) is used to simulate direct coupled problems numerically. The software packages based on FEM are adapted to create the optimization system. Suitable interfaces between optimization algorithm and the FEM software are created. They use internal script languages embedded in preprocessors of the FEM systems. Different types of functionals are formulated on the basis of the results obtained from coupled-field analysis. Functionals depending on the volume of the structure are also proposed. Parametric NURBS curves are used to model some optimized structures. Numerical examples for bi-objective and three-objective optimization problems are presented.

1 INTRODUCTION

Modern techniques of multiobjective optimization [4,20,21] have great importance in designing process and structures. In order to perform optimization task, proper functionals have to be defined. For coupled field problems, functionals based on the different quantities, deriving from different multiphysics fields are often contradictory. Moreover for such problems objectives may be strongly multimodal. Efficient optimization requires the usage of techniques, which are resistant for stuck in local optima and allows to obtain solutions useful for designers. Application of bio-inspired optimization techniques such as Evolutionary Algorithms (EAs), Artificial Immune Systems (AISs), Particle Swarm Optimizers (PSOs) increases probability of finding global solution [14]. In comparison with to the other techniques Bio-inspired algorithms give a proper balance between two conflicting aspects needed in successful optimization: exploitation near the likely optimum and the exploration of
the search space. In many practical cases this attitude is the only choice. EAs are ideal candidates for finding the Pareto optimal solutions in multiobjective optimization, because they work on the population of potential solutions in each generation. Comparing to single objective optimization, in the multiobjective optimization the result is not one, but set of solutions. In order to obtain set of Pareto-optimal solutions with the use of scalarization techniques (weighting sum method, ε-constrained method, etc.) optimization task have to be performed many times. It could be inadequate and inefficient comparing to method which uses Pareto classifications of solutions.

The paper is devoted to the multiobjective optimization of different types of actuators. Such devices produce movement and force by transformation different types of energy. Numerical simulation requires solving appropriate type of type multiphysics problem.

2 FORMULATION OF THE MULTIOBJECTIVE OPTIMIZATION PROBLEM

The process of finding a vector of decision variables that satisfies some restrictions and optimizes the vector of functionals is called multiobjective optimization (MOO). A MOO problem is formulated as follows:

find the vector \( \mathbf{x} = [x_1, x_2, \ldots, x_n]^T \) which satisfies the \( m \) inequality constrains:

\[
g_i(\mathbf{x}) \geq 0 \quad i = 1, 2, \ldots, m
\]

(2.1)

and the \( p \) equality constrains

\[
h_i(\mathbf{x}) = 0 \quad i = 1, 2, \ldots, p
\]

(2.2)

which minimizes the vector of \( k \) objective functions:

\[
f(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})]^T
\]

(2.3)

where \( n \) is the number of design variables, \( k \) is the number of objective functions.

On each design variable box constraints are imposed

\[
x_i^L \leq x_i \leq x_i^R
\]

(2.4)

where \( x_i^L \) and \( x_i^R \) are minimum and maximum acceptable values for the variable \( x_i \) respectively. Multiobjective optimization deals with multiple conflicting objectives and usually the optimal solution for one of the objectives is not necessarily the optimum for any of the other objectives. For such a case, instead of one optimal solution like in single-objective optimization problem, many solutions are incomparable and simultaneously optimal. These solutions are called Pareto-optimal ones.

3 OPTIMIZATION METHOD

3.1 MOOPTIM algorithm

In-house implementation of the MultiObjective OPTIMization tool based on evolutionary algorithm (MOOPTIM) is used for optimization [3,7,8,9]. Pseudo code of the algorithm is presented in Fig.3.1. In the initialization step, besides determining all settings of the algorithm, populations \( Q_i \) and \( P_i \) (of the same size) are generated and the fitness functions are
evaluated for population $Q_i$. In the main loop, after evaluation fitness of functions for $P_i$, populations $Q_i$ and $P_i$ are combined into set $R_i$. Next selection is performed on the set $R_i$. The individuals from the population $R_i$ are selected to $P_{i+1}$ on the basis of the nondomination level and the crowding coefficient. Individuals from $P_{i+1}$ are copied to $Q_{i+1}$ and then evolutionary operators change the population $P_{i+1}$.

**MOOPTIM algorithm**

```
begin
  1  i←0
  2  randomly generate population $Q_i$
  3  evaluate objective functions for $Q_i$
  4  randomly generate population $P_i$
  5  while (not termination condition) do
      6    evaluate objective functions for $P_i$
      7    join population $Q_i$ and $P_i$ ($R_i=Q_i + P_i$)
      8    use selection (choose $P_{i+1}$ from $R_i$)
      9    copy $P_{i+1}$ to $Q_i$
     10    apply evolutionary operators for $P_{i+1}$
     11  i←i+1
end
```

*Figure 3.1: Pseudo code of the MOOPTIM*

The proposed algorithm is similar to NSGAII. It uses nondominated sorting procedure for classification of the individuals in population and a crowding coefficient to preserve diversity in the population [6]. The main difference between MOOPTIM and NSGA II is based on changes in selection mechanism and the application of different evolutionary operators. The proposed implementation has more evolutionary operators comparing to the NSGAII. Two types of mutation (uniform and Gaussian once) and two types of crossover operators (simple and arithmetical) are used.

It should be emphasized that Gaussian mutation has significant influence on the effectiveness of searching by the algorithm. This operator requires an extra parameter (besides mutation probability) in the form of the mutation range (from 0 to 1). It was observed that higher values of the probability and range of this operator improved the convergence of the algorithm, especially for more difficult tasks.

The other difference between these algorithms is related to the formation of population $P_{i+1}$. There is no binary tournament selection operator in MOOPTIM, but individuals are selected on the basis of nondomination level and crowding coefficient.

MOOPTIM was tested on several benchmarks (SCH, ZDT1, ZDT2, ZDT3, ZDT4, ZDT6, CONSTR, SRN, TNK) and also on some engineering problems. The results obtained using MOOPTIM in most cases are better in comparison with the results obtained by means of NSGA-II. Detailed comparison between MOOPTIM and NSGA-II can be found in [7]. Functionals defining for engineering coupled problems, which is solved by using FEM, are usually strongly multimodal. Ability of finding global solutions by optimization algorithm for such problems is essential.
3.2 Fitness functions evaluation

In the initialization part of the algorithm and in the main loop for each individual, fitness functions have to be calculated. Chromosomes genes represent the design variables in optimization tasks. The box constraints are imposed on each gene. The design variables are used for preparation of the geometry of the structure. That can be specific dimensions of some parts of the model (eq. length, width, distances) or control points of the parametric curves. There exist a lot of methods for geometry modelling. In the proposed approach NURBS (Non-Uniform Rational B-Spline) curves are used to model the geometry of some structures [18]. By using this curves in the preparation of the model it is possible to reduce the total number of design variables and to obtain regular smooth shapes of the structures.

Functionals are formulated on the basis of results obtained from coupled field analysis. As mentioned before boundary-value problems are solved by means of FEM [2,13,19]. To solve considered coupled problems FEM Ansys Multiphysics and MSC.Mentat/Marc software packages are used [1,17]. FEM software packages has been joined with the algorithm MOOPTIM. In order to combine optimization algorithm and FEM packages proper interface codes are created. Majority of FEM software packages consist of pre- and post-processors (usually used for preparation of the geometry, generation of the finite element mesh, introducing boundary conditions, etc.), as well as solvers used for numerical simulation of a particular physical phenomena. Preprocessors and solvers MSC.Mentat/Marc and Ansys Muliphysics are used to create the optimization system. In-house C++ and script languages codes implemented in preprocessors of the FEM packages. APDL (Ansys Parametric Design Language) is used for Ansys Multiphysics, while Python is used for MSC.Mentat software.

Fig.3.2 shows procedure of fitness functions evaluation. Communication between MOOPTIM algorithm, preprocessors, own procedures and FEM solvers is performed via files., the geometry of the optimized structure is generated on the basis of chromosome genes. In the next steps the finite element mesh is generated, the boundary conditions are applied and others settings of the analysis are defined. All these steps are performed with the use of an own in-house software and codes written in script languages (Python and APDL). After solving coupled analysis, the values of functionals are calculated on the basis of output files generated by MSC.Marc and Ansys. It should be underlined that efficient calculation in this step is critical taking into account the time of the optimization. Such approach allows avoiding external meshing procedure. Preprocesors are run in the batch mode, so the time necessary for fitness functions calculation is relatively short.

Figure 3.1: Procedure of fitness functions evaluation
4 CONSIDERED COUPLED PROBLEMS

Three types of boundary value problems are considered in the paper: piezoelectricity, electrostatic-mechanical analysis and electrical-thermal-mechanical analysis [2,12,15,16]. These problems are described by the appropriate electrical, heat conduction and elastostatic partial differential equations:

\[ \phi_{\mu} - \rho / \varepsilon_0 = 0 \]  
\[ \lambda_p T = 0 \]  
\[ \mu u_{i,\mu} + (\mu + \lambda) u_{i,\mu} + b_i = 0 \]

where: \( \phi \) electric potential, \( \rho \) is electric charge density, \( \varepsilon_0 \) is permittivity of free space (electric constant), \( \lambda_p \) is thermal conductivity, \( T \) is temperature, \( Q \) is internal heat source, \( u \) is displacement, \( b \) is body force per unit volume, \( \mu, \lambda \) are Lamé constants.

The equations with arbitrary geometries have to be fulfilled with appropriate boundary conditions. Fig.4.1 shows general form of the considered elastic body under electrical, thermal and mechanical loads.

\[ \phi_{\mu} = \phi_{\mu}^0 \quad \omega = \omega_t \]  
\[ T = T_i \quad q = q_i \quad u = u_i \]  
\[ \lambda_p T = \alpha (T - T^o) \]  
\[ \mu u_{i,\mu} + (\mu + \lambda) u_{i,\mu} + b_i = 0 \]

where \( \phi_{\mu}^0, \omega, u_i, T_i, q_i, \alpha, T^o \) is known electric potential, electric charge, displacements, tractions, temperatures, heat fluxes, heat conduction coefficient and ambient temperature respectively.

Separate parts of the boundaries must fulfill the following relations:

\[ \Gamma = \Gamma_{\phi} \cup \Gamma_{\omega} = \Gamma_{\phi} \cup \Gamma_{\omega} = \Gamma_{T} \cup \Gamma_{q} \cup \Gamma_{u} \]  
\[ \Gamma_{\phi} \cap \Gamma_{\omega} = \emptyset, \quad \Gamma_{\omega} \cap \Gamma_{\omega} = \emptyset, \quad \Gamma_{T} \cap \Gamma_{q} \cap \Gamma_{u} = \emptyset \]
Electrostatic-mechanical and electrical-thermal-mechanical analysis are weakly coupled, so it requires solving electrical, thermal and mechanical analysis separately. Coupling is carried out by transferring loads between the considered analysis using staggered procedures.

Piezoelectricity couples electrical and mechanical fields. This problem is strongly coupled and requires solving following system of partial differential equations:

\[ C_{ijkl} u_{k,j} + e_{ij} \phi_j + b_j = 0 \]
\[ e_{kl} u_{k,j} - \varepsilon_{il} \phi_{li} = 0 \] (4.6)

The tensors \( C_{ijkl}, e_{ij}, \varepsilon_{il} \) denote elastic moduli, piezoelectric constants and dielectric constants respectively.

FEM is used to solve boundary-value problems in all cases. For the electrostatic-structural analysis following system of linear equation are solved:

\[ K_e V = I \]
\[ K_m u = F + F_E \] (4.7)

where: \( K_e \) is electrical conductivity matrix, \( K_m \) is stiffness matrix, \( V, I, u, F, F_E \) denotes nodal vector of voltage, current, displacements, mechanical and electrostatic forces respectively.

For the electrostatic-mechanical analysis first electrical problem is solved. On the basis of nodal results from the electrical analysis, electrostatic forces vector \( F_E \) is calculated. This vector is treated as additional forces in mechanical analysis. For the electrical-thermal-mechanical analysis following system of linear equation are solved:

\[ K_e V = I \]
\[ K_T T = Q + Q_E \]
\[ K_m u = F + F_T \] (4.8)

where: \( K_T \) is thermal conductivity matrix, \( T, Q, Q_E, F_T \) are nodal vector of temperatures, heat fluxes, heat fluxes due to current flow, force due to thermal strain vector respectively. For this problem three analyses (electrical, thermal and mechanical) are solved separately. The thermal and mechanical problems are coupled through thermal strain loads \( F_T \). Coupling between the electrical and thermal problems is done by heat generation due to the electrical flow \( Q_E \).

Matrix equations of static piezoelectricity can be expressed as follows:

\[
\begin{bmatrix}
K_{uu} & K_{up} \\
K_{pu} & K_{pp}
\end{bmatrix}
\begin{bmatrix}
u \\
\Phi
\end{bmatrix} =
\begin{bmatrix}
F_u \\
\rho_p
\end{bmatrix}
\] (4.9)

where \( K_{uu} \) is mechanical stiffness matrix, \( K_{up}, K_{pu} \) are piezoelectric stiffness matrices, \( K_{pp} \) is dielectric stiffness matrix, \( F_u \) is force vector and \( \rho_p \) is charge flux vector.

Piezoelectric and electrostatic-mechanical analysis are solved with the use of Ansys Multiphysics, whereas electrical-thermal-mechanical analysis with the use of MSC.Mentat/Marc software.
5 FORMULATION OF THE FUNCTIONALS

Algorithm MOOPTIM is applied to the shape optimization of different structures by the minimization or maximization of appropriate functionals. Different functionals based on the results derived from coupled field analyses are formulated. For the considered problems functionals can be calculated on the basis of nodal results of electrical, thermal and mechanical quantities. Four different functionals are proposed:

- The minimization of the volume of the structure
  \[ \min_{\Omega} f_1 = \int_{\Omega} d\Omega \]  
  \hspace{1cm} (5.1)

- The minimization of the maximal value of the equivalent stress
  \[ \min_{\Omega} f_2 = \max(\sigma_{eq}) \]  
  \hspace{1cm} (5.2)

- The maximization of the average of electrostatic force
  \[ \max_{\Omega} f_3 = \overline{F_{el}} \]  
  \hspace{1cm} (5.3)

- The maximization of the maximal value of vertical deflection of the structure at node \( i \)
  \[ \max_{\Omega} f_4 = \max(u_i) \]  
  \hspace{1cm} (5.4)

6 NUMERICAL EXAMPLES

MOOPTIM and NSGA II algorithms are used to solve considered optimization tasks. In all examples parameters of MOOPTIM are as follows: probabilities of arithmetic crossover, simple crossover and uniform mutation are set to 0.1, The probability of Gaussian mutation is 0.7 and range of Gaussian mutation is 0.5. For NSGAII the crossover probability is set to 0.9 and mutation probability is equal to 0.1 as suggested in the papers [5, 6]. All tests are run with the same size and number of generations for both algorithms.

6.1 Example 1 – electrostatic-mechanical coupling

A part of a comb drive actuator is considered. Electrostatic comb drives are frequently used for actuation in MEMS [10,11]. A pair one finger and stator of the actuator is modeled (Fig.6.1). Stator is fixed, whereas rotor is attached to the spring (spring constant \( k=2.83e-4N/\mu m \)) and allowed to move. The actuator is fabricated from polycrystalline silicon. Potential difference between rotor and stator is 4V. Electrostatic force generated by the actuator (besides potential difference) depends on shape of the finger, distance between rotor and stator and its bilateral overlapping \( x_0 \). Functional (5.4) is calculated for the selected 4 points of the distance \( x_0 \) equal to 20\( \mu m \), 40\( \mu m \), 60\( \mu m \), 80\( \mu m \) respectively (Fig.6.2a). Shape of the finger is modeled with the use of NURBS curve consist of 4 control points (Fig.6.2b). Geometry of the model is symmetrical along the horizontal axis. Vertical coordinates of this points are design variables. Box constraints imposed on each design variable are between -4\( \mu m \) and -2\( \mu m \) (assuming that value 0\( \mu m \) gives finger thickness equal to 10.0\( \mu m \) - see Fig.6.1).
Figure 6.1: Model of the finger in the combrive electrostatic actuator

Figure 6.2: a) Relation between electrostatic force and overlapping distance between rotor and stator for the straight shape of the finger, b) parametrization of the finger

Figure 6.3: Set of Pareto optimal solution for functionals a) (5.1) and (5.2), b) (5.1) and (5.3)

Table 6.1: Shape of the rotor, stress distribution, values of design variables and functionals for 3 points

<table>
<thead>
<tr>
<th>functionals</th>
<th>Point A</th>
<th>Point B</th>
<th>Point C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_2$ and $f_1$</td>
<td>$Z_1=1.48 \ \mu m; \ Z_2=-4.0 \ \mu m; \ Z_3=-1.91 \ \mu m; \ Z_4=-3.01 \ \mu m$</td>
<td>$Z_1=-4.0 \ \mu m; \ Z_2=-4.0 \ \mu m; \ Z_3=-3.06 \ \mu m; \ Z_4=-3.03 \ \mu m$</td>
<td>$Z_1=-4.0 \ \mu m; \ Z_2=-4.0 \ \mu m; \ Z_3=-4.0 \ \mu m; \ Z_4=-3.49 \ \mu m$</td>
</tr>
<tr>
<td>$f_2$ and $f_3$</td>
<td>$f_1=1.05e-5\ \text{MPa}; \ f_1=826.3\ \mu m^2$</td>
<td>$f_1=1.16e-5\ \text{MPa}; \ f_1=622.7\ \mu m^2$</td>
<td>$f_1=1.47e-5\ \text{MPa}; \ f_1=517.2\ \mu m^2$</td>
</tr>
<tr>
<td>$f_3$ and $f_2$</td>
<td>$Z_1=0.08 \ \mu m; \ Z_2=2.0 \ \mu m; \ Z_3=1.95 \ \mu m; \ Z_4=0.95 \ \mu m$</td>
<td>$Z_1=-3.06 \ \mu m; \ Z_2=0.94 \ \mu m; \ Z_3=0.44 \ \mu m; \ Z_4=1.4 \ \mu m$</td>
<td>$Z_1=-3.06 \ \mu m; \ Z_2=0.94 \ \mu m; \ Z_3=-3.74 \ \mu m; \ Z_4=1.40 \ \mu m$</td>
</tr>
<tr>
<td>$f_3$ and $f_1$</td>
<td>$f_1=4.34e-5\ \text{N}; \ f_1=1627.5\ \mu m^2$</td>
<td>$f_1=2.92e-5\ \text{N}; \ f_1=964.8\ \mu m^2$</td>
<td>$f_1=1.61e-5\ \text{N}; \ f_1=559.3\ \mu m^2$</td>
</tr>
</tbody>
</table>
The multiobjective optimization problem is solved for pairs of the functionals (5.1), (5.2) and (5.2), (5.4). Problems have been solved with the use of MOOPTIM and NSGA II algorithm. Fig. 6.3 shows set of obtained Pareto-optimal solution, whereas Table 6.1 contains optimal shapes and values of design variables for three selected points on the fronts.

6.2 Example 2 – electric-thermal-mechanical coupling

The model of microelectrothermal actuator is considered (Fig. 6.4). The actuator is fabricated from polycrystalline silicon. The deflection of the actuator occurs when the electrical potential difference is applied across two electrical pads (EP1 and EP2). It is possible due to material properties - high electrical resistivity and different thermal expansion between thin and wide arms.

Pads are fixed in all degrees of freedom, potential difference between pads are 5V. The length of the actuator is equal to 260 microns, while electrical pads are 20x20 microns wide. The multiobjective problem concerns determining the specified dimensions of the actuator shape, which minimize or maximize 3 functionals simultaneously (5.1), (5.2) and (5.4). For the functional (5.4) node at the right upper corner is considered. Six design variables are assumed (Fig.6.4).
Box constraints imposed on design variables $Z_1$, $Z_2$, $Z_3$ are between $1\mu m$ and $3\mu m$, for $Z_4$ between $12\mu m$ and $18\mu m$, for $Z_5$ between $30\mu m$ and $100\mu m$ and for $Z_6$ between $2\mu m$ and $8\mu m$. As in previous example problem has been solved with the use of algorithm MOOPTIM and NSGA II. Fig.6.5 presents a set of Pareto optimal solutions, whereas Table 6.2 contains final shapes and values of design variables for extreme points on the front.

**Table 6.2:** Shapes of the actuator and values of design variables for extreme points on the Pareto-front.

<table>
<thead>
<tr>
<th>functionals</th>
<th>minimum of $f_1$</th>
<th>minimum of $f_2$</th>
<th>maximum of $f_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$; $f_2$; $f_4$</td>
<td>$Z_1=1.0\mu m$; $Z_2=1.05\mu m$; $Z_3=1.0\mu m$; $Z_4=12.0\mu m$</td>
<td>$Z_1=1.24\mu m$; $Z_2=1.91\mu m$; $Z_3=1.01\mu m$; $Z_4=14.3\mu m$</td>
<td>$Z_1=1.0\mu m$; $Z_2=1.0\mu m$; $Z_3=100.0\mu m$; $Z_4=2.0\mu m$</td>
</tr>
</tbody>
</table>

6.3 Example 3 - piezoelectricity

The piezoelectric actuator presented in Fig.6.1 is considered. Actuator is made of piezoelectric ceramic material PZT-5H. Left side (segment AF) is fixed in all degrees of freedom. On the segments AF and BC electric potential $-1000V$ and $1000V$ are applied respectively.

Figure 6.6: Model of the piezoelectric actuator

The aim of the multiobjective optimization is to find optimal shape of the actuator by minimizing or maximizing functionals $(5.1)$, $(5.2)$ and $(5.4)$. Four design variables have been distinguished: $Z_1$ (vertical coordinate of point A) with limitation $-6.0 \div 0mm$, $Z_2$ (horizontal coordinate of point B) with limitation $1.0 \div 5.0mm$, $Z_3$ (vertical coordinate of point B) with limitation $-6.0 \div 0mm$ and $Z_4$ (horizontal coordinate of point C) with limitation $-5.0 \div 9.0mm$.

Figure 6.3: Set of Pareto optimal solution for functionals a) $(5.1)$ and $(5.2)$, b) $(5.1)$ and $(5.3)$
Optimization tasks are performed with the use of both algorithms (MOOPTIM and NSGA II) for the pairs of the functionals: (5.1), (5.2) and (5.2), (5.4). Fig. 6.7 shows set of obtained Pareto-optimal solution, whereas Table 6.3 contains optimal shapes and values of design variables for three selected points on the fronts.

Table 6.3: Shape of the actuator, stress distribution, values of design variables and functionals for the points

<table>
<thead>
<tr>
<th>functionals</th>
<th>Point A</th>
<th>Point B</th>
<th>Point C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_2$ and $f_1$</td>
<td>$Z_1=-0.47 \ \mu m; Z_2=4.83 \ \mu m; Z_3=2.06 \ \mu m; Z_4=7.46 \ \mu m$</td>
<td>$Z_1=0 \ \mu m; Z_2=3.02 \ \mu m; Z_3=-1.0 \ \mu m; Z_4=5.0 \ \mu m$</td>
<td>$Z_1=-0.02 \ \mu m; Z_2=1.06 \ \mu m; Z_3=-1.0 \ \mu m; Z_4=5.0 \ \mu m$</td>
</tr>
<tr>
<td></td>
<td>$f_2=5.35 \ \text{MPa}; f_1=24.99 \ \text{mm}^2$</td>
<td>$f_2=9.23 \ \text{MPa}; f_1=16.51 \ \text{mm}^2$</td>
<td>$f_2=30.69 \ \text{MPa}; f_1=15.54 \ \text{mm}^2$</td>
</tr>
<tr>
<td>$f_2$ and $f_4$</td>
<td>$Z_1=-0.27 \ \mu m; Z_2=4.34 \ \mu m; Z_3=-1.76 \ \mu m; Z_4=6.89 \ \mu m$</td>
<td>$Z_1=-0.09 \ \mu m; Z_2=1.96 \ \mu m; Z_3=-1.25 \ \mu m; Z_4=5.18 \ \mu m$</td>
<td>$Z_1=-0.01 \ \mu m; Z_2=1.00 \ \mu m; Z_3=-1.0 \ \mu m; Z_4=5.0 \ \mu m$</td>
</tr>
<tr>
<td></td>
<td>$f_2=5.85 \ \text{MPa}; f_1=-1.8 \ \text{mm}$</td>
<td>$f_2=13.16 \ \text{MPa}; f_1=-2.51 \ \text{mm}$</td>
<td>$f_2=32.64 \ \text{MPa}; f_1=-3.1 \ \text{mm}$</td>
</tr>
</tbody>
</table>

6 CONCLUSIONS

The designing of the real structures, especially for multiphysics problems, when different criteria are taken into account, belongs to difficult optimization tasks. Intuitive solutions can be found only for some cases (simple geometry, simple boundary conditions, low number of design variables, etc.). The application of classical methods based on gradient algorithms may be also unsuccessful due to the multimodality of functionals for real problems.

The direct problems concern coupling between mechanical, thermal and electrical fields. For solving direct coupled problems MSC.Mentat/Marc and Ansys Multiphysics are applied. Interfaces between MOOPTIM and FEM software’s are prepared with the use of in-house software and codes written in C++ and script languages.

MOOPTIM algorithm has been proposed for multiobjective shape optimization of different structures in the paper. The algorithm is in-house implementation based on evolutionary algorithms. It belongs to global optimization methods. The efficiency of proposed optimization method was tested on several benchmark problems. In most cases MOOPTIM gives better results comparing to the NSGA-II algorithm [7,9]. For problems considered in the paper MOOPTIM gives better results also. Set of Pareto-optimal solutions are: more converged, more uniformly distributed and wider comparing to NSGA II.

Different functionals based on the results derived from coupled field analyses are formulated. The application of the FEM software requires evaluation in several steps for each single solution (the modification of the geometry, creating finite element mesh, etc.). It can be very-time-consuming task, especially for more complicated geometries. Moreover the solution of the coupled problems is more time-consuming in comparison to the single-field problems. In order to reduce the time of the computation application of approximate surrogate model or/and parallelization of the fitness function evaluations are needed.
REFERENCES

COUPLING GENERALISED-\(\alpha\) METHODS: ANALYSIS, ADAPTIVITY, AND NUMERICS

JOACHIM RANG∗

∗Institute of Scientific Computing
TU Braunschweig
Hans-Sommer Str. 65, 38106 Braunschweig, Germany
e-mail: j.rang@tu-bs, web page: http://www.wire.tu-bs.de

Key words: Coupled Problems, generalised-\(\alpha\) methods

Abstract. In this article we consider the generalised-\(\alpha\) methods, make an analysis of the methods and apply them to a coupled model problem. A new adaptive timestep control is presented.

1 INTRODUCTION

Coupled problems appear in different research areas. One common example is the interaction of structure and fluid [DR08], e.g. the numerical simulation of offshore wind turbines, see [MM04], or of biomechanical processes. Coupled problems consist of two or more different physical problems which are in general space and time dependent. The discretisation in space leads to a high dimensional system of ordinary differential equations (ODEs) or differential algebraic equations (DAEs). The computation of the numerical solution needs the simultaneous solution of the strong coupled equations of each problem. But often for each subproblem different discretisation schemes are used. In the case of fluid-structure interaction the fluid is often discretised with Finite Volumes, and the structure with Finite Elements. To build a monolithic solver [RB00] it is often difficult to find a cost free available software system which processes different discretisation methods for different problem classes.

This is one reason to use a modular approach and partitioned methods [RB00, FP80, MW01, PFL95, MS02, MNS06], i.e. the subproblems are solved by different codes which communicate with each other. The communication between the solvers can be realised with the help of the Component Template Library (CTL), i.e. the solvers are transformed into software components and are controlled with an independent central unit. In [RSM09] the CTL is used to solve FSI problems.

In this paper we consider the generalised-\(\alpha\) methods, which are introduced for first order ODEs in [JWH00], and for second order ODEs in [CH93]. The generalised-\(\alpha\) methods
are in general of second order and allow the damping of high frequencies, which can be controlled by certain parameters. An analysis for first order problems can be found in [DP03]. In the case of second order ODEs many papers can be found, which analyse the generalised-α method, for example [EBB02]. It is well known that the generalised-α method for first order problems can be formulated as onestep and as multistep method. In the case of second order ODEs this statement is only true, if the ODE is linear in the first derivative (see [EBB02]). For both classes of multistep methods second order can be achieved if a further condition is satisfied. Together with the stability conditions (see [EBB02]) a robust and effective class of methods is obtained. If these parameter sets are used for onestep methods theoretically only first order can be reached. But the error constant is very small so that the observed numerical order of convergence is two. Moreover, in our experience the onestep versions obtain better results than the multistep versions.

In this paper we apply the generalised-α methods for first and second order ODEs on a damped mass system (see [JDP10]). In this paper we show that it is possible to couple the multistep versions of the generalised-α method. The numerical results are a little better than those of onestep versions. But the coupling of onestep methods has the advantage that it allows to easily compute adaptive timestep sizes, which is introduced in this paper, too.

This paper is structured as follows: First we introduce the generalised-α methods for first and second order ODEs. A short analysis about convergency and stability is given. Then we apply both generalised-α methods on the damped mass spring system of [JDP10] and analyse the linear systems.

2 THE GENERALISED-α METHOD FOR 1ST ORDER ODES

In the following we consider the ODE

\[ \dot{u} = f(t, u), \quad u(0) = u_0. \]  

(1)

The numerical solution of (1) is determined by the generalised-α method, which is given by the formulas (see [JWH00, DP03])

\[ \dot{u}_{n+\alpha_m} = f(t_{n+\alpha_f}, u_{n+\alpha_f}), \]  

(2)

\[ u_{n+1} = u_n + \tau \dot{u}_n + \tau \gamma (\dot{u}_{n+1} - \dot{u}_n), \]  

(3)

\[ \dot{u}_{n+\alpha_m} = \dot{u}_n + \alpha_m (\dot{u}_{n+1} - \dot{u}_n), \]  

(4)

\[ u_{n+\alpha_f} = u_n + \alpha_f (u_{n+1} - u_n). \]  

(5)

It is well known that the generalised-α method can be formulated as onestep and as twostep methods.
2.1 The formulation as onestep method and its analysis

First we manipulate the formulas (2)–(5) to obtain a non-linear system consisting of two decoupled equations. To abbreviate we define $f_{n+\alpha_f} := f(t_{n+\alpha_f}, u_{n+\alpha_f})$. A simple calculation gives us

$$u_{n+1} = u_n + \tau \left(1 - \frac{\gamma}{\alpha_m}\right) \dot{u}_n + \tau \frac{\gamma}{\alpha_m} f_{n+\alpha_f},$$

(6)

and

$$\dot{u}_{n+1} = \frac{1}{\tau \gamma} (u_{n+1} - u_n - \tau (1 - \gamma) \dot{u}_n),$$

(7)

if $\alpha_m \neq 0$. We call the scheme (6), (7) the onestep generalised-\(\alpha\) method. The starting value value $\dot{u}_0$ can be computed from the ODE (1). Next we want to determine the order of consistency. For this the numerical solution $u_{n+1}$ can be expanded in a Taylor series as follows

$$u_{n+1} = u_n + \tau \dot{u}_n + \frac{\tau^2 \gamma}{\alpha_m} f_n + O(\tau^3).$$

For consistency of order 2 we get the condition $\frac{\gamma}{\alpha_m} = \frac{1}{2}$. Since $u_{n+1}$ depends on $\dot{u}_n$ (see equation (6)) we use equation (7) for expanding $\dot{u}_{n+1}$ in a Taylor series and get

$$\dot{u}_{n+1} = \dot{u}_n + \frac{\tau \gamma}{\alpha_m} \ddot{u}_n + O(\tau^2),$$

i. e. $\dot{u}_{n+1}$ is of order 1 if $\frac{\gamma}{\alpha_m} = 1$. Summarising our results we have consistency of order 2 if $\alpha_m = \alpha_f$ and $\gamma = 1/2$. The generalised-\(\alpha\) method is zero-stable if $\alpha_m > 1/2$. In other words our method is convergent if $\alpha_m > 1/2$.

2.2 Formulation as multistep method and its analysis

The generalised-\(\alpha\) method can be formulated as a two-step method as follows

$$u_{n+1} = \frac{2\alpha_m - 1}{\alpha_m} u_n - \frac{\alpha_m - 1}{\alpha_m} u_{n-1} + \frac{\tau (1 - \gamma)}{\alpha_m} f_{n-1+\alpha_f} + \frac{\tau \gamma}{\alpha_m} f_{n+\alpha_f}.$$  

(8)

For $\alpha_m = 3/2$, $\alpha_f = 1$, and $\gamma = 1$ we obtain the backward difference formula (BDF) from Gear (see [HW96]). Next we expand $u_{n+1}$ in a Taylor expansion and compare it with the exact solution. Then we have

$$u_{n+1} = u_n + \tau \dot{u}_n + \frac{\tau^2}{2} \frac{2\alpha_f - \alpha_m + 2\gamma - 1}{\alpha_m} \ddot{u}_n + O(\tau^3).$$

Comparing the Taylor expansions for $u(t_{n+1})$ and $u_{n+1}$ leads to the condition for accuracy of order 2

$$\gamma = \frac{1}{2} - \alpha_f + \alpha_m,$$

(9)
which is already known from [JWH00, EBB02, CH93]). The generalised-\(\alpha\) method in form (8) is convergent of order 2 if \(\alpha_m > 1/2\) and condition (9) holds. For stability reasons often the setting

\[
\alpha_f = \gamma = \frac{1}{1 + \rho_\infty}, \quad \alpha_m = \frac{3 - \rho_\infty}{2(1 + \rho_\infty)},
\]

is used (see [JWH00, DP03]). Note that the condition (9) is automatically satisfied. For \(\rho_\infty = 0\) we get the BDF-2 method.

3 THE GENERALISED-\(\alpha\) METHOD FOR SECOND ORDER ODES

3.1 Formulation as onestep method

In the following we consider the second order ODE

\[
\ddot{u} = f(t, u, \dot{u}), \quad u(0) = u_0, \dot{u}(0) = \dot{u}_0.
\]

The generalised-\(\alpha\) method can be written as

\[
\begin{align*}
u_{n+\alpha_f} &= \alpha_f u_{n+1} + (1 - \alpha_f) u_n, \\
\dot{u}_{n+\alpha_f} &= \alpha_f \dot{u}_{n+1} + (1 - \alpha_f) \dot{u}_n, \\
\ddot{u}_{n+\alpha_m} &= \alpha_m \ddot{u}_{n+1} + (1 - \alpha_m) \ddot{u}_n,
\end{align*}
\]

\[
\begin{align*}
u_{n+1} &= u_n + \tau \dot{u}_n + \tau^2 \left[ \frac{1}{2} - \beta \right] \dddot{u}_n + \beta \ddot{u}_{n+1} \\
\dot{u}_{n+1} &= \dot{u}_n + \tau \left[ (1 - \gamma) \dddot{u}_n + \gamma \ddot{u}_{n+1} \right] \\
\ddot{u}_{n+1} &= \ddot{u}_n + \tau \left[ (1 - \gamma) \dddot{u}_n + \gamma f_{n+\alpha_f} \right],
\end{align*}
\]

where \(t_{n+\alpha_f} = t_n + \tau \alpha_f\). To abbreviate we write

\[
f_{n+\alpha_f} := f(t_{n+\alpha_f}, \alpha_f u_{n+1} + (1 - \alpha_f) u_n, \alpha_f \dot{u}_{n+1} + (1 - \alpha_f) \dot{u}_n).
\]

First we determine the order of consistency and use equations (14) and (17) for manipulating (15). We obtain

\[
\begin{align*}
u_{n+1} &= u_n + \tau \dot{u}_n + \tau^2 \left[ \frac{1}{2} - \beta \right] \dddot{u}_n + \beta \ddot{f}_{n+\alpha_f}, \\
\dot{u}_{n+1} &= \dot{u}_n + \tau \left[ (1 - \gamma) \dddot{u}_n + \gamma \ddot{f}_{n+\alpha_f} \right], \\
\ddot{u}_{n+1} &= \frac{1}{\alpha_m} \left[ \dddot{u}_{n+\alpha_m} - (1 - \alpha_m) \dddot{u}_n \right] = \frac{1}{\alpha_m} \left[ f_{n+\alpha_f} - (1 - \alpha_m) \ddot{u}_n \right].
\end{align*}
\]
Next we expand these three expression into Taylor expansions and get
\[ u_{n+1} = u_n + \tau \dot{u}_n + \frac{1}{2} \tau^2 \ddot{u}_n + O(\tau^3), \]
\[ \dot{u}_{n+1} = \dot{u}_n + \tau \ddot{v}_{n+1} + \gamma \frac{\alpha}{\alpha_m} \tau^2 \ddot{u}_n + O(\tau^3), \]
\[ \ddot{u}_{n+1} = \ddot{u}_n + \tau \frac{\alpha}{\alpha_m} \ddot{u}_n + O(\tau^2). \]

It follows that the method is of order 2 if
\[ \frac{\alpha}{\alpha_m} = 1 \quad \text{and} \quad \gamma \frac{\alpha}{\alpha_m} = \frac{1}{2}. \]

This is the same result as in the previous section.

3.2 Formulation as multistep method

As in the previous section the generalised-\( \alpha \) method can be written as a multistep method if the ODE (11) is linear in \( \dot{u} \). Therefore we consider the problem as in [EBB02]
\[ M \ddot{u} + C \dot{u} + S(u) = F(t). \]

Then equation (17) reads as
\[ M \ddot{u}_{n+a_m} = F(t_{n+a_f}) - S(\alpha_f u_{n+1} + (1 - \alpha_f) u_n) - C(\alpha_f \dot{u}_{n+1} + (1 - \alpha_f) \dot{u}_n). \]

The generalised-\( \alpha \) method can be formulated as a multistep method with the help of (15), (16), and (22). These formulas are evaluated at time \( t_n, t_{n+1}, \) and \( t_{n+2} \) (see for example [EBB02]). Then we get
\[ \sum_{j=0}^{3} [M \alpha_j + \tau C \gamma_j] u_{n+j} + \tau^2 \sum_{j=0}^{2} \delta_j [S_{n+j+a_f} - F(t_{n+j+a_f})] = 0, \]
where
\[ \alpha_0 = 1 - \alpha_m, \quad \alpha_1 = 3\alpha_m - 2, \quad \alpha_2 = 1 - 3\alpha_m, \quad \alpha_3 = \alpha_m, \]
\[ \gamma_0 = (1 - \alpha_f)(\gamma - 1), \quad \gamma_1 = 1 - 2\alpha_f - 2\gamma + 3\gamma \alpha_f, \gamma_2 = \alpha_f + \gamma - 3\gamma \alpha_f, \quad \gamma_3 = \alpha_f \gamma, \]
\[ \delta_0 = \frac{1}{2} + \beta - \gamma, \quad \delta_1 = \frac{1}{2} - 2\beta + \gamma, \quad \delta_2 = \beta, \]
and
\[ F_{n+j-a_f} = F(\alpha_f t_{n+j+1} + (1 - \alpha_f) t_{n+j}) = F(t_{n+j} + \alpha_f \tau) \]
\[ S_{n+j+a_f} = \alpha_f S(u_{n+j+1}) + (1 - \alpha_f) S(u_{n+j}). \]
The method has consistency order 2 if \( \gamma = \frac{1}{2} + \alpha_m - \alpha_f \). It is zero-stable and convergent if \( \alpha_m \geq 1/2, \alpha_f \leq 1/2 \) and \( \gamma \leq 1/2 \) (see [EBB02]). For stability reasons often the setting

\[
\beta = \frac{(1 + \alpha_m - \alpha_f)^2}{4}, \alpha_f = \frac{1}{1 + \rho_\infty}, \alpha_m = \frac{2 - \rho_\infty}{1 + \rho_\infty}
\]

is used (see [CH93, EBB02, JDP10]).

4 Coupling of generalised-\( \alpha \) methods

In the following we consider the model problem

\[
\ddot{u} + 2\xi \omega \dot{u} + \omega^2 u = 0, \quad u(0) = u_0, \dot{u}(0) = v_0,
\]

where \( \xi \) is a given damping factor and \( \omega \) a given frequency.

4.1 The onestep methods

As in [JDP10] we introduce a partitioning of the problem as follows

\[
(1 - \alpha)\ddot{u}^s + \omega^2 u^s = -\alpha \ddot{u}^f - 2\xi \omega \dot{u}^f, \tag{24}
\]

where the left-hand side of (24) is integrated with the generalised-\( \alpha \) for second order problems and the the right-hand side of (24) is integrated with the generalised-\( \alpha \) for first order problems. In the case of onestep formulation we have

\[
(1 - \alpha)\ddot{u}_{n+1}^s + \omega^2 u_{n+1}^s = -\alpha \ddot{u}_{n+1}^f - \xi \omega \dot{u}_{n+1}^f
\]

Quantities \( \dot{u}_{n+1}^f \) and \( \dot{u}_n^f \) represent the interface velocities at time \( t_{n+1} \) and \( t_n \) (see [JDP10]).

We insert the formulas for \( \ddot{u}_{n+1}^s, \ddot{u}_{n+1}^f, \) and \( \ddot{u}_{n+1}^f \) and get

\[
(1 - \alpha) \left[ \alpha_m \ddot{u}_{n+1}^s + (1 - \alpha_m) \ddot{u}_n^s \right] + \omega^2 \left[ \alpha_f u_{n+1}^s + (1 - \alpha_f) u_n^s \right] = -\alpha \left[ \alpha_m \ddot{u}_{n+1}^f + (1 - \alpha_m) \ddot{u}_n^f \right] - 2\xi \omega \left[ \alpha_f \ddot{u}_{n+1}^f + (1 - \alpha_f) \ddot{u}_n^f \right].
\]

Moreover we have

\[
\dot{u}_{n+1}^s = u_n^s + \tau \ddot{u}_n^s + \tau^2 \left[ \frac{1}{2} - \beta^s \right] \ddot{u}_n^s + \beta_s \dot{u}_{n+1}^s,
\]

\[
\dot{u}_{n+1}^f = \dot{u}_n^f + \tau \left[ \gamma \ddot{u}_{n+1}^f + (1 - \gamma) \ddot{u}_n^f \right],
\]

\[
\dot{u}_{n+1}^f = \dot{u}_n^f + \tau \left[ \gamma \ddot{u}_{n+1}^f + (1 - \gamma) \ddot{u}_n^f \right].
\]
Since the left- and right-hand side of (24) are taken at different times we set

\[
F_{n+\alpha_f}^s = (1 - \alpha)\ddot{u}_{n+\alpha_m}^s + \omega^2 u_{n+\alpha_m}^s, \\
F_{n+\alpha_f}^f = -\alpha \ddot{u}_{n+\alpha_m}^f - 2\xi \omega \dot{u}_{n+\alpha_m}^f,
\]

where \(F_{n+\alpha_f}^s = \alpha_f F_{n+1} + (1 - \alpha_f) F_n\) and \(F_{n+\alpha_f}^f = \alpha_f F_{n+1} + (1 - \alpha_f) F_n\). It then follows

\[
\alpha_f^s F_{n+1} + (1 - \alpha_f^s) F_n = (1 - \alpha)\ddot{u}_{n+\alpha_m}^s + \omega^2 u_{n+\alpha_m}^s, \\
\alpha_f^f F_{n+1} + (1 - \alpha_f^f) F_n = -\alpha \ddot{u}_{n+\alpha_m}^f - 2\xi \omega \dot{u}_{n+\alpha_m}^f.
\]

Then our problem reads as

\[
(1 - \alpha) \left[ \alpha_m^s \ddot{u}_{n+1}^s + (1 - \alpha_m^s) \ddot{u}_n^s \right] + \omega^2 \left[ \alpha_m^f \ddot{u}_{n+1}^f + (1 - \alpha_m^f) u_n^s \right] = \alpha_f^s F_{n+1} + (1 - \alpha_f^s) F_n, \\
\alpha_f^f F_{n+1} + (1 - \alpha_f^f) F_n = -\alpha \left[ \alpha_m^f \ddot{u}_{n+1}^f + (1 - \alpha_m^f) u_n^f \right] \\
- 2\xi \omega \left[ \alpha_f^f \ddot{u}_{n+1}^f + (1 - \alpha_f^f) \ddot{u}_n^f \right].
\]

\[
u_{n+1} = A_1^{-1} A_2 \nu_n
\]

with

\[
A_1 = 
\begin{pmatrix}
\omega^2 \alpha_f^s & 0 & (1-\alpha)\alpha_m^s \tau^2 & 0 & -\alpha_f^s \\
0 & 2\xi \omega \alpha_f^f & 0 & \alpha_m^f \alpha_f^f \tau^2 & 0 \\
1 & 0 & -\beta^s & 0 & 0 \\
0 & \frac{1}{\tau} & -\frac{\alpha_f^f}{\tau} & 0 & 0 \\
0 & \frac{1}{\tau} & -\frac{\alpha_f^f}{\tau} & 0 & 0
\end{pmatrix},
\]

\[
A_2 = 
\begin{pmatrix}
-\omega^2 (1 - \alpha_f^s) & 0 & (1-\alpha)(1-\alpha_m^s) \tau^2 & 0 & 1-\alpha_f^s \\
0 & -2\xi \omega (1 - \alpha_f^f) & 0 & -\alpha (1-\alpha_m^f) \tau^2 & 0 \\
1 & 1 & \frac{1}{2} - \beta^s & 0 & 0 \\
0 & \frac{1}{\tau} & 0 & \frac{1-\gamma_f}{\tau} & 0 \\
0 & \frac{1}{\tau} & 0 & \frac{1-\gamma_f}{\tau} & 0
\end{pmatrix},
\]

\[
u_n = (u_n^s, \tau \ddot{u}_n^f, \tau^2 \ddot{u}_n^s, \tau^2 \ddot{u}_n^f, \tau^2 F_n)^\top.
\]
Using a computer algebra package we can compute $u_{n+1}$ and expand it into a Taylor series. Then the local error reads $\delta_\tau = O(\tau^2)$. We test the implementation on two different settings, which are considered in [JDP10], too. In setting 1 we set $\omega = 1$, $\xi = 0.001$, and $\alpha = 0.5$, in setting 2 we have $\omega = 1$, $\xi = 0.01$, and $\alpha = 0.8$. The parameters $\rho_f$ and $\rho_s$ are taken as variables and we test the above method with each combination of $\rho_f$ and $\rho_s$. The numerical results are presented in Figure 1. We see that the coupled method produces accurate results. It can be observed that the monolithic approach does not always give the best results. The best result is obtained with setting $\rho_f = \rho_s = 1$.

4.2 The multistep methods

In the next step we couple the multistep versions. First we have

$$u_{n+1} = \frac{2\alpha_m - 1}{\alpha_m} u_n - \frac{\alpha_m - 1}{\alpha_m} u_{n-1} + \frac{\tau(1 - \gamma)}{\alpha_m} f_{n-1+\alpha_f} + \frac{\tau\gamma}{\alpha_m} f_{n+\alpha_f}.$$ 

for the first order problem and

$$\sum_{j=0}^{3} [M_{\alpha_j} + \tau C_{\gamma_j}] u_{n+j} + \tau^2 \sum_{j=0}^{2} \delta_j [S_{n+j+\alpha_f} - F(t_{n+j+\alpha_f})] = 0.$$ 

for the second order problem. For coupling $u_{n+1}$ and $\dot{u}_{n+1}$ we use the difference quotient

$$11u_{n+3} - 18u_{n+2} + 9u_{n+1} - 2u_n = 6\tau \dot{u}_{n+3}.$$ 

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{figure1a}
\includegraphics[width=0.4\textwidth]{figure1b}
\caption{Numerical errors of the coupled method: setting 1 (left) and setting 2 (right)}
\end{figure}
Finally we have a coupled system of three equations

\[
(1 - \alpha) \sum_{j=0}^{3} \alpha_j \mathbf{u}_{n+j}^f + \tau^2 \sum_{j=0}^{2} \delta_j \omega^2 \mathbf{u}_{n+j} + \mathbf{F}(t_{n+j}) = 0
\]

Using a computer algebra package we can compute \( u_{n+1} \) and expand it into a Taylor series. Then the local error reads \( \delta_r = O(\tau^2) \). As for the onestep methods the coupled method is applied on two different settings. Again, in setting 1 we set \( \omega = 1, \xi = 0.001, \) and \( \alpha = 0.5 \), and in setting 2 we have \( \omega = 1, \xi = 0.01, \) and \( \alpha = 0.8 \). Parameters \( \rho_f \) and \( \rho_s \) are taken as variables and we test the above method with each combination of \( \rho_f \) and \( \rho_s \). The numerical results are presented in Figure 2. We see that the coupled multistep methods produce more accurate results than the previous approach with the onestep methods. It can be observed that the monolithic approach does not always give the best results. The best result is obtained with setting \( \rho_f = \rho_s = 1 \).

5 Adaptness

In [Ran13] an adaptive timestep control for the generalised-\( \alpha \) method is introduced for onestep and for multistep versions. In this article only adaptivity for onestep methods is considered. With the help of the backward Euler method a second solution can be computed. Then the next timestep size \( \tau_{n+1} \) is proposed to be

\[
\tau_{n+1} = \rho \frac{\tau_n^{2}}{\tau_{n-1}} \left( \frac{TOL \cdot r_n}{r_{n+1}^2} \right)^{1/p}, \tag{26}
\]
where $\rho \in (0, 1]$ is a safety factor, $TOL > 0$ is a given tolerance, and $r_{n+1} := \|u_{n+1} - \hat{u}_{n+1}\|$. For details about the numerical error and the implementation of automatic steplength control, we refer to [HW96, Lan01]. For our coupled problem (25) the algorithm reads as follows:

- First compute the solution of (25). Then we have $u_{n+1}^s$ and $\dot{u}_{n+1}^I$.
- Compute $\ddot{u}_{n+1}$ by evaluating the model problem.
- Compute a second solution with the backward Euler method, i.e. $u_{n+1}^s = u_{n}^s + \tau \dot{u}_{n+1}^I$ and $\dot{u}_{n+1}^I = \ddot{u}_{n+1} + \tau \ddot{u}_{n+1}$.
- Compute the numerical error $r_{n+1}$ and approximate the new timestep length $\tau_{n+1}$ with (26).
- If the numerical error is smaller than the given tolerance the timestep is accepted otherwise it is rejected and has to recomputed with the new timestep length $\tau_{n+1}$.

As before we test our numerical method on the problem with settings 1 and 2. We choose different tolerances and compute the numerical errors with respect to the computing time (see Figure 3). It can be observed that our approach produces stable numerical results.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.pdf}
\caption{Numerical errors of the adaptive method}
\end{figure}
6 Summary and Outlook

In this article we introduced a new coupling scheme of generalised-\(\alpha\) methods, which use multistep formulation. As in the case of onestep methods order 2 is theoretically not reached. In the second part of the paper we developed an adaptive timestep control for the partitioned approach, which gives good numerical results.

In a future work this approach should be applied on other problems like, for example, FSI problems.

REFERENCES


FAST SOLVERS AND EFFICIENT NUMERICAL CFD TECHNIQUES FOR DYNAMIC POROUS MEDIA PROBLEMS

S. TUREK*, A. OBAID* AND B. MARKERT†

*Institute of Applied Mathematics (LS3)
Dortmund University of Technology
Vogelpothsweg 87, 44227 Dortmund, Germany
e-mail: {stefan.turek,aobaid}@math.tu-dortmund.de, web page:
http://www.mathematik.tu-dortmund.de/lsiii/cms/de/lehrstuhl3.html

†Institute of Applied Mechanics (CE)
University of Stuttgart
Pfaffenwaldring 7, 70569 Stuttgart, Germany
e-mail: bernd.markert@mechbau.uni-stuttgart.de - Web page:
http://www.mechbau.uni-stuttgart.de/ls2/index.en.html

Key words: Multigrid solver, Porous media dynamics, Coupled problem, Monolithic solution, Fully implicit, Finite element method, CFD

Abstract. We present a fully implicit, monolithic finite element solution scheme to efficiently solve the governing set of differential algebraic equations of incompressible poroelastodynamics. Thereby, we proceed from a two-dimensional, biphasic, saturated porous medium model with intrinsically coupled and incompressible solid and fluid constituents. Our approach, motivated by well-accepted CFD techniques and originally developed for the efficient simulation of incompressible flow problems, is characterized by the following aspects: (1) a special treatment of the algebraically coupled volume balance equation leading to a reduced form of the boundary conditions; (2) usage of a higher-order accurate mixed LBB-stable finite element pair with piecewise discontinuous pressure for the spatial discretization; (3) application of the fully implicit 2nd-order Crank-Nicolson scheme for the time discretization; (4) use of a special fast multigrid solver for the resulting discrete linear equation system. For the purpose of validation and to expose the merits and benefits of our new solution strategy to other established approaches, canonical one- and two-dimensional wave propagation problems are solved. Finally, a large-scale, dynamic soil-structure interaction problem serves to reveal the efficiency of the special multigrid solver in combination with the chosen finite element discretization.
1 Governing Equations

In the framework of the Theory of Porous Media (TPM) [4], we proceed from a continuum-mechanical description of a fluid-filled porous body consisting of a solid matrix saturated by a single pore fluid. Thereby, the binary aggregate is treated as a macroscopic mixture $\varphi (\alpha = S : \text{solid}; \alpha = F : \text{fluid})$, so that $\varphi = \varphi^S \cup \varphi^F$ at any macroscopic spatial point $x(t)$ at any time $t \in [t_0, T]$. The local composition of the biphasic continuum is described by volume fractions $n^\alpha(x, t):= dV^\alpha / dV \in (0, 1)$ of $\varphi$ (solidity; $n^S$ : porosity) defined as the ratios of the partial to the total volume elements of $\varphi$. Assuming fully saturated conditions, the saturation constraint obviously yields $\sum_\alpha n^\alpha = n^S + n^F = 1$. Closely related is the introduction of two density functions, namely an effective density $\rho^R(\mathbf{x}, t)$ and a partial density $\rho^\alpha(\mathbf{x}, t)$ relating the local mass of $\varphi^\alpha$ to the partial or the bulk volume element. The considered biphasic model excludes thermal effects as well as mass exchanges (inert $\varphi^\alpha$) and proceeds from intrinsically incompressible constituents ($\rho^R = \text{const}$). In particular, the arising purely mechanical, binary model with $\alpha = \{S, F\}$ is governed by the following constituent balance equations:

$$ \rho^S(\mathbf{v}_S)' = \text{div} \mathbf{T}^S - n^S \text{grad} p + \rho^S \mathbf{b} + \frac{(n^F)^2 \gamma^{FR}}{k^F} (\mathbf{v}_F - \mathbf{v}_S) \quad (1) $$

$$ \rho^F(\mathbf{v}_F)' + \rho^F(\text{grad} \mathbf{v}_F)(\mathbf{v}_F - \mathbf{v}_S) = \text{div} \mathbf{T}^F - n^F \text{grad} p + \rho^F \mathbf{b} - \frac{(n^F)^2 \gamma^{FR}}{k^F} (\mathbf{v}_F - \mathbf{v}_S) \quad (2) $$

$$ \text{grad} n^F \cdot \mathbf{v}_F + \text{grad} n^S \cdot \mathbf{v}_S + n^F \text{div} \mathbf{v}_F + n^S \text{div} \mathbf{v}_S = 0 \quad (3) $$

Here, div (·) is the divergence operator related to grad (·), $\mathbf{T}^\alpha = (\mathbf{T}^\alpha)^T$ is the symmetric extra stress assuming non-polar constituents, $\mathbf{b}$ is the mass-specific body force acting on the overall aggregate. $n^S$ and $n^F$ are assumed to be constant, which is acceptable for the small deformation case, such that the blue terms are dropped out. Furthermore, proceeding from a geometrically linear description, the (red) nonlinear convective term becomes negligible. To continue with linear PDEs, the pore fluid is assumed to be Newtonian and incompressible leading to the following relation:

$$ \text{div} \mathbf{T}^F = \nu \Delta \mathbf{v}_F \quad (4) $$

In spite of its negligible influence (cf. [6]) in all our performed numerical tests so far, this term containing the (small) fluid viscosity is nevertheless considered in our subsequent discretization and solution approaches, particularly in view of future large scale simulations which shall be able to involve all physically relevant effects. Restricting the presentation...
to the small strain regime, the solid extra stress is determined by the Hookean elasticity law

$$T_S^E = 2\mu^S \varepsilon_S + \lambda^S (\varepsilon_S \cdot I) I$$

with

$$\varepsilon_S = \frac{1}{2}(\text{grad} u_S + \text{grad}^T u_S)$$

as the geometrically linear solid strain tensor and $\mu^S, \lambda^S$ being the macroscopic Lamé constants of the porous solid matrix. Note that the chosen primary unknowns for this set of PDE are $u_S, v_F$ and $p$. Hence, $v_S(u_S)$ as well as $T^E(u_S), T^F(v_F), n^S(u_S)$ and $n^F(u_S)$ represent the secondary variables of the problem. Additionally, a reduction in the order of the PDE to order-one in time is achieved using

$$\frac{d u_S}{d t} = v_S$$

which eliminates the second time derivative of the solid displacement from (1), and allows the applicability of a wide range of fundamental time-stepping algorithms. This is a short presentation and for a more detailed discussion, the interested reader is referred to [2, 3, 7] and the citations therein.

2 Weak formulation and discretization in space and time

Our subsequent variational form of the $uvp$ approach, inspired by weak formulations that are typical in the CFD community for treating the incompressible Navier-Stokes equations, is created by multiplying (1)-(3) with the displacement test function $\delta u_S$, the velocity test function $\delta v_F$, the pressure test function $\delta p$, integrating over the whole domain $\Omega$ and performing partial integrations. Finally, we obtain the following weak form, which is similar to the standard one for porous media (see, for instance, page 1349 of [7]):

$$\int_{\Omega} \text{grad} \delta u_S : T_S^E \, dv - \int_{\Omega} \frac{(n^F)^2 \gamma^{FR}}{k^F} \delta u_S \cdot v_F \, dv + \int_{\Omega} n^S \text{div} \delta u_S \, p \, dv + \int_{\Gamma_{in}} \delta u_S \cdot \mathbf{t}^S \, da = 0$$

$$\int_{\Omega} \nu \text{grad} \delta v_F : \text{grad} v_F \, dv + \int_{\Omega} \frac{(n^F)^2 \gamma^{FR}}{k^F} \delta v_F \cdot v_F \, dv - \int_{\Omega} n^F \text{div} \delta v_F \, p \, dv - \int_{\Gamma_{in}} \delta v_F \cdot \mathbf{t}^F \, da = 0$$

$$\int_{\Omega} n^S \delta p \, \text{div} v_S \, dv + \int_{\Omega} n^F \delta p \, \text{div} v_F \, dv = 0$$

Here, the red-colored terms represent slight differences to [7], namely the additional fluid viscosity term and the natural shape of the weak form of the volume balance. Finally, we multiply (6) with $\delta u_S$ and integrate over $\Omega$:

$$\int_{\Omega} \delta u_S \cdot \{(u_S)'_S - v_S\} \, dv = 0$$
The boundary $\Gamma = \partial \Omega$ is divided into Dirichlet (essential) and Neumann (natural) boundaries, respectively, resulting in $\Gamma = \Gamma_u \cup \Gamma_s$ for the solid momentum balance and in $\Gamma = \Gamma_v \cup \Gamma_f$ for the fluid momentum balance, wherein the tractions are defined as:

\[
\mathbf{t}^s = \left( \mathbf{T}^s - n^s \mathbf{pI} \right) \cdot \mathbf{n}, \quad \mathbf{t}^f = \nu \frac{\partial \mathbf{v}_f}{\partial n} - n^f \mathbf{p}_n
\]

Keep in mind that due to the fact that the pressure (as Lagrange multiplier regarding the incompressibility constraint) provides typically less regularity than displacement and velocity, the pressure derivatives in the weak formulation have been eliminated by partial integration. For the same reason and as usual for the treatment of the incompressible Navier-Stokes equations, no integration by parts has been carried out in (9).

Using such a weak form, which avoids derivatives acting on the pressure functions, one can use standard FEM pairs for velocity/displacement and pressure as typical for incompressible flow problems, which are based on piecewise discontinuous pressure approximations (as shown in Figure 1), and the boundary conditions are imposed in a slightly modified way in which the fully drained boundaries are represented by the typical CFD ‘Do-nothing’ (see [5]) boundary condition ($\mathbf{t}^f = 0$) while the volume effluxes and values for the pressure as boundary conditions are not needed anymore. Therefore, we can choose the boundary conditions independently. As a candidate for LBB-stable Stokes elements, we apply in the following (2D) simulations the well-known (non-parametric) Q2/P1 element, that means biquadratic velocities and displacements and piecewise linear (discontinuous) pressure approximations (cf. [11]), which belongs currently to the ‘best’ FEM choices for incompressible flow problems with respect to efficiency, accuracy and robustness.

Next, based on the discretization with the introduced FEM spaces, equations (7)-(10) can be written in the following matrix-vector notation:

\[
\mathbf{M} \dot{\mathbf{y}} + \mathbf{K} \mathbf{y} = \mathbf{f}
\]

Figure 1: The discontinuous linear pressure element P1 (left) and the 9-node Lagrange biquadratic element Q2 (right) that we use for our $\mathbf{uvp}(3)$-TR method.
In more detail with mass and stiffness matrices and right hand side vectors, one obtains

\[
\begin{pmatrix}
M_{us} & 0 & 0 \\
0 & M_{us} & 0 \\
0 & 0 & M_{v_F} \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\dot{u}_S \\
\dot{v}_S \\
\dot{v}_F \\
\dot{p}
\end{pmatrix}
+ 
\begin{pmatrix}
0 & K_{v_S} & 0 & 0 \\
K_{u_S} & K_{u_S} & K_{u_S} & K_{u_S} \\
0 & K_{v_F} & K_{v_F} & K_{v_F} \\
0 & K_{p_{v_F}} & K_{p_{v_F}} & 0
\end{pmatrix}
\begin{pmatrix}
u_S \\
v_F \\
f \\
p
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
f_{u_s} + b_S \\
f_{v_F} + b_F \\
0
\end{pmatrix}
\tag{13}
\]

with the following matrices and right hand side vectors:

\[
K_{u_S} = \int_{\Omega} \text{grad} \delta u_S : T_E^S \, dv, \quad K_{v_S} = \int_{\Omega} \left(\frac{n^F}{k_F} \gamma_{FR}^F\right) \delta u_S \cdot v_S \, dv
\]

\[
K_{u_S} = -\int_{\Omega} \left(\frac{n^F}{k_F} \gamma_{FR}^F\right) \delta u_S \cdot v_F \, dv, \quad K_{u_S} = -\int_{\Omega} \n_S^S \delta u_S \cdot v_F \, dv
\]

\[
K_{v_S} = -\int_{\Omega} \delta u_S \cdot v_S \, dv, \quad K_{v_F} = -\int_{\Omega} \left(\frac{n^F}{k_F} \gamma_{FR}^F\right) \delta v_F \cdot v_S \, dv
\]

\[
K_{v_F} = \int_{\Omega} \nu \text{grad} \delta v_F : \text{grad} v_F \, dv + \int_{\Omega} \left(\frac{n^F}{k_F} \gamma_{FR}^F\right) \delta v_F \cdot v_F \, dv
\]

\[
K_{v_F} = -\int_{\Omega} n^F \delta v_F \, dv, \quad K_{p_F} = \int_{\Omega} n_S^S \delta p \, dv
\]

\[
M_{u_S} = \int_{\Omega} \delta u_S \cdot (u_S)' \, dv, \quad M_{v_F} = \int_{\Omega} \left\{n_S^S \rho_{SR}^F\right\} \delta v_F \cdot (v_F)' \, dv
\]

\[
f_{u_s} = \int_{\Gamma_s} \delta u_S \cdot \bar{t}^S \, da, \quad f_{v_F} = \int_{\Gamma_F} \delta v_F \cdot \bar{t}^F \, da
\]

\[
b_S = \int_{\Omega} \left\{n_S^S \rho_{SR}^F\right\} \delta u_S \cdot b \, dv, \quad b_F = \int_{\Omega} \left\{n_F^F \rho_{FR}^F\right\} \delta v_F \cdot b \, dv
\]

In the next step, regarding the time integration, equations (12) or (13) are treated in a monolithic implicit way leading to a fully coupled system. In our approach, we apply the standard one-step \(\theta\)-scheme to these systems, which leads to

\[
M_{y_{n+1}} + \frac{\gamma_{y_{n+1}}}{\Delta t} + \theta K_{y_{n+1}} = - \left(1 - \theta\right) K_{y_{n}} + \theta f_{n+1} + \left(1 - \theta\right)f_{n}.
\]

In the subsequent more detailed description, the red-colored continuity equation and the blue-colored pressure \(p\) as corresponding Lagrange multiplier are treated fully implicitly.
\[(\theta = 1) \text{ (cf. [11])}.
\]

\[
\begin{pmatrix}
    M_{v_S u_S} & \theta \Delta t K_{v_S v_S} & 0 & 0 \\
    \Delta t \theta K_{u_S u_S} & M_{u_S v_S} + \theta \Delta t K_{u_S v_S} & \theta \Delta t K_{u_S v_F} & K_{u_S p} \\
    0 & \theta \Delta t K_{v_F v_S} & M_{v_F v_F} + \theta \Delta t K_{v_F v_F} & K_{v_F p} \\
    0 & K_{p v_S} & K_{p v_F} & 0
\end{pmatrix}
\begin{pmatrix}
    u_S \\
    v_S \\
    v_F \\
    p
\end{pmatrix}
= 
\begin{pmatrix}
    u_S \\
    v_S \\
    v_F \\
    \bar{p}
\end{pmatrix}_{n+1}
\]

\[
\begin{pmatrix}
    M_{v_S u_S} & (\theta - 1) \Delta t K_{v_S v_S} & 0 & 0 \\
    (\theta - 1) \Delta t K_{u_S u_S} & M_{u_S v_S} + (\theta - 1) \Delta t K_{u_S v_S} & (\theta - 1) \Delta t K_{u_S v_F} & 0 \\
    0 & (\theta - 1) \Delta t K_{v_F v_S} & M_{v_F v_F} + (\theta - 1) \Delta t K_{v_F v_F} & 0 \\
    0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
    u_S \\
    v_S \\
    v_F \\
    \bar{p}
\end{pmatrix}_n
+ \theta \Delta t f_{n+1} + (1 - \theta) \Delta t f_n
\]

Note that the time steps \((\Delta t)\), supposed to be in front of the (blue) pressure matrices, are absorbed into \(\bar{p} = \Delta t p\), as usually done in CFD, leading to the following saddle-point problem with \(U^T = [u_S^T \ v_S^T \ v_F^T]\) that we solve for every time step:

\[
\begin{pmatrix}
    \hat{A} & B \\
    B^T & 0
\end{pmatrix}
\begin{pmatrix}
    U \\
    \bar{p}
\end{pmatrix}_{n+1}
= \text{RHS}
\]

After solving the above saddle-point systems, the pressure is scaled back using the relation \(p = \bar{p}/\Delta t\). Setting \(\theta = \frac{1}{2}\), we recover the second-order Crank-Nicolson scheme (in time), which is based on the well-known trapezoidal rule (TR). However, also fully L-stable 2nd order schemes like BDF(2) or Fractional-Step-Theta-schemes can be used in an analogous way.

### 3 Numerical validation

To validate and to evaluate our discussed formulations (which all have been realized in our in-house code FEATFLOW\(^1\)), two numerical examples taken from [7] are introduced and implemented in order to compare with well-established methods. Our \(uvp(3)\)-TR-Q2/P1 approach stands for the described monolithic solver for the \(uvp\) formulation based

\(^1\)http://www.featflow.de
on the weak forms (7)-(10) using the fully implicit Crank-Nicolson ($\theta = \frac{1}{2}$) time integration scheme as shown in (16) and the mixed finite element pairs $Q2/P1$ shown in Figure 1. The number $3$ in $uvp(3)$-TR-$Q2/P1$ is used to distinguish our solution algorithms from those in Table I in [7].

3.1 Results I: Saturated poroelastic column under harmonic load

In this example, the response of a homogeneous and isotropic, water-saturated, poroelastic column is analyzed under plane-strain, confined compression conditions. The geometry with boundary conditions and the loading path are illustrated in Figure 2 and the physical parameters, adopted from literature [1], are listed in Table 1.

![Figure 2: Left: Geometry, boundary conditions Center: isotropic Cartesian mesh Right: loading path of the dynamic confined compression of a saturated poroelastic column.](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
<th>SI unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Lamé constant of solid skeleton</td>
<td>$\mu^S$</td>
<td>$5.583 \times 10^8$</td>
<td>N/m$^2$</td>
</tr>
<tr>
<td>2nd Lamé constant of solid skeleton</td>
<td>$\lambda^S$</td>
<td>$8.375 \times 10^6$</td>
<td>N/m$^2$</td>
</tr>
<tr>
<td>Effective density of dense solid</td>
<td>$\rho^S$</td>
<td>2000</td>
<td>kg/m$^3$</td>
</tr>
<tr>
<td>Effective density of pore fluid</td>
<td>$\rho^F$</td>
<td>1000</td>
<td>kg/m$^3$</td>
</tr>
<tr>
<td>Initial volume fraction of solid</td>
<td>$n^S = n_0^S$</td>
<td>0.67</td>
<td>–</td>
</tr>
<tr>
<td>Darcy permeability</td>
<td>$k^F$</td>
<td>$10^{-2}, 10^{-5}, 10^{-10}$</td>
<td>m/s</td>
</tr>
<tr>
<td>Fluid dynamic viscosity (Figure 2 &amp; 4)</td>
<td>$\nu$</td>
<td>$10^{-3}$</td>
<td>Pa.s</td>
</tr>
<tr>
<td>Fluid dynamic viscosity (Figure 7)</td>
<td>$\nu$</td>
<td>$10^{-14}$</td>
<td>Pa.s</td>
</tr>
</tbody>
</table>
From Figure 3, we notice that, for this problem the proposed \( \mathbf{uvp}(3) \)-TR-Q2/P1 method provides almost the most accurate solutions at all selected heights.

<table>
<thead>
<tr>
<th>height</th>
<th>LL</th>
<th>QL</th>
<th>QQ</th>
<th>Q2/P1</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.7</td>
<td>0.12</td>
<td>0.019</td>
<td>0.0055</td>
<td>0.0003</td>
</tr>
<tr>
<td>9.8</td>
<td>0.13</td>
<td>0.0062</td>
<td>0.0193</td>
<td>0.0010</td>
</tr>
<tr>
<td>9.9</td>
<td>0.8</td>
<td>0.43</td>
<td>0.13</td>
<td>0.077</td>
</tr>
<tr>
<td>10</td>
<td>0.00</td>
<td>1.1</td>
<td>0.00</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Figure 3: Solid displacement (left) and absolute errors in \( \mu m \) (right) for the first half meter below the top surface for the isotropic Cartesian mesh (10 elem/m) (cf. Figure 2, center) for \( k^F = 10^{-5} \text{m/s} \) at \( t = 0.15 \text{s} \). (All the data except Q2/P1 are taken from [7]).

### 3.2 Two-dimensional wave propagation

In this second example, we study the 2D dynamical wave propagation in a rectangular symmetric domain under plane-strain conditions (Figure 4) as presented in [2]. The material parameters are the same as before (Table 1) and the ‘earthquake event’ is represented by the applied distributed impulse force

\[
f(t) = 10^5 \sin(25 \pi t) \left[1 - H(t - \tau)\right] \text{[N/m}^2\text{]} \tag{18}\]

with \( H(t - \tau) \) being the Heaviside step function and \( \tau = 0.04 \text{s} \). The symmetry of the problem can be exploited to reduce the problem size. However, the computation was performed on the full problem for the Q2/P1 approach and extremely low permeability of \( k^F = 10^{-10} \text{m/s} \), which further demonstrates the merits of the considered Q2/P1 approach. For this case both \( \mathbf{uvp}(2) \)-TB2-QQ and \( \mathbf{uvp}(2) \)-TB2-LL do not converge and the monolithic solution requires LBB-stable mixed FE formulations such as QL [7] and Q2/P1 element pairs. Based on the results shown in Figure 3, the direct comparison between the fully converged QL solutions and the fully converged Q2/P1 solutions (see Figure 5) reveals the less accurate displacement solution of the QL approach. In contrast to the TR-QL approach, our TR-Q2/P1 approach does not produce large pressure oscillations as seen in Figure 6. Such large oscillations are extremely reduced even for the trapezoidal rule (TR) by using a LBB stable element with equal order approximations of \( \mathbf{u}_S, \mathbf{v}_S \) and \( \mathbf{v}_F \) such as the Q2/P1 element.
\[ \bar{t}_y^{\text{S}} = f(t) \]

![Figure 4: Geometry, boundary conditions and mesh level 1 of the symmetric 2D wave propagation problem (left). Total number of elements and unknowns for the \( uv^p(3) \)-TR-Q2/P1 approach (right).](image)

<table>
<thead>
<tr>
<th>#Elements</th>
<th>#DOFs ( (\text{Q2/P1}) )</th>
<th>#DOFs ( (\text{QL}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>21-10</td>
<td>6048</td>
<td>3498</td>
</tr>
<tr>
<td>42-20</td>
<td>23430</td>
<td>13289</td>
</tr>
<tr>
<td>84-40</td>
<td>92214</td>
<td>51771</td>
</tr>
<tr>
<td>168-80</td>
<td>365862</td>
<td>102611</td>
</tr>
</tbody>
</table>

Figure 5: Pressure history at point B and displacement history at point A using \( uv^p(3) \)-TR-Q2/P1 and \( uv^p(2) \)-TB2-QL for \( k^F = 10^{-10} \text{m/s}, \Delta t = 10^{-3} \text{s}, t \in [0 0.2] \text{s} \) and mesh level 3 for Q2/P1 and QL.

![Figure 6: Pressure history at point B for \( k^F = 10^{-10} \text{m/s}, \Delta t = 10^{-3} \text{s} \) and mesh level 2 for \( uv^p(3) \)-TR-Q2/P1, \( uv^p(2) \)-TR-QL and \( uv^p(2) \)-TB2-QL](image)
4 Fast Multigrid solvers

During each time step, most of the elapsed CPU time is consumed by solving the corresponding linear systems in (17). Typically, by accuracy reasons which requires small mesh widths, the arising block systems are too large to be handled by direct solvers, such that iterative schemes have to be preferred. However, due to the nature of the involved partial differential equations, the condition numbers of the arising matrices typically scale with the problem size and are quite large, such that standard single-grid schemes, for instance Krylov-space methods like BICGSTAB or GMRES (cf. [8, 13]), are too slow. Moreover, due to the elliptic character of the incompressibility constraint, the choice of small time steps does not help since the condition numbers do not scale with the time step size due to the incompressibility. Therefore, an excellent alternative is to solve (17) via geometrical multigrid (MG) solvers (see [9, 10, 14]), which require a hierarchy of refined mesh levels and corresponding intergrid transfer operators, which are selected w.r.t. the chosen FEM spaces. What is special for the described saddle-point problems in (17) is the choice of the so-called ‘smoothing operator’, which in our case can be traced back to the early work by Vanka [12].

\[
\begin{align*}
\text{Structure} & \quad \bar{u}_S = 0 \\
& \quad u_{Sy} = 0 \\
& \quad u_{Sx} = 0 \\
& \quad v_{Fx} = 0 \\
\text{Soil} & \quad \bar{u}_S = 0 \\
& \quad u_{Sy} = 0 \\
& \quad u_{Sx} = 0 \\
& \quad v_{Fx} = 0 \\
\end{align*}
\]

Figure 7: Geometry of the 2D structure-soil problem with prescribed boundary conditions. The domain is composed of a structure, represented by an elastic block (4 × 2 m²) founded on an infinite domain of elastic soil replaced by fixed domain (size: 40 × 40 m²) with rigid boundaries. \( f(t) = 10^4 [1 - \cos(20 \pi t)] [1 - H(t - \tau)] \) [N/m²] with \( H(t - \tau) \) being the Heaviside step function and \( \tau = 0.1 \text{s} \) and the material properties are found in Table 1. The time period is set to \( t \in [0, 1.0] \text{s} \) and \( \Delta t = 2 \times 10^{-3} \).
The corresponding (basic) iterative schemes can be interpreted as block Gauß-Seidel methods applied to mixed formulations of saddle-point problems. In the following, we perform multigrid iterations of F-cycle type, applying a fixed number of pre- and postsmoothing steps for Cartesian (equidistant) grids of the problem depicted in Figure 7. Typically, we will show results for a sequence of consecutively refined meshes, which are constructed by connecting opposite midpoints of the corresponding coarser meshes, starting from a basic mesh on mesh level 1. Sample results, found in Table 2, demonstrate the very efficient convergence behavior for several parameter configurations and they illustrate the typical convergence behavior of multigrid solvers, namely to be more or less independent of the mesh size and the time step.

Table 2: Averaged number of iterations (iter.) and elapsed CPU time (CPU) per time step for the described multigrid solver for uvp(3)-TR-Q2/P1 for $t \in [0, 1.0]$ s and $\Delta t = 2$ ms for the Cartesian grid case.

<table>
<thead>
<tr>
<th>mesh level</th>
<th>no. elements</th>
<th>no. DOFs</th>
<th>$k^F = 10^{-2}$ m/s iter.</th>
<th>$k^F = 10^{-5}$ m/s iter.</th>
<th>$k^F = 10^{-10}$ m/s iter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 : (1 elem/m)</td>
<td>1608</td>
<td>44406</td>
<td>2</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>3 : (2 elem/m)</td>
<td>6432</td>
<td>175638</td>
<td>3</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>4 : (4 elem/m)</td>
<td>25728</td>
<td>698598</td>
<td>3</td>
<td>89</td>
<td>2</td>
</tr>
<tr>
<td>5 : (8 elem/m)</td>
<td>102912</td>
<td>2786502</td>
<td>3</td>
<td>353</td>
<td>3</td>
</tr>
<tr>
<td>6 : (16 elem/m)</td>
<td>411648</td>
<td>11130246</td>
<td>3</td>
<td>1446</td>
<td>3</td>
</tr>
</tbody>
</table>

5 Conclusion

Based on the comprehensive investigation of several test cases and the quantitative comparison with the results presented in [7], we recommend our fully implicit, monolithic approach using the $uvp(3)$-TR-Q2/P1 formulation in combination with the described special multigrid components. The proposed scheme does not only demonstrate excellent numerical results regarding accuracy and robustness, but is also less prone to stability issues (L-stability) of the time integrator even for coarser meshes.

Acknowledgements

We thank Michael Köster for the intensive support in the stage of implementing the presented model into FEATFLOW. Moreover, this work was partially supported by the German Research Foundation (DFG) under grant TU102/34-1.

References


NUMERICAL SIMULATION OF INDUCTIVE HEATING PROCESSES

T. GLEIM, D. KUHL AND B. SCHRÖDER

Institute of Mechanics and Dynamics
University of Kassel
Mönchebergstrae 7, 34109 Kassel, Germany
e-mail: kuhl@uni-kassel.de, web page: http://www.uni-kassel.de/fb14/mechanics

Key words: finite element method, heat equation, heat induction, Maxwell equations

Abstract. For product optimization regarding weight reduction, material properties have to be adapted efficiently. To achieve this, new compositions of materials can be created or the manufacturing process can be changed in a way that heterogeneous distributions of material properties are enabled. An example for such an improved process chain is the production of thermo-mechanically graded structures like shafts. The manufacturing method mainly consists of three stages. The first one is characterized by a local temperature increase of the workpiece due to inductive heating. In the second phase the workpiece is deformed and simultaneously cooled throughout the contact with the forming die. In the last step, however, a high pressured air stream is applied, leading to a partial cooling of the workpiece.

The inductive heating step is controlled by an alternating current inducing a high frequency magnetic field, which causes a temperature increase due to the resulting eddy currents. To analyse this process, the coupling between the electric and the magnetic field is described by the fully coupled Maxwell equations. Moreover the heat conduction equation is considered to describe thermal effects. To solve this multifield the equations are in the first step decoupled using an additional time differentiation. In the second step an axisymmetric case is considered, motivated by the fact that the inductive heating process of a cylindrical shaft is analysed. Afterwards the resulting equations are spatially discretized by the Galerkin finite element method. The temporal discretization is carried out via the Newmark method so that afterwards the electrical source distribution can be achieved. As a consequence the temperature evolution is determined in a postprocessing step.

1 INTRODUCTION

In order to optimize material properties, new material composites or new fabrication sequences are developed. An example for the latter mentioned aspect is a thermo-
mechanical forming process, where locally varying material properties are achieved due to heterogenous temperature fields during the manufacturing process. As an illustration the production of a shaft within an integrated fabrication process is treated. The first step consists of an inductive heating, afterwards the shaft is formed as well as cooled due to the contact with the forming die and in the end a cold pressured air stream is applied, compare Figure 1.

![Figure 1: Thermomechanical forming process with inductive heating [6]](image)

The inductive heating step is an essential part of the production sequence, since due to this a material gradation is enabled and the forming process is facilitated. Thus this process step will be analyzed in what follows solving the coupled Maxwell equations using a finite element spatial discretization and a time discretization.

2 MAXWELL EQUATIONS AND CONSTITUTIVE ASSUMPTIONS

To describe electromagnetic phenomena mathematically, for instance that magnetic sources do not exist (2), the Maxwell equations

\[ \nabla \cdot D = \rho_R \quad (1) \]
\[ \nabla \cdot B = 0 \quad (2) \]
\[ \nabla \times H = J + \dot{D} \quad (3) \]
\[ \nabla \times E = -\dot{B} \quad (4) \]

are used. Herein \( B \) denotes the magnetic and \( D \) the electric flux density. Moreover, \( H \) as well as \( E \) are called magnetic and electric field intensity, respectively. The variable \( J \) represents the electric current density and \( \rho_R \) the electric charge density. Nevertheless not all variables included in the Maxwell equations are independent of each other, thus they are linked, assuming the following constitutive equations, [3].

\[ D = \varepsilon E \quad (5) \]
\[ B = \mu H \quad (6) \]
\[ J = \sigma E + J_1 \quad (7) \]

Herein
\[ \epsilon = \epsilon_0 \epsilon_R \]  

represents the permittivity, consisting of the electric permittivity in the vacuum \( \epsilon_0 \) and a material dependent part \( \epsilon_R \). Likewise the variable
\[ \mu = \mu_0 \mu_R \]

embodies the permeability. The constant \( \sigma \) symbolizes the electric conductivity and the variable \( J_i \) the intrinsic current density.

Having a closer look on the equations above the basis of an inductive heating process can be explained. In general in an inductive heating process the workpiece is surrounded by an induction coil, on which an alternating current is applied, [5]. In accordance with equation (3) this alternating current \( J \) produces an alternating magnetic field, which is, following equation (4) and (7), the source of induced eddy currents in the workpiece. If in addition the heat equation
\[ \rho \sigma c \frac{\partial \Theta}{\partial t} + \sigma \nabla \cdot q = J^2 \]  

with the heat capacity \( c \), the density \( \rho \) and the constitutive equation
\[ q = -\lambda \nabla \Theta \]  

connecting the heat flux vector \( q \) with the temperature \( \Theta \) throughout the heat conduction coefficient \( \lambda \), is considered, then it can be followed that these eddy currents are the reason for the temperature increase in the workpiece. Since the equations for the inductive heating process are now all known, the next chapter will show one possible solution strategy, [1].

3 AXISSYMMETRIC FORMULATION

The following analysis is motivated by the inductive heating of a shaft with the help of an induction coil. Both of this parts can be considered as axisymmetric domains, so that instead of using a three-dimensional model an axisymmetric formulation considering cylindrical coordinates \((r, \phi, z)\) is used. Thus a meridian half-plane of the shaft and the induction coil, described by the coordinates \((r, z)\) and setting \( \phi = \text{const.} \), is observed, compare Figure 2. Before this assumption is exploited further, the MAXWELL equations are reformulated. Therefore, the constitutive laws (5) as well as (6) are inserted in equation (3)
\[ \nabla \times \left( \frac{1}{\mu} B \right) = J + \epsilon \dot{E}. \]  

The resulting relation is time differentiated and equation (4) is furthermore considered. This leads to
\[ \sigma \frac{\partial E}{\partial t} + \epsilon \frac{\partial^2 E}{\partial t^2} + \frac{1}{\mu} \nabla \times \nabla \times E = -\frac{\partial J_i}{\partial t}. \]
An analogous treatment of equation (4) results in a similar relation

$$\epsilon \frac{\partial^2 B}{\partial t^2} + \frac{1}{\mu} \nabla \times \nabla \times B = \nabla \times J. \quad (14)$$

The other two MAXWELL equations (1) as well as (2) are not affected and stay unaltered. Thus by considering geometric aspects in the framework of an axisymmetric formulation a decoupling of AMPERÉ’s (3) and FARADAY’s law (4) is achieved, if in addition the current density $J$ is known and adequate initial conditions are assumed, [2]. On the explicit formulation of the initial conditions as well as of the boundary conditions will be commented later on.

Since the aim is to describe the temperature evolution, moreover, just the electric equation (13) together with the heat equation (10) are considered and their weak forms are generated.

### 3.1 WEAK FORM OF THE ELECTRIC FIELD EQUATION AND THE HEAT EQUATION

To enable the usage of the Finite Element Method, partial differential equations have to be formulated weakly throughout a multiplication with a test function i.e. $\delta E$ or $\delta \Theta$ and an integration over a volume $\Omega$. For the electric field evolution equation and the temperature evolution equation this leads to:

![Figure 2: Shaft with induction coil](image-url)
\[
\int_\Omega \sigma \delta \mathbf{E} \cdot \dot{\mathbf{E}} \, dV + \int_\Omega \epsilon \delta \mathbf{E} \cdot \ddot{\mathbf{E}} \, dV + \frac{1}{\mu} \int_\Omega \delta \mathbf{E} \cdot \nabla \times \nabla \times \mathbf{E} \, dV + \int_\Omega \delta \mathbf{E} \cdot \dot{\mathbf{J}}_1 \, dV = 0 \quad (15)
\]
\[
\int_\Omega \delta \mathbf{E} \cdot \nabla \Theta \, dV + \int_\Omega \epsilon \delta \mathbf{E} \cdot \nabla \cdot \mathbf{q} \, dV - \int_\Omega \delta \mathbf{E} \cdot \dot{\mathbf{J}}^2 \, dV = 0 \quad (16)
\]

Using Green's formula
\[
\int_\Omega \delta \mathbf{E} \cdot \nabla \times \nabla \times \mathbf{E} \, dV = \int_\Omega \nabla \times \delta \mathbf{E} \cdot \nabla \times \mathbf{E} \, dV + \int_\Gamma \delta \mathbf{E} \cdot \mathbf{n} \times \nabla \times \mathbf{E} \, dA \quad (17)
\]
and the divergence theorem, equations (15) and (16) can be determined to
\[
\int_\Omega \sigma \delta \mathbf{E} \cdot \dot{\mathbf{E}} \, dV + \int_\Omega \epsilon \delta \mathbf{E} \cdot \ddot{\mathbf{E}} \, dV + \frac{1}{\mu} \int_\Omega \nabla \times \delta \mathbf{E} \cdot \nabla \times \mathbf{E} \, dV + \int_\Gamma \delta \mathbf{E} \cdot \mathbf{n} \times \nabla \times \mathbf{E} \, dA + \int_\Omega \delta \mathbf{E} \cdot \dot{\mathbf{J}}_1 \, dV = 0 \quad (18)
\]
\[
\int_\Omega \delta \mathbf{E} \cdot \nabla \Theta \, dV - \int_\Omega \nabla \delta \Theta \cdot \mathbf{q} \, dV + \int_\Gamma \delta \mathbf{E} \cdot \mathbf{n} \cdot \mathbf{q} \, dA - \int_\Omega \delta \mathbf{E} \cdot \dot{\mathbf{J}}^2 \, dV = 0 \quad (19)
\]

### 3.2 Spatial Discretization

Due to the spatial discretization the whole domain \( \Omega \) is divided into several finite elements and the continuous field variables are described with the help of discrete node values and shape functions. For this, standard Lagrange functions \( N^i \) are employed, [7].

\[
\mathbf{E}^e \approx \sum_{i=1}^{NN} \mathbf{E}^{ei} N^i \quad \Theta^e \approx \sum_{i=1}^{NN} \Theta^{ei} N^i \quad (20)
\]
\[
\dot{\mathbf{E}}^e \approx \sum_{i=1}^{NN} \dot{\mathbf{E}}^{ei} N^i \quad \dot{\Theta}^e \approx \sum_{i=1}^{NN} \dot{\Theta}^{ei} N^i \quad (21)
\]
\[
\delta \mathbf{E}^e \approx \sum_{i=1}^{NN} \delta \mathbf{E}^{ei} N^i \quad \delta \Theta^e \approx \sum_{i=1}^{NN} \delta \Theta^{ei} N^i \quad (22)
\]

If further the geometrical aspect is considered, that the current density \( \mathbf{J} \) has just a component in \( \mathbf{e}_\phi \) direction, then the spatial discretized weak forms result in
\[
\sum_{i=1}^{N_N} \sum_{j=1}^{N_N} \delta E^{ei} \int_{\Omega_e} N^i N^j \dot{E}^{ej} r dr dz + \frac{\delta E^{ei}}{\mu} \int_{\Omega_e} \left[ \frac{\partial N^i}{\partial z} \frac{\partial N^j}{\partial z} + \frac{1}{r^2} \frac{\partial (r N^i)}{\partial r} \frac{\partial (r N^j)}{\partial r} \right] E^{ej} r dr dz = 0 
\]

\[
\sum_{i=1}^{N_N} \sum_{j=1}^{N_N} \delta \Theta^{ei} \int_{\Omega_e} \rho c \sigma \dot{\Theta}^{ej} N^i r dr dz + \sigma \dot{\Theta}^{ei} \int_{\Omega_e} \left[ \frac{\partial N^i}{\partial r} \frac{\partial N^j}{\partial r} + \frac{\partial N^i}{\partial z} \frac{\partial N^j}{\partial z} \right] r dr dz + \delta \Theta^{ei} \int_{\Omega_e} J^2 N^i r dr dz = 0 
\]

for each element. To get the above listed equations, adequate boundary conditions have to be exploited. For the electric field this is fulfilled by setting

\[
E = 0 \quad \text{on} \quad \Gamma_s \cup \Gamma_E. 
\]

For the boundary \(\Gamma_s\) this represents a symmetry condition, whereas on the the rest of \(\Gamma_E\) the influence of the electric field is neglected. The boundary condition for the temperature evolution equation is

\[
q \cdot n = 0 \quad \text{on} \quad \Gamma_\Theta. 
\]

If all elements are assembled to the desired structure and matrix notations are used, then equations (24) and (25) result in

\[
M \ddot{E} + D_E \dot{E} + K_E E + r_E = 0 
\]

\[
D_\Theta \dot{\Theta} + K_\Theta \Theta + r_\Theta = 0.
\]

Moreover these equations have to be discretized temporarily.

### 3.3 TEMPORAL DISCRETIZATION

In a first step the time interval \([0, T]\) with the time step \(\Delta t\) is examined using as a time discretization the Generalized-\(\alpha\) method, \([4]\), with \(\alpha_m = \alpha_t = 0.5\)

\[
\dot{E}_{n+\frac{1}{2}} = \frac{1}{2} \left[ \frac{\gamma}{\beta \Delta t} \left[ E_{n+1} - E_n \right] - \gamma - \frac{\beta}{\beta} \dot{E}_n + \dot{E}_n \right] 
\]

\[
\ddot{E}_{n+\frac{1}{2}} = \frac{1}{2} \left[ \frac{1}{\beta \Delta t^2} \left[ E_{n+1} - E_n \right] - \frac{1}{\beta \Delta t} \dot{E}_n - \frac{1 - 2\beta}{2\beta} \ddot{E}_n + \ddot{E}_n \right] 
\]

\[
\dot{\Theta}_{n+\frac{1}{2}} = \frac{1}{2} \left[ \frac{\gamma}{\beta \Delta t} \left[ \Theta_{n+1} - \Theta_n \right] - \gamma - \frac{\beta}{\beta} \dot{\Theta}_n + \dot{\Theta}_n \right], 
\]

For the boundary \(\Gamma_s\) this represents a symmetry condition, whereas on the the rest of \(\Gamma_E\) the influence of the electric field is neglected. The boundary condition for the temperature evolution equation is

\[
q \cdot n = 0 \quad \text{on} \quad \Gamma_\Theta. 
\]

If all elements are assembled to the desired structure and matrix notations are used, then equations (24) and (25) result in

\[
M \ddot{E} + D_E \dot{E} + K_E E + r_E = 0 
\]

\[
D_\Theta \dot{\Theta} + K_\Theta \Theta + r_\Theta = 0.
\]

Moreover these equations have to be discretized temporarily.
assuming $2\beta = \gamma = \frac{1}{2}$ and considering

\[
\dot{E}(t_0) = \dot{E}_0, \quad E(t_0) = E_0
\]

and

\[
\Theta(t_0) = \Theta_0
\]

as appropriate initial conditions. Inserting the above mentioned formulations leads to two linear systems of equations

\[
\hat{K}_E \dot{E} = \hat{r}_E \left( \ddot{E}_n, \dot{E}_n, E_n \right)
\]

and

\[
\hat{K}_\Theta \dot{\Theta} = \hat{r}_\Theta \left( \dot{\Theta}_n, \Theta_n \right)
\]

which can be solved in two steps. In the first step the electric field equation is solved, determining the current density $J$ at each node. Afterwards the temperature evolution equation is solved in a postprocessing step, containing the current density $J$ as prescribed load vector and determining the temperature field.

Since now the theoretical aspects are explained in the next chapter a simulated inductive heating process is described.

### 4 NUMERICAL EXAMPLE

As a numerical example for the above deduced theory the inductive heating process of the shaft, see Figure 2, with three rings as induction coil will be examined. Therefor a sinusoidal current with a current $I$ and a frequency $f$ will be applied to each of the induction rings. Then the evolution of the electric field and the temperature field will be calculated. The necessary material parameters and load factors are listed in the table below.

<table>
<thead>
<tr>
<th>$\mu_{\text{Rshaft}}$</th>
<th>$\mu_{\text{Rcoil}}$</th>
<th>$\epsilon_{\text{Rshaft}}$</th>
<th>$\epsilon_{\text{Rcoil}}$</th>
<th>$\sigma$</th>
<th>$\lambda$</th>
<th>$c$</th>
<th>$I$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>600 Vs Am</td>
<td>1 Vs Am</td>
<td>0 As Vm</td>
<td>1 As Vm</td>
<td>$10^6$ 1 $\Omega$m</td>
<td>39,88 W mK</td>
<td>620 Nm kgK</td>
<td>5000A</td>
<td>8100 Hz</td>
</tr>
</tbody>
</table>

The following figures, compare Figure 3, are achieved for the electrical as well as for the temperature field. It can be seen that with increasing distance $r$ from the shaft the influence of the electrical field decreases. Moreover, the three induction rings cause a symmetric distribution of the electric field inside the shaft with a very small induction depth. At the interface between shaft and air an unsteady electric field is noticeable, which is explicable due to the different material properties.

For the calculation of the temperature field only the shaft itself was considered and the influence of the surrounding air was neglected. The temperature field distribution is symmetric but in contrast to the electric field distribution not only the periphery is affected. The whole shaft, at the height of the induction coils, is heated up.
5 CONCLUSION AND OUTLOOK

In this paper a method is shown to determine the temperature field during an inductive heating process. Therefore the MAXWELL equations are used to derive an axisymmetric formulation for the electric field intensity, whose solution enables the determination of the current density distribution. Furthermore, in a postprocessing step, the heat equation is solved, considering the current density as heat source.

To achieve the before mentioned, first the weak form of the electric field intensity equation as well as the weak form of the temperature evolution equation has to be generated. Then the finite element method is used as spatial discretization and the generalized-α method is carried out as temporal discretization. Afterwards linear systems of equations are solved and moreover, the realization is demonstrated.

In the next step a monolithic scheme should be developed to solve the electrical field equation and the heat equation at the same time. This approach towards a nonlinear system of equations will make it possible to consider temperature dependent material properties.

6 ACKNOWLEDGEMENT

Financial support was provided by the German National Science Foundation (DFG) in the framework of project C1 of the collaborative research center (SFB/TR TRR 30). This support is gratefully acknowledged.
REFERENCES


NUMERICAL ANALYSIS OF THE COUPLING BETWEEN MECHANICAL STRAIN AND THERMAL CONDUCTIVITY OF CERAMIC MATRIX COMPOSITES.

DAVID R. HAYHURST

School of Mechanical, Aerospace, & Civil Engineering, The University of Manchester, George Begg Building C-004, MANCHESTER, M13 9PL, U.K.
e-mail: d.r.hayhurst@manchester.ac.uk

Key words: Ceramic matrix composites, Tows, Unit cell, 0-90 woven composites, Stress-strain response. Thermal conductivity degradation, Damage mechanisms.

Abstract. The paper addresses the numerical modeling of the coupling between mechanical strain and thermal conductivity in Ceramic Matrix Composites (CMCs) when subjected to uni-axial mechanical straining.

A computationally economic finite element-based multi-linear elastic orthotropic mechanical materials description combined with a multi-linear discretisation of thermal conductivity-strain response has been developed to predict the stress-strain, fracture, and thermal conductivity-strain behaviour of a ceramic matrix composites with strain-induced damage. The finite element analysis utilises a solid element to represent an homogenised orthotropic medium of a heterogeneous uni-directional tow. The discretised non-linear multi-axial stress-strain curves and non-linear multi-axial strain dependent thermal behaviour have been implemented by a user defined subroutine in the commercial finite element package ABAQUS. The model has been used to study the performance of a carbon fibre/carbon matrix-SiC matrix (C/C-SiC) plain weave laminate DLR-XT. With the effects of fibre waviness included, the global stress-strain curves, with catastrophic fracture behaviour, and the thermal conductivity-strain response have been predicted. Excellent comparisons have been made between predictions and experimental data, with fibre waviness included.

1 INTRODUCTION

The superior material properties of Ceramic Matrix Composites (CMCs), e.g. low density and good mechanical and thermal properties at high temperatures, make them favourable materials for use in: rocket nozzles; thermal protection systems; and gas turbine engines [1]. Increased operating temperatures, from 900-1200°C for coated metallic superalloys to above 1300°C, for CMCs, have the potential to achieve higher thermal efficiencies and lower emissions [2]. The economic use of CMCs in engineering components and structures requires an ability to simulate the material, and the component response at both the design and manufacturing stages [3]. To perform optimal design of these CMC engineering components, an efficient yet computationally accurate model is a necessary requirement.

The present paper addresses the development of a conceptually simple and computationally economic finite element, FE, technique, which emphasises the effects of
strain-induced damage modes and their interactions on the mechanical behaviour. The approach adopted by [4] of modelling the composite using tows, i.e. a collection of thousands of fibres embedded in at least one matrix, has been adopted here. The tow is captured in a single orthotropic finite element, and the approach used to model the multi-axial stress-strain response, and property degradation for 0º/90º plain weave composites.

The DLR–XT, 10 high plain weave laminae [5], has been previously studied; and, an idealised unit cell of the material, that includes four tows, is shown schematically in Fig.1.

2 MECHANICAL BEHAVIOUR OF A UNI-DIRECTIONAL CMC TOW
Damage modes and their interactions on the stress-strain response are now introduced.

Figure 2: Schematic drawings of: (a) a uni-directional tow under longitudinal extension; (b) longitudinal stress-strain and Poisson’s ratio-strain curves; and (c)–(f) corresponding damage modes.
2.1 Longitudinal stress-strain response
When a uni-directional fibre-reinforced ceramic matrix composite tow undergoes uniform straining along the fibre direction, Fig.2a, it exhibits the longitudinal stress-strain response of Fig.2b [6]. Four characteristic stages in the stress-strain curve are [4]: **AB**: linear elastic material; **BC**: periodic matrix cracking occurs, Fig.2c; **CD**: fibre-matrix interface debonding and wake debonding, Fig.2d; and weaker fibre failure, Fig.2e; **DE**: the majority of fibres fail and are pulled out against the frictional stress along the wake debonding interface, Fig.2f. Figure 2b shows the variation of Poisson’s ratio with longitudinal strain.

2.2 Transverse stress-strain response
When a uni-directional tow is subjected to transverse extension, Fig.3a, the material response is assumed to be linear elastic. The transverse Young’s modulus, defined in Fig.3b, and the Poisson’s ratio, $\nu_{21}$ shown in Fig.3c, are constant until the fibre-matrix debonding and/or matrix cracking take place, Fig.3d, at which point the damage is instantaneous and causes complete loss in strength, and the stress immediately drops to zero [7].

2.3 Shear stress-strain response
The shear stress-strain response, Fig.4a, is similar to the transverse response. The shear stress increases linearly with the shear strain Fig.4b, until a critical value is reached, when catastrophic shear failure, Fig.4c, occurs, and the shear stress immediately drops to zero, [8].

Figure 3: Schematic drawings of: (a) a uni-directional tow under transverse straining; (b) and (c) transverse stress-strain and Poisson’s ratio curves; and (d) schematic of corresponding damage modes.

Figure 4: (a) a uni-directional tow under shear; (b) shear stress-strain curve; and (c) damage modes.
2.4 Longitudinal stress-strain response under multi-axial loading

When a uni-directional tow is subjected to two or more uni-axial loading cases, e.g. longitudinal, transverse, and shear loading, the effects of transverse tension or shear stress on the longitudinal stress-strain response are reflected in CMC experiments [5] addressed here.

For a uni-directional tow under longitudinal uni-axial loading, the material exhibits ductile behaviour, demonstrated as the gentle decreasing stress-strain curve after the peak in Fig.2b and Fig.5a [8]. This is due to the fibre-matrix interface wake debonding that occurs gradually, and creates a partially intact interface, which allows a failed fibre to pull out against the frictional stress along the wake debonding interface. It takes place within individual blocks, where a block is defined by a single fibre and matrix between two adjacent matrix cracks [4].

The variation of the normalised number of wake debonded blocks, \(N/N_T\) with respect to the total number of blocks \(N_T\) in a tow \((N/N_T)\), with composite strain, \((\varepsilon_{33})_\infty\), is shown in Fig.5a. A positive transverse stress or a shear stress advances wake debonding, and also degenerates the partially intact interface to a fully separated gap. Hence the frictional stress reduces to zero and the pullout mechanism is then deactivated.

Even though the pullout mechanism is negated, it cannot explain the catastrophic fibre failure observed in the experiments [5]. Another mechanism, dynamic fibre failure by instantaneous pullout deactivation [8], is responsible for this brittle failure. The mechanism is that approximately one half of all blocks in a tow \((N/N_T = 0.5)\), at the strain \(\varepsilon_{wd}\) shown in Fig.5a simultaneously undergo instantaneous wake debonding and fibre pullout deactivation.

By virtue of the way in which a ductile CMC is designed, the average wake debonding strain, \(\varepsilon_{wd}\), is slightly less than the peak composite strain, and hence at this point the fibres are stressed to a high fraction of the average fibre failure stress. When the mechanism is triggered, a shear stress wave travels from the matrix cracks associated with the block; this induces a tension stress wave in the fibre to maintain equilibrium. The two tension waves
meet in the fibre at the centre of the block, and reflect as a wave with twice the amplitude of
the incident wave. This doubling of the tensile stresses in the fibres therefore causes complete
failure of the tow. By using this mechanism, the original curve, Fig.5b, switches to the
degraded curve, which is for a tow subjected to multi-axial loading.

The variation of transverse, shear and axial tow properties with tow axial strain are
assumed to fall to zero at failure. This dynamic failure is consonant with experimentally
observed behaviour [5]. The transverse, and shear properties may not immediately drop to
zero at failure; however, the worst condition of brittle failure has been assumed.

3 FORMULATION OF THE FINITE ELEMENT MODEL
A uni-directional tow is chosen to be the basic constituent in the FE model, and therefore an
entire tow can be represented by a single 8-node solid finite element. This can be seen in
Fig.2a where the 8 corners define the nodes. The material properties are assumed to be multi-
linear elastic and the stress-strain and Poisson’s ratios-strain curves, shown in Figs.2, 3 and 4,
are used in the constitutive equations. The FE package ABAQUS, SIMULIA [9] with a user-
defined subroutine UMAT is used to carry out the simulation.

3.1 Homogenisation of a uni-directional tow.
A heterogeneous uni-directional tow or lamina is homogenised to a single block which has
the same overall dimensions and equivalent orthotropic material properties. There are nine
independent material properties: Young’s moduli: \( E_3, E_2, \) and \( E_1 \); Poisson’s ratios: \( \nu_{31}, \nu_{32}, \) and \( \nu_{12} \); and, Shear moduli: \( G_{12}, G_{23}, \) and \( G_{13} \).

3.2 Discretisation of the stress-strain curve and constitutive equations
In order to perform a finite element analysis, the non-linear longitudinal stress-strain
relationship is discretised to a multi-linear curve. The loading is imposed in terms of the
displacement boundary conditions. The applied displacement is divided into many small
increments. For each increment, the constitutive equation for an orthotropic material is [7]:

\[
\{\Delta \sigma\} = [C(\Delta \varepsilon)] \{\Delta \varepsilon\};
\]  

(1)

\( \{\Delta \sigma\} \) and \( \{\Delta \varepsilon\} \) are the incremental stress and strain vectors, and \( [C(\Delta \varepsilon)] \) is the stiffness matrix.

3.3 Assumption made for the activation of catastrophic fracture failure
To activate dynamic fibre failure by instantaneous pullout deactivation, Fig.5b, two
assumptions are made: (a) the switch from the original stress-strain curve to the degraded
curve occurs by a small finite positive transverse stress, or by a small shear stresses; and (b)
as a consequence, the Poisson’s ratios, \( \nu_{31} \) and \( \nu_{32} \) start to decrease at the matrix cracking
strain, and fall to zero at, \( e_{wd} \) in Fig.5a; as opposed to the peak load strain at D, in Fig.2b.

3.4 UMAT implementation in ABAQUS SIMULIA [9]
The multi-axial elastic properties, the constitutive equations, and the catastrophic failure
activation are defined in the ABAQUS UMAT. For every increment, the UMAT reads the
strains at each material point and then assigns the corresponding Young’s moduli, Poisson’s
ratios and shear moduli. The constitutive equations are then used to update the stress fields.
Solution-dependent state variables, STATEV, record different damage modes and control
their interactions. Automatic incrementation algorithms and redefinition variable, PNEWDT, are updated to ensure the discretised points are located exactly at the end of the increments.

Nine state variables are used to record the respective damage states associated with the nine material properties. The initial values of state variables, STATEV, are set to be 0. When the damage state changes, i.e. transverse cracking, shear failure, or dynamic fibre failure occurs, their corresponding STATEV value is set to be 1. The STATEV values associated with the Poisson’s ratios, $\nu_{31}$ and $\nu_{32}$, are controlled by dynamic fibre failure. Once STATEV values are equal to 1, the material properties become a very small finite value, no matter what the local strains are. In the UMAT, the Jacobian matrix for mechanical loading is required and the expressions used for the current constitutive equation (Equation (1)) are given in [7].

3.5 Load path algorithm

Post-peak behaviour of composites and the associated loss of uniqueness, can pose a problems. This leads to zones of composite unloading with adjacent loading zones, whilst both zones are subjected to increasing strain, hence compatibility can be satisfied whilst equilibrium may be violated. The author has developed an approach that identifies the load-path within each tow that links many finite elements, and ensures that lines of force have continuity through the unit cell. Failure in one tow location is automatically transmitted along the tow in a way that ensures satisfaction of both compatibility and equilibrium.

To model the catastrophic dynamic fibre failure mechanism, load paths have been introduced into the FE analysis. A load path consists of a series of elements which form a continuous tow or lamina. It is assumed that: (a) In a load path, once the local longitudinal strain of any integration point (master point) reaches 90% of the wake debonding strain, $\varepsilon_{wd}$, the Poisson’s ratios, $\nu_{31}$ and $\nu_{32}$, of all the other integration points (slave points) in this load path decrease proportionally with the reduction in the corresponding variable at the master point. The Poisson’s ratios, $\nu_{31}$ and $\nu_{32}$, of all the integration points reduce to zero when the longitudinal strain of the master point reaches the wake debonding strain, $\varepsilon_{wd}$. (b) After that, elements in all load paths fail simultaneously and all their stress components unload to zero. The unloading slope is chosen to be as steep as possible, but consistent with maintaining numerical stability.

3.6 Implementation of the incremental constitutive law

For each FE increment, the constitutive law [7] has been implemented as follows:
1) Determine the material properties from the local strains and the multi-linear curves.
2) Identify the damage state and update the state variable, STATEV. If the dynamic fibre failure criterion is satisfied, then the load path algorithm is activated.
3) Update the stress increments using the incremental constitutive equations.
4) Update the material Jacobian matrix using equation [7].

4 MECHANICAL BEHAVIOUR OF PLANE WEAVE DLR-XT COMPOSITES

This section addresses the prediction of the stress-strain response of a plain weave DLR-XT composite, Fig.1, under uni-axial straining along the 0º tows. Effects of the tow waviness on the overall stress-strain curve of the composite are modelled. The material properties of the tow are used as input data to predict the overall mechanical behaviour of the composite.
4.1 Formulation of the finite element Model
The FE mesh for a single tow, consists of 12 elements: two 8-node brick elements, six 6-node wedge elements, and four 4-node tetrahedral elements. The DLR-XT unit cell model is an assembly of four tow models, Fig.6, bonded together through the fully contacted interfaces. Inevitably, the small number of elements is achieved with some loss of accuracy. The fidelity of the model will be assessed by using the experimental data in Section 4.5.

4.2 Modelling of waviness
Waviness in the DLR-XT material and the misalignment angle, $\xi = 7^\circ$, has been measured over a large region of available micrographs; and used to create the material orientation of the tows in the finite element model. Figure 7 shows the fibre directions.

4.3 Material properties
Blacklock and Hayhurst [8,10] reported the initial elastic properties and multi-axial failure of the tows. The longitudinal stress strain curve of uni-directional tows were obtained using the physical model of Hayhurst et al., [6] and Tang et al., [4] and the fracture behaviour by dynamic fibre failure mechanism by instantaneous pullout deactivation due to [8].

4.4 Elastic Properties
All the elastic properties used for the finite element model are given in Fig.8. The discretised longitudinal stress-strain curves are shown in Fig.8a, the solid line is for a uni-directional

![Figure 6: Unit cell model: (a) exploded view and (b) assembled view. 1-direction tows have light shading, and 3-direction tows are dark shaded.](image)

![Figure 7: Tow waviness: (a) plan view of 1-3 plane and (b) side elevation of 2-3 plane. Misalignment angle is shown as $\xi = 7^\circ$.](image)
Figure 8: Material properties of a uni-directional DLR-XT tow: (a) longitudinal stress-strain curve (b) transverse stress-strain curves (c) Poisson’s ratios, $\nu_{12}$, $\nu_{31}$ and $\nu_{32}$ and (d) shear stress-strain curves.

tow under longitudinal straining, and the broken line is the degraded curve, which considers the dynamic fibre failure at $\varepsilon_{33} = \varepsilon_{\text{wd}}$. The variation of Poisson’s ratios, $\nu_{12}$, $\nu_{31}$ and $\nu_{32}$, with longitudinal strain are shown in Fig.8b. The transverse and shear stress-strain curves, shown in Fig.8c and 8d, are smoothed around the peaks of the curves.

4.5 Predictions

Using the FE-based multi-linear elastic model and the material properties in Fig.8, two predictions are made for $\xi = 0^\circ$ and $7^\circ$ as defined in Section 4.2. Fig.9 shows a comparison between the predictions and experimental data [5]. The effects of tow waviness on the stiffness of the stress-strain curves are significant. The model $\xi = 0^\circ$ predicts a higher stiffness, and failure stress, but a lower failure strain than for $\xi = 7^\circ$. In comparison with the experimental data, only the prediction for $\xi = 7^\circ$, correlates well with experimental data. Hence the effect of tow waviness on the strength and fracture strain is significant.

Figure 9: Comparison between predicted stress-strain curves and experimental data for a plain weave DLR-XT laminate.
5 THERMAL BEHAVIOUR OF A UNI-DIRECTIONAL CMC TOW

FE analysis has been done using coupled stress-thermal-strain strain data.

5.1 Tow longitudinal thermal conductivity

Tang et al. [4] showed that the longitudinal thermal conductivity is controlled by the air gaps introduced on matrix cracking and fibre failure. They derived equations, based on the thermo-mechanical properties of the constituent materials, which can be numerically integrated to produce variations of the local tow longitudinal thermal conductivity, \( k_{\text{Long}}^l \), with local tow strain, \( \varepsilon_{33}^l \). The relevant equations for variation of \( k_{\text{Long}}^l / k_{\text{Long}}^{\text{Initial}} \) with the tow damage variable \( \omega \) and for the variation of \( k_{\text{Long}}^l / k_{\text{Long}}^{\text{Initial}} \) with \( \varepsilon_{33}^l \) are obtained by coupled numerical integration of the governing equations [4]. Using these relations, and values of \( k_{\text{Long}}^{\text{Initial}} \), the variation of \( k_{\text{Long}}^l \) with \( \varepsilon_{33}^l \) have been obtained from the base material parameters and are given in Fig.10a. An initial value of \( k_{\text{Long}}^{\text{Initial}} = k_{33}^l = 20.38 \text{ Wm}^{-1}\text{K}^{-1} \) has been used to determine Fig.10a, shown discretised as multi-linear curves using typically 25 data crosses.

The first step reduction in \( k_{\text{Long}}^l \) shown in Fig.10a is due to matrix cracking and the subsequent monotonic decrease that results from fibre failure. The data in Fig.10a is used as the discretised multi-linear materials input data for the finite element model.

5.2 Tow transverse thermal conductivity

Tang et al. [4] have shown that the degradation of transverse thermal conductivity is controlled by the process of wake debonding, Fig.2d, which produces a cylindrical air gap at the interface between fibre and matrix in a block of material associated with a single fibre located between two adjacent matrix cracks. Poor transverse thermal conductivity of the cylindrical air gap prevents transverse heat flow through the fibre and some of the matrix. The degradation of the local transverse thermal conductivity, \( k_{\text{Tan}}^l \), with local tow strain, \( \varepsilon_{33}^l \), is related to the ratio, \( N/N_T \). The variation of \( k_{\text{Tan}}^l / k_{\text{Tan}}^{\text{Initial}} \) with \( \varepsilon_{33}^l \) has been determined from Tang et al [4]; with \( k_{\text{Tan}}^{\text{Initial}} = k_{11}^l = k_{22}^l = 12.56 \text{ Wm}^{-1}\text{K}^{-1} \). The graph of \( k_{\text{Tan}}^l \) against \( \varepsilon_{33}^l \) is given in Fig.10b, which is used in the FE analysis discretised by 25 data crosses.
5.3 Effect of shear strain on the longitudinal and through-thickness thermal conductivity

In the composite weave section, Fig. 11a, stress and strain components are defined in the local coordinate system, where $\sigma_\infty$ and $\varepsilon_\infty$ are the global remote stress values. Failure of the fibre/matrix interface is shown in Fig. 11b due to local shear and normal stress. When $\varepsilon_{32}^f$, is greater than the critical failure strain, $\varepsilon_{32}^{fs}$, the continuity of the fibre-matrix interface is broken and partial failure is assumed. The partial failure affects the mechanical behaviour, but the thermal response is unchanged since the failure surfaces are in contact. If the normal stress, $\sigma_{22}^f$, is less than a small positive lower level, $\sigma_{22}^{UL}$, i.e. $\sigma_{22}^f < \sigma_{22}^{UL}$, the fibres and matrix are assumed to be in sufficient contact for heat transfer to take place. However, when $\sigma_{22}^f > \sigma_{22}^{UL}$, an air gap is assumed to form, and transverse heat transfer is prevented. The effects of a through-thickness shear failure condition, $\varepsilon_{32}^f > \varepsilon_{32}^{fs}$, and $\sigma_{22}^f > \sigma_{22}^{UL}$ on transverse conductivity are assumed to dominate over the effects of wake debonding that results in the separation of the two surfaces on the failure plane due to $\sigma_{33}$. The following mechanisms will be considered: (i) Wake debonding driven by longitudinal tension $\sigma_{33}^f$; or (ii) Shear failure induced by shear or transverse tension, $\varepsilon_{32}^f > \varepsilon_{32}^{fs}$, with $\sigma_{22}^f > \sigma_{22}^{UL}$.

6 FORMULATION OF THE FINITE ELEMENT MODEL

The coupled thermal-mechanical stress model reported here has been formulated using strain dependent thermal material properties, Fig. 10. The details of the formulation for mechanical behaviour are given in earlier sections, and can also be found in Zhang and Hayhurst [7]. The formulation of the model for thermal behaviour is now addressed [11].

6.1 Heat flow in a homogenised uni-directional tow

A heterogeneous uni-directional tow or lamina was homogenised to a single block e.g. Fig. 2a. For the present three dimensional steady-state heat conduction problem, there are three independent tow thermal conductivities: $k_1$, $k_2$, and $k_3$, defined relative to the local tow axes given in Fig. 2a. For anisotropic continua the steady-state heat conduction equation is given by Fourier’s law, particularised for composites by White and Knutsson [12].
\[
\begin{bmatrix}
q_1 \\
q_2 \\
q_3 
\end{bmatrix} = \begin{bmatrix}
k_{11} & 0 & 0 \\
0 & k_{22} & 0 \\
0 & 0 & k_{33} 
\end{bmatrix} \begin{bmatrix}
\frac{\partial T}{\partial x_1} \\
\frac{\partial T}{\partial x_2} \\
\frac{\partial T}{\partial x_3} 
\end{bmatrix}
\]

where \( q_i \) is the heat flux; \( k_{ij} \) the thermal conductivity; and \( \frac{\partial T}{\partial x_i} \) the thermal gradient (\( i, j = 1, 2, 3 \)). In local coordinates, c.f. Fig.2a, \( k_{33} \) is the thermal conductivity along the tow, \( k_{22} \) is the out-of-plane (through-thickness), and \( k_{11} \), is the in-plane transverse to tow.

### 6.2 Determination of Unit Cell Heat Flux

To perform the unit cell FE analyses the discretised multi-linear curves of Fig.10 were used. Mechanical straining in the 3-direction, c.f. Figs.2a and 6, was imposed in terms of the boundary displacements as a succession of small strain increments; and over each increment the mechanical-thermal problem was solved for a unity thermal gradient between the upper and lower surfaces of the unit cell. For each strain increment, element thermal conductivities were evaluated by ABAQUS from Fig.10 using the mechanical strain fields at the start of the increment; and these values were used to compute heat flux (Watts m\(^{-2}\)) at nodes over the unit cell lower face, at the end of the current strain increment. These local values were integrated over the lower surface to determine the total heat flux, to give the transverse thermal conductivity of the unit cell, since a unity thermal gradient boundary condition was imposed.

### 6.3 Implementation of Subroutine USDFLD in ABAQUS, SIMULIA, [9]

The present thermo-mechanical model has been implemented using the FE package, ABAQUS/standard and a user defined subroutine, USDFLD. The discretised strain dependent tow thermal conductivities, c.f. Fig.10 have been defined as functions of the strain field variables in the input file. For each increment, the subroutine USDFLD read the local strains at each integration point and defined the tow thermal conductivity properties. Automatic incrementation algorithms and the increment redefinition variable, PNEWDT, defined in ABAQUS, were continuously updated to ensure that values of the field variables at the end of the relevant strain increments were located exactly at the discretised points of Fig.10.

### 7 THERMAL CONDUCTIVITY-STRAIN RESPONSE OF DLR-XT COMPOSITE

#### 7.1 Formulation of the finite element Model

The mesh and geometry of the DLR-XT unit cell model is exactly as described in Section 4.1 for the model used in the mechanical analysis. The fidelity of the model for thermal flux is assessed by comparison with experimental data in Section 7.2.

Periodic displacement boundary condition was applied to the unit cell to simulate a uni-axial strain along the 0º tow direction. Steady state heat conduction was modelled by a unity thermal gradient between the top and bottom faces of the unit cell, all other faces were lagged.

#### 7.2 Predicted thermal response

For the materials tested by Sheikh et al. [5], no results were obtained for in-plane composite thermal conductivity, and no attempt was made to predict in-plane thermal
Figure 12: Predicted through-thickness thermal conductivities for DLR-XT ($\xi = \pm 7^\circ$) with effects of shear failure when $\varepsilon_{23}^f > \varepsilon_{23}^{bf}$ and $\sigma_{22}^f > 3.5 \text{MPa}$.

conductivities. Predictions of the variation of composite transverse thermal conductivity, $k_{22}$, with strain, $(\varepsilon_{33})_c$, are given in Fig.12 for $\xi = \pm 7^\circ$ with consideration of shear strain failure. The shear failure mechanism, introduced in Section 2.3 is assumed to be triggered when the transverse stress exceeds a lower bound, $\sigma_{22}^f > \sigma_{22}^{lf}$. Since no data is available for $\sigma_{22}^{lf}$, it has been determined by best fitting to experimental data, and a value of $\sigma_{22}^{lf} = 3.5 \text{MPa}$ was determined. The curve for $\xi = \pm 7^\circ$ in Fig.12 predicts the experimental data well.

Closer examination of the prediction in Fig.12 shows a sharp drop in the curve at $(\varepsilon_{33})_c \approx 0.17\%$. This is due to the shear failure mode being activated, $\varepsilon_{23}^f > \varepsilon_{23}^{bf} = 0.1152\%$ and overriding the reduction in thermal conductivity due to wake debonding [11]. Therefore, the plain weave DLR-XT has its out-of-plane transverse thermal conductivity actively reduced by the shear failure mechanism, coupled with a waviness angle of $\xi = \pm 7^\circ$.

8 CONCLUSIONS

Mechanical response: for the two levels of waviness considered $\xi = 0^\circ$ and $7^\circ$, the effect of increased waviness is to increase flexibility and increase failure strains. The predicted stress-strain curves for $\xi = \pm 7^\circ$ agree best with experimental results.

Thermal Response: with inclusion of failure due to out-of-plane shear very good predictions of the composite transverse thermal conductivity-strain curve have been achieved with waviness $\xi = \pm 7^\circ$. Hence thermal conductivity degradation takes place by out-of-plane shear, and by the wake debonding that accompanies fibre failure.

The coupled computational algorithm has successfully and accurately predicted both the mechanical and thermal responses of the DLR-XT Ceramic Matrix Composite.

9 REFERENCES


DECOMPOSITIONS OF THE STRESS AND THE RATE OF DEFORMATION TENSORS FOR MATERIALS UNDERGOING PHASE TRANSFORMATIONS

VASSILIS P. PANOSKALTSIS*

* Department of Civil Engineering
Demokritos University of Thrace
12 Vassilissis Sofias Street, Xanthi 67100, Greece
e-mail: vpanoska@civil.duth.gr

Key words: Phase Transformations, Shape Memory Alloys, Multi Surface Generalized Plasticity, Loading – Unloading Criteria, Lie Derivative, Finite Deformations, Internal Variables, Duhamel – Neumann Hypothesis, Rate of Deformation Tensor.

Summary: An extension of the “Duhamel-Neumann hypothesis” for materials undergoing phase transformations and for arbitrary magnitudes of strains and rotations is provided.

1 INTRODUCTION

Generalized plasticity theory has been successfully used by the author and his co-workers in order to model phase transformations of shape memory alloy materials. Our work has been initially carried out within small strains (Panoskaltsis et al. [1], Ramanathan et al. [2]) and very recently within finite deformations and rotations and under both isothermal and non-isothermal conditions (Panoskaltsis [3]). In this paper we will develop, for the first time, two important decompositions applied to materials undergoing phase transformations. In the first decomposition (Theorem 1) the rate of the Kirchhoff stress tensor is given in terms of the rate of deformation tensor and the rate of the temperature. In the second decomposition (Theorem 2), which can be thought of as a conjugate to the first one, the rate of deformation tensor is expressed as a sum of the (objective) rate of the stress tensor and the rate of the temperature. In the next section we will review the formulation of
generalized plasticity theory for modeling phase transformations in finite deformations and in both reference and current configurations of the body (for an exhaustive development, see Panoskaltsis [3]). Finally, in the last section, we will prove Theorems 1 and 2.

2 GENERALIZED PLASTICITY FOR PHASE TRANSFORMATIONS

2.1 Formulation of the governing equations in the reference configuration

Generalized plasticity is a local internal variable theory of rate – independent behavior which is based primarily on loading – unloading irreversibility. In the theory it is assumed that the local thermomechanical state in a body is determined uniquely by the couple \((G, Q)\) where \(G\) stands for the vector of the controllable state variables and \(Q\) stands for the vector of the internal variables, which are related to phase transformations. In this work, we follow a material (referential) approach within a strain – space formulation. Accordingly, \(G\) may be identified by \((E, T)\) where \(E\) is the referential (Green – St. Venant) strain tensor and \(T\) is the (absolute) temperature. Depending on the nature of the (material) internal variable vector \(Q\), the theory may, in principle, be formulated equivalently with respect to the macro –, meso –, or micro – scale structure of the material.

The central concept of generalized plasticity is that of the elastic range (e.g., see Lubliner [4]) which is defined at any material state, as the region in the strain – temperature space comprising the strains which can be attained elastically (i.e., with no change in the internal variables) from the current strain – temperature point. It is assumed that the elastic range is a regular set in the sense that it is the closure of an open set. The boundary of this set is defined as a loading surface at \(Q\) (see Eisenberg and Phillips [5], Lubliner [4]). In turn a
material state may be defined as elastic if it is an interior point of its elastic range and inelastic if it is a boundary point of its elastic range; in the latter case the material state lies on a loading surface. It should be added that the notion of process is introduced implicitly here. By assuming that the loading surface is smooth at the current strain - temperature point and by invoking some basic axioms and results from set theory and topology, Lubliner [4] showed that the rate equations for the evolution of the internal variable vector may be written in the form:

\[
\dot{Q} = H(L(G, \mathbf{Q}))\langle \mathbf{N} : \mathbf{G} \rangle,
\]

(1)

where \(\langle \cdot \rangle\) stands for the Macauley bracket which is defined as:

\[
\langle x \rangle = \begin{cases} 
  x & \text{if } x > 0 \\
  0 & \text{if } x \leq 0 
\end{cases}
\]

and \(H\) denotes a scalar function of the state variables enforcing the defining property of an inelastic state. Accordingly, the value of \(H\) must be positive at any inelastic state and zero at any elastic one. Finally, \(L\) denotes a non-vanishing (tensorial) function of the state variables, which is associated with the properties of the phase transformation and \(\mathbf{N}\) is the outward normal to the loading surface at the current state, while the symbol \(:\) between two second order tensors denotes their double contraction. Furthermore, the set of the material states defined as \(H = H(G, \mathbf{Q}) = 0\), which comprises all the elastic states, is called the elastic domain and its projection on the set defined by \(\mathbf{Q} = \text{constant}\) is defined at the elastic domain at \(\mathbf{Q}\). In general, the elastic domain at \(\mathbf{Q}\) is a subset of the elastic range (Lubliner [4]). The particular case where the two sets coincide corresponds to classical plasticity and the
boundary of the elastic domain, that is the initial loading surface, constitutes the yield surface (see Eisenberg and Phillips [5], Lubliner [4], Panoskaltsis et al. [6]).

It is emphasized that Equation (1) has been derived under the assumption of a smooth loading surface at the current strain – temperature point, which implies that only one loading mechanism can be considered. On the other hand, the phase transformations include multiple and sometimes interacting loading mechanisms, which may result in the appearance of a vertex or a corner at the current strain – temperature point. This fact calls for an appropriate modification of the rate Equation (1).

For this purpose we assume that the loading surfaces are defined in the state space by a number – say n – of smooth surfaces, which are defined by expressions of the form:

\[ \Phi_i(G, Q) = 0, \quad i = 1, 2, \ldots, n \]  

(2)

These surfaces can be either disjoint, or intersect in a possibly non-smooth fashion. Each of these surfaces is associated with a particular transformation mechanism which may be active at the current strain – temperature point. Then by assuming that each equation \( \Phi_i(G, Q) = 0 \) defines independent (non-redundant) active surfaces at the current stress temperature point the rate equations for the evolution of the internal variables in view of Equation (1) can be stated in the following general form:

\[ Q = \sum_{i=1}^{n} H_i L_i (G, Q) \langle N_i \cdot G \rangle, \]  

(3)

where \( H_i, L_i \) and \( N_i \) are functions of the state variables defined as in Equation (1) and each set of them – defined by the index \( i \) – refers to the specific transformation associated with the part of the loading surface defined by \( \Phi_i(G, Q) = 0 \). From Equation (3) one can deduce
directly the loading – unloading criteria for the proposed formulation as follows: Lets denote
\( n_{\text{adm}} \leq n \) the number of loading surfaces that may be active at an inelastic state i.e., \( H_i > 0 \),
and lets denote by \( J_{\text{adm}} \) the set of \( n_{\text{adm}} \) indices associated with those surfaces, i.e.,
\[
J_{\text{adm}} = \{ \alpha \in \{1,2,\ldots,n\}/H_\alpha > 0 \}.
\]
Then Equation (3) implies the following loading – unloading conditions:

If \( J_{\text{adm}} = \emptyset \), then \( \dot{Q} = 0 \).
If \( J_{\text{adm}} \neq \emptyset \), then:

i. If \( N_\alpha : \dot{\mathbf{G}} \leq 0 \) for all \( \alpha \in J_{\text{adm}} \) then \( \dot{Q} = 0 \),

ii. If \( N_\alpha : \dot{\mathbf{G}} > 0 \) for at least one \( \alpha \in J_{\text{adm}} \) then \( \dot{Q} \neq 0 \).

So, if we denote further by \( n_{\text{act}} \leq n_{\text{adm}} \) the number of parts for which (ii) holds, and we set:
\[
J_{\text{act}} = \{ \alpha \in J_{\text{adm}}/N_\alpha : \dot{\mathbf{G}} > 0 \},
\]
the loading criteria in terms of the sets \( J_{\text{adm}} \) and \( J_{\text{act}} \) may be stated as:

\[
\begin{align*}
\text{If } J_{\text{adm}} = \emptyset : & \quad \text{elastic state.} \\
\text{If } J_{\text{adm}} \neq \emptyset \text{ and } J_{\text{act}} = \emptyset : & \\
& \quad \text{i. If } N_\alpha : \dot{\mathbf{G}} < 0 \text{ for all } \alpha \in J_{\text{adm}} : \quad \text{elastic unloading,} \\
& \quad \text{ii. If } N_\alpha : \dot{\mathbf{G}} = 0 \text{ for at least one } \alpha \in J_{\text{adm}} : \quad \text{neutral loading,} \\
\text{If } J_{\text{adm}} \neq \emptyset \text{ and } J_{\text{act}} \neq \emptyset : & \\
& \quad \text{inelastic loading.}
\end{align*}
\] (4)

2.2 Equivalent spatial formulation

The equivalent assessment of the governing equations in the spatial configuration can be
done on the basis of a push – forward operation (e.g., see Marsden and Hughes [5], pp. 67 –
68) to the basic equations. Hence, by performing a push – forward operation onto Equation
(3) the latter is written in the form:
\[ L_v q = \sum_{i=1}^{n} h_i (g, q, F) \langle r_i \rangle, \]  

(5)

where \( F \) stands for the deformation gradient and \( g \) denotes the vector of the controllable variables in the spatial configuration. These are, the Almansi strain tensor \( e \) – defined as the push-forward of the Green – St. Venant strain tensor – and the temperature \( T \). Moreover, in Equation (5) \( q \) stands for the push forward of the internal variable vector \( Q \) and \( L_v (\cdot) \) denotes the Lie derivative of its argument (e.g., see Marsden and Hughes [5], pp. 93 – 104), defined as the convected derivative relative to the spatial configuration. The use of Lie derivatives guarantees the objectivity of rate equations in the current configuration (Marsden and Hughes [5], p. 99). Finally, \( h_i \) denotes the expression of the scalar invariant functions \( H_i \) in terms of the spatial variables \( (e, T, q) \) and the deformation gradient \( F \), while \( l_i \) denotes the push-forward of the tensorial functions \( L_i \), and \( r_i \) stands for the (scalar invariant) loading rates which are written in the form:

\[ r_i = \frac{\partial \varphi_i}{\partial e} : L_v e + \frac{\partial \varphi_i}{\partial T} T, \]  

(6)

where \( \varphi_i \) is the expression for the loading surface associated with the index \( i \) in terms of the spatial variables. The (spatial) loading – unloading criteria follow naturally from Equations (5) as:

\[
\begin{align*}
\text{If } j_{\text{adm}} &= \emptyset: & \text{elastic state.} \\
\text{If } j_{\text{adm}} \neq \emptyset \text{ and } j_{\text{act}} = \emptyset: & \\
\begin{cases}
\text{i. If } r_{\alpha} < 0 \text{ for all } \alpha \in j_{\text{adm}}: & \text{elastic unloading,} \\
\text{ii. If } r_{\alpha} = 0 \text{ for at least one } \alpha \in j_{\text{adm}}: & \text{neutral loading,} \\
\end{cases} \\
\text{If } j_{\text{adm}} \neq \emptyset \text{ and } j_{\text{act}} \neq \emptyset: & \text{inelastic loading.}
\end{align*}
\]  

(7)
where the sets $j_{adm}$ and $j_{act}$ are now defined in terms of the spatial variables as:

$$j_{adm} = \{ \alpha \in \{1,2,...,n\} / \eta_\alpha > 0 \}$$

and

$$j_{act} = \{ \alpha \in J_{adm} / r_\alpha > 0 \}.$$

### 3 STRESS AND RATE OF DEFORMATION DECOMPOSITIONS

In this section, we derive decompositions of the rate of the Kirchhoff stress tensor and of the rate of deformation tensor into mechanical and thermal parts. The approach presented herein extends the work of Marsden, Hughes and Pister, which has been carried out within the theory of finite deformation nonlinear elasticity (see Marsden and Hughes [5], pp. 204 – 206). Their work has itself generalized the so called “Duhamel – Neumann hypothesis” for linearized (infinitesimal) elasticity (Sokolnikoff [9], p. 359). Our derivations are for inelastic materials with internal variables and in particular for shape memory alloys, within the framework of finite strains and rotations. Our results are given by Theorems 1 and 2. In Theorem 1 the Kirchhoff stress tensor is decomposed in terms of the rate of deformation tensor and the rate of temperature. Our proof is based on a manipulation of the constitutive equation for the Kirchhoff stress tensor $\tau$, which is defined as:

$$\tau = \rho_{ext} \frac{\partial \psi}{\partial e},$$

where $\psi$ is the free energy in spatial coordinates. The Kirchhoff stress tensor is related to Cauchy stress $\sigma$ by $\tau = J \sigma$, where $J$ is the determinant of the deformation gradient $F$, and can be also thought of as the push – forward onto the spatial configuration of the second Piola
Kirchhoff stress tensor. In our derivation we will employ the concept of Lie derivative, a Legendre transformation and the rate Equations (5). The procedure is the following:

By considering an inelastic process and taking Lie derivatives of both members of Equation (8) and noting that the material time derivative of the mass density in the reference configuration is zero, \( \rho_{\text{ref}} = 0 \), we derive:

\[
L_\nu \tau = \rho_{\text{ref}} \left( \frac{\partial^2 \psi}{\partial \varepsilon^2} : L_\nu e + \frac{\partial^2 \psi}{\partial \varepsilon \partial T} \partial_{\varepsilon} L_\nu e + \frac{\partial^2 \psi}{\partial \varepsilon \partial q} : L_\nu q \right),
\]

which, on substituting for the Lie derivative of \( q \) from Equation (5) yields:

\[
L_\nu \tau = \rho_{\text{ref}} \left[ \frac{\partial^2 \psi}{\partial \varepsilon^2} : L_\nu e + \frac{\partial^2 \psi}{\partial \varepsilon \partial T} : \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial \varepsilon} : L_\nu e \right] =
= \rho_{\text{ref}} \left[ \frac{\partial^2 \psi}{\partial \varepsilon^2} + \frac{\partial^2 \psi}{\partial \varepsilon \partial q} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial \varepsilon} : L_\nu e + \frac{\partial^2 \psi}{\partial \varepsilon \partial T} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial T} \right].
\]

Note, that the double contraction between a fourth and a second order tensor, denoted by the symbol \( : \), produces a second order tensor. We now define the tangential stiffness (a fourth order tensor) and the thermal stress coefficient tensor (a second order tensor) as

\[
a = \frac{\partial^2 \psi}{\partial \varepsilon^2} + \frac{\partial^2 \psi}{\partial \varepsilon \partial q} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial \varepsilon}, \\
m = \frac{\partial^2 \psi}{\partial \varepsilon \partial T} + \frac{\partial^2 \psi}{\partial \varepsilon \partial q} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial T},
\]

respectively. It is instructive and helpful to give here their indicial expression also:

\[
a_{abcd} = \frac{\partial^2 \psi}{\partial \varepsilon_{ab} \partial \varepsilon_{cd}} + \frac{\partial^2 \psi}{\partial \varepsilon_{ab} \partial q_{\eta \ldots \tau}} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial \varepsilon_{ab}}, \\
m_{ab} = \frac{\partial^2 \psi}{\partial \varepsilon_{ab} \partial T} + \frac{\partial^2 \psi}{\partial \varepsilon_{ab} \partial q_{\eta \ldots \tau}} \sum_{i=1}^{n} h_i(\varepsilon,T,q,F) \frac{\partial \varphi_i}{\partial T}.
\]

By taking into account these definitions and by noting that the Lie derivative of the Almansi strain tensor equals the rate of deformation tensor \( d \) (e.g., see Holzapfel [7], p. 107), we can state the following theorem:
**Theorem 1:** For the rate–independent SMA material with internal variables, whose evolution in the course of martensitic transformations is described by the rate Equations (5) (or equivalently by Equations (3)), the thermomechanical constitutive equation relating the (objective) rate of the Kirchhoff stress tensor to the rate of deformation tensor and to the rate of the temperature is given by

\[ -q_{\tau} + \rho_{\text{ref}} (a : d + m) = 0, \]

where the tangential stiffness \( a \) and the thermal coefficients \( m \) are dependent on the martensitic phase transformation properties, through their dependence on the internal variables.

We will now obtain the counterpart of Equation (13), i.e., an expression for the rate of deformation tensor in terms of the rate of the Kirchhoff stress tensor and the temperature rate. This will be accomplished as follows. First, we define the complementary free energy, by using the Legendre transformation technique, as:

\[ \chi = \frac{1}{\rho_{\text{ref}}} (\tau : e) - \psi(e, T, q). \]

We assume that the change of variables from \( e \) to \( \tau \) is invertible and that the evolution of the internal variables can be given in the stress space in the course of inelastic loading by a rate equation of the form

\[ L_\chi = \sum_{i=1}^n t_i c_i (\tau, T, q, F) [x_i (\tau, T, q, F), L_\chi, \tau + y_i (\tau, T, q, F)^T], \]

where the functions \( t_i, c_i \) have an analogous meaning with the functions \( h_i, l_i \) in the rate Equations (5) and \( x_i, y_i \) are tensorial functions of the (new) state variables. By taking partial derivatives in Equation (14) with respect to the Kirchhoff stress we obtain:
\[ \rho_{\text{ref}} \frac{\partial \hat{\chi}}{\partial \tau} = -\frac{\partial e}{\partial \tau} - \rho_{\text{ref}} \frac{\partial \psi_{e}}{\partial \tau}, \]  

(16)

which in light of the constitutive equation for the Kirchhoff stress tensor (Equation (8)) yields the constitutive relation for the Almansi strain tensor as a function of the Kirchhoff stress tensor:

\[ e = \rho_{\text{ref}} \frac{\partial \hat{\chi}}{\partial \tau}. \]  

(17)

By taking the Lie derivative of both hand sides in Equation (17) we obtain

\[ L_{\chi} e = \rho_{\text{ref}} \left( \frac{\partial^2 \chi}{\partial \tau^2} : L_{\chi} \tau + \frac{\partial^2 \chi}{\partial \tau \partial T} T + \frac{\partial^2 \chi}{\partial \bar{\tau} \partial \bar{q}} : L_{\chi} \bar{q} \right) = \]  

(18)

which in turn, upon substitution of the rate Equation (15) takes the form

\[ L_{\chi} e = \rho_{\text{ref}} \left( \frac{\partial^2 \chi}{\partial \tau^2} : L_{\chi} \tau + \frac{\partial^2 \chi}{\partial \tau \partial T} T + \frac{\partial^2 \chi}{\partial \bar{\tau} \partial \bar{q}} : L_{\chi} \bar{q} \right) = \]  

(19)

We now define the fourth and second order material compliance tensors \( \mathbf{r} \) and \( \mathbf{s} \) as

\[ \mathbf{r} = \frac{\partial^2 \chi}{\partial \tau^2} + \frac{\partial^2 \chi}{\partial \bar{\tau} \partial \bar{q}} \sum_{i=1}^{n} l \mathbf{c}_i (\tau, T, q, F) \mathbf{x}_i (\tau, T, q, F) : L_{\chi} \tau, \]  

(20)

\[ \mathbf{s} = \frac{\partial^2 \chi}{\partial \tau \partial T} + \frac{\partial^2 \chi}{\partial \bar{\tau} \partial \bar{q}} \sum_{i=1}^{n} l \mathbf{c}_i (\tau, T, q, F) y_i (\tau, T, q, F), \]  

(21)

respectively, with components

\[ r_{abcd} = \frac{\partial^2 \chi}{\partial \tau_{ab} \partial \tau_{cd}} + \frac{\partial^2 \chi}{\partial \bar{\tau}_{ab} \partial q_{f_{1}...n_{t}}} x_{ic},  

s_{abcd} = \frac{\partial^2 \chi}{\partial \tau_{ab} \partial T} + \frac{\partial^2 \chi}{\partial \bar{\tau}_{ab} \partial q_{f_{1}...n_{t}}} y_{i}. \]  

(21)
Therefore, and since the Lie derivative of $\mathbf{e}$ is equal to $\mathbf{r}$, Equation (19) with the help of Equations (20) yields for the rate of deformation tensor

$$\mathbf{d} = \rho_{\text{ref}} (\mathbf{r} : \mathbf{L} \mathbf{r} + \mathbf{s}^T).$$

We have therefore proved the following theorem:

**Theorem 2:** For a rate – independent shape memory alloy material with internal variables, whose evolution in the course of martensitic transformations is described in strain – space and in the current configuration by the rate Equations (5) (or equivalently in the reference configuration by Equations (3)), its rate of deformation tensor can be decomposed in terms of the Lie derivative of its stress tensor and the rate of its temperature, according to Equation (22).

### 4 CONCLUSIONS

The formulation of a finite strain multi surface generalized plasticity theory in strain space and in referential and current configurations for the description of phase transformations, as well as the important loading-unloading criteria have been reviewed. The main thrust of this work is the derivation of a decomposition of the rate of the Kirchhoff stress tensor and of the rate of deformation tensor into mechanical and thermal rates, for materials with internal variables, for arbitrary strains and rotations and for non isothermal conditions.

### REFERENCES


Determination of 3D Wind Induced Vibration of Cables for Cable–Stayed Bridges

ALEJANDRO HERNÁNDEZ* AND JESÚS G. VALDÉS†

* Departamento de Ingeniería Civil, División de Ingenierías, Campus Guanajuato, Sede Belén
Universidad de Guanajuato
Av. Juárez 77, Zona Centro, Guanajuato, Gto., Mexico
e–mail: alejandro.hernandez@ugto.mx, www.ugto.mx

† Departamento de Ingeniería Civil, División de Ingenierías, Campus Guanajuato, Sede Belén
Universidad de Guanajuato
Av. Juárez 77, Zona Centro, Guanajuato, Gto., Mexico
e–mail: valdes@ugto.mx, www.ugto.mx

Key words: Cable–stayed bridge, wind–structure interaction, Vibration, Aeroelastic, Resonance.

Abstract. This paper is focused in the analysis of cable stays for Viaducto Zaragoza Bridge located in the city of Puebla, Mexico. Computational fluid dynamics based on finite element method is used to simulate wind forces acting on the cable stays, which are coupled to a structural model of each model analyzed for two wind velocities, associated with serviceability limit state (50 km/hr or 14 m/s) and maximum wind velocity (140 km/hr or 14 m/s). Thus, a stabilized fluid flow formulation is presented to solve an ALE fluid flow while a geometrically non–linear solid elements are used to model the cable stays. Both solutions are coupled using a strong coupling technique to perform an aeroelastic analysis of the cable stays. The results suggest that transversal vibration to wind action can generate undesirable vibrations conditions as the resonance phenomenon in cable stays.

1 INTRODUCTION

Cable–stayed bridges are a highly nonlinear structural system where superstructure deck is supported on several points of its length by cables anchored directly to a support column. Nowadays different cable geometries are commonly used to transmit loads between deck and the pylon (or support column). The most typical cable geometries are shown in Figure 1.

The concept of cable–stayed bridges can be traced to the XVII century [1], the economic viability of this kind of bridges had to wait the development of two aspects: (a) Decks constructed of steel and concrete, and (b) High resistance steel cables.

While the materials used in structural components of these structures are under linear–elastic behaviour for normal operation conditions, the overall performance of the entire structure is often highly nonlinear due to: (i) Nonlinear behaviour of cable stays for axial load–strain relationship caused by its self–weight; (ii) Nonlinear relationship for elements under axial load – bending; and (iii) Large displacements in the structure even for normal
operation loads.

Figure 1 Typical geometries for cable–stayed bridges

1.1 General Bridge Description

The geometry employed in the Viaducto Zaragoza Bridge requires a system of inclined and horizontal cables to support the superstructure deck to the inclined pylon or mast as can be shown in Figure 2. Table 1 shows normal operating conditions for each cable stay of the bridge

Wind actions are estimated according local code requirements specified in [2]. According the characteristics of the project, cable stays analyses must be performed for two wind velocities:
1) An average wind velocity \( V_{\text{med}} = 14 \text{ m/s} \) representing the average operating conditions of the bridge
2) Maximum probable wind velocity \( V_{\text{max}} = 40 \text{ m/s} \) representing the maximum wind speed expected for the operations of the bridge.
Table 1 Cable stays characteristics for normal operation conditions

<table>
<thead>
<tr>
<th>Designation</th>
<th>Length (m)</th>
<th>Strands</th>
<th>Weight (kN)</th>
<th>Tensioning Force (MN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01N</td>
<td>17.732</td>
<td>28</td>
<td>12.66</td>
<td>784.5</td>
</tr>
<tr>
<td>T01S</td>
<td>17.639</td>
<td>28</td>
<td>12.59</td>
<td>784.5</td>
</tr>
<tr>
<td>T02N</td>
<td>23.440</td>
<td>48</td>
<td>28.68</td>
<td>431.5</td>
</tr>
<tr>
<td>T02S</td>
<td>23.205</td>
<td>48</td>
<td>28.40</td>
<td>451.1</td>
</tr>
<tr>
<td>T03N</td>
<td>32.331</td>
<td>48</td>
<td>39.57</td>
<td>725.6</td>
</tr>
<tr>
<td>T03S</td>
<td>31.934</td>
<td>48</td>
<td>39.08</td>
<td>745.3</td>
</tr>
<tr>
<td>T04N</td>
<td>42.518</td>
<td>48</td>
<td>52.03</td>
<td>1078.7</td>
</tr>
<tr>
<td>T04S</td>
<td>41.854</td>
<td>48</td>
<td>51.22</td>
<td>1304.2</td>
</tr>
<tr>
<td>T05N</td>
<td>53.072</td>
<td>34</td>
<td>46.01</td>
<td>1588.6</td>
</tr>
<tr>
<td>T05S</td>
<td>52.214</td>
<td>34</td>
<td>45.26</td>
<td>1588.6</td>
</tr>
<tr>
<td>H01</td>
<td>39.049</td>
<td>39</td>
<td>77.65</td>
<td>1314.0</td>
</tr>
<tr>
<td>H02</td>
<td>45.362</td>
<td>39</td>
<td>90.21</td>
<td>990.4</td>
</tr>
<tr>
<td>H03</td>
<td>51.145</td>
<td>39</td>
<td>101.71</td>
<td>2510.3</td>
</tr>
</tbody>
</table>

2 FINITE ELEMENT ANALYSIS OF THE WIND ACTION

An incompressible fluid formulation has been used to simulate wind action due to fluid velocity is lower than 0.3 Mach. Navier–Stokes equations are used to model flow as shown in Eq. (1).

\[
M \ddot{v}_{n+1} + KV_{n+1} - Gp_{n+1} = f_{ext}^{n+1}, G^T \dot{v}_{n+1} = 0
\]  

(1)

where:
- \(v\) = Velocity
- \(p\) = Pressure field
- \(\ddot{v}\) = Acceleration
- \(M\) = Mass matrix
- \(K\) = Matrix with convective and viscous terms
- \(G\) = Matrix to include pressure terms or to consider a compressible flow

For dynamic fluid flow analysis Eq. (1) is rewritten by Gunzburger [3] as shown in Eq. (2). To perform a faster calculation, Eq. (2) is decoupled using fractional step method proposed by Codina [4] and considering the equation as complete Eulerian formulation, expression is transformed using and ALE formulation, as can be found in Belytschko et al. [5] to take into account the structure movement in the domain of analysis.

\[
\begin{bmatrix}
(v_{n+1}^{p+1}, w_{n+1}^{p+1}) + \frac{c}{\delta t} (v_{n+1}^{p+1}, w_{n+1}^{p+1}) - b(p_{n+1}^{p+1}, w_{n+1}^{p+1}) + a(v_{n+1}^{p+1}, w_{n+1}^{p+1}) = \left(b_{n+1}^{p+1}, w_{n+1}^{p+1}\right) \\
b(q_{n+1}, v_{n+1}^{p+1}) = 0
\end{bmatrix}
\]  

(2)

Decoupled and stabilized equations are expressed as can be shown in Eq.(3), which are formulated in four implicit steps for each time increment. The first step is to solve the system at an intermediate velocity, which is a nonlinear formulation. The final pressure is computed...
in the second step. In the third step the final velocity is calculated and the complete system is stabilized in the fourth step.

\[
\left( \mathbf{\dot{u}}_{\text{ext}^{n-1}} \cdot \mathbf{w}_n \right) + c \left( \mathbf{\ddot{e}}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{v}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{w}_n \right) = -b \left( \mathbf{p}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{w}_n \right) + a \left( \mathbf{\mathbf{\dot{v}}}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{v}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{w}_n \right) + \tau \left( \mathbf{\mathbf{\dot{e}}}_{\text{ext}^{n-1}}^{n-1} \cdot \nabla \mathbf{v}_{\text{ext}^{n-1}}^{n-1} \cdot \mathbf{v}_{\text{ext}^{n-1}}^{n-1} \cdot \nabla \mathbf{w}_n \right) = \left( \mathbf{u}_{\text{ext}^{n+1}}^{n+1} \cdot \mathbf{w}_n \right)
\]

\[
-\frac{\Delta t}{\alpha} \left( \nabla \left[ \mathbf{p}_{\text{ext}^{n+1}}^{n+1} - \mathbf{p}_{\text{ext}^{n+1}}^{n+1} \right] \cdot \nabla \mathbf{q}_n \right) - \tau \left( \mathbf{\mathbf{\dot{e}}}_{\text{ext}^{n+1}}^{n+1} \cdot \nabla \mathbf{v}_{\text{ext}^{n+1}}^{n+1} \cdot \mathbf{v}_{\text{ext}^{n+1}}^{n+1} \cdot \nabla \mathbf{w}_n \right) = \mathbf{b} \left( \mathbf{q}_n, \mathbf{v}_{\text{ext}^{n+1}}^{n+1} \right)
\]

(3)

Figure 3 shows a mesh used to model wind for the analysis of cable H01. Figure 3(b) shows that finite elements near to the contour are smaller to model the boundary layer.

3 FINITE ELEMENT ANALYSIS OF THE CABLE STAYS

To estimate the behavior of the cable stays a geometrically nonlinear model of solid is used. The properties of solids for each cable stay is estimated from the real properties showed in Table 1, considering the accessories elements for strands protection (as shown if Figure 4) only provide mass, thus, the cable stays are modeled with solids with the same equivalent properties to all elements.

The expression that describes the behavior of cable stays are obtained from the equation of lineal momentum, and discretized using FEM. The computations consist of solve the following expression

\[
f^{\text{int}} \left( \mathbf{u}_{\text{ext}} \right) + \mathbf{M} \mathbf{u}_{\text{ext}} = \mathbf{f}^{\text{ext}} \left( \mathbf{u}_{\text{ext}} \right)
\]

(4)

where:

- \( \mathbf{f}^{\text{ext}} \) = Internal forces
To solve the dynamics of cable stays in time Generalized–α method is used to integrate Eq. (4) due to other traditional methods like β–Newmark or θ–Wilson produce inconsistent results with nonlinear finite element formulation proposed.

One important issue of this work is the modeling of the cable stays tension for the elements. Considering that the spatial position and tension force for each cable stay are known, the one-dimensional stress is transformed in a tensional stress in a 3D solid according to the reference system used for analysis. Thereafter, internal forces due to tension stress are added to the ordinary 3D stress tensor.

Furthermore, to reduce the memory required to solve the equation system, meshes for each cable stay is always horizontal or vertical to coincide with the global reference system. This consideration helps to the development of mesh analysis but leads the problem of cable self-weight. To solve this problem gravity is considered as a unit vector that can be oriented in any direction making the cables have the correct deformation profile according its real position in the bridge.

4  WIND–CABLE STAY INTERACTION ANALYSIS

The cable stay and the fluid are solved in a domain containing both models, as can be shown in Figure 3, solving both problems in a coupled way, as occur in real world.

To solve both systems, a partitioned approach is employed, i.e., for each time step behavior of cable stay and fluid are computed independently using Aitken schemes (Wüncher, [6]) to ensure convergence of both coupled systems. This methodology has been employed by Valdés [7] given excellent results for aeroelastic problems of several structures, such as those studied by Valdés et al. [8] and Hernández and Valdés [9].

The procedure for the solution of wind–cable stay interaction is as follows:

1. Solve the cable stay to predict the displacements according the acting external forces. Initial step consider tension force and gravity only.
2. Obtained displacements of cable–stays are passed to the fluid mesh, adjusting the mesh to match with deformed profile of the cable.
3. Wind dynamics is solved to estimate acting forces on cable–stay surface.
4. Updating fluid forces acting on cable–stay surface.

This procedure is repeated until convergence criteria is reached for each time step. Detailed information for above procedure can be found in [7].
5 OBTAINED RESULTS

Figure 5 shows pressure distribution for certain time step at half-length of cable stay T03N. Both figures shown suction due to vortices generation that induces transversal vibration on cable stays respect to wind action.

Vortices are not statically and “move” through the cable length, as can be seen in Figure 6 which shows pressure distribution in a cross section along the entire length of cable stay T03N for both analysis conditions.

Figure 5 Transversal views for pressure distribution at half length of cable stay T03N

Figure 6 Longitudinal views of pressure distribution for cable stay T03N
Variation of the acting forces on the cable length in conjunction with operational conditions for each case cause the vibration of cable stay with certain profile deformation and vibrational frequency.

As can be noted in Figure 7 the displacements in transversal direction to wind action dominate the total displacement at this position, making evident that the cable has a typical resonant behavior. Furthermore, the displacements shown in Figure 8 suggest that transverse displacements do not dominate the total displacements, however, is evident that transversal displacements increases with time, but their influence on the total displacements is minimal. It is noteworthy that in the presented case, there is a greater potential that resonance phenomena occurs at wind velocity less than the maximum analyzed, which is more probably to occur in bridge life.

Obtained results suggest that the most critical conditions for cable stays are transversal vibration respect to wind action, and is not necessary have a high wind speeds to induce resonance in cable stays, which may lead in fatigue on the strands or their anchorage systems.

6 CONCLUSIONS

To determinate displacements and forces acting on a cable–stayed bridge, traditional analysis methods are not recommended to use due to highly geometrically nonlinear behavior of this kind of structures. In order to improve the functionality and security for cases like above mentioned, techniques that can predict the behavior more accurately are needed. The presented FEM application permits to visualize vibrations characteristics for cable stays and determinate if they are susceptible to undesirable frequencies, as those associated with
resonance phenomena.

Figure 8 Displacements history at half length of T03N cable for maximum wind velocity (40 m/s)

Obtained results suggest that transversal vibration respect to wind action is very important to determinate safety operation condition of the cable stays, currently most of methodologies which allow the determination vibration periods only can predict frequencies in the direction of wind action, due to the models are considered in 2D (Au et al. [10], Starossek [11] among others) making impossible to determinate the vibrations characteristics in transverse direction to wind action.

For presented case of Viaducto Zaragoza Bridge, resonance phenomena is identified for some cable stays, viewing the need to change their operations conditions by adding frictional dampers at deck anchorages. These devices modify the natural vibration frequencies of the cables due to added mass of the device, and reduce the displacements by the damper, decreasing the probability of occurrence of resonance.

7 ACKNOWLEDGMENTS

The authors acknowledge the support the University of Guanajuato for the work development presented in this paper.

8 REFERENCES


Electricidad CFE, Mexico, (2008).


MESH DEFORMATION WITH EXACT SURFACE RECONSTRUCTION USING A REDUCED RADIAL BASIS FUNCTION APPROACH

DANIEL S. C. KOWOLLIK, MATTHIAS C. HAUPT AND PETER HORST

Institute of Aircraft Design and Lightweight Structures (IFL)
TU Braunschweig
Herrmann-Blenk-Str. 35, 38106 Braunschweig, Germany
e-mail: d.kowollik@tu-braunschweig.de, web page: http://www.tu-braunschweig.de/ifl

Key words: Mesh deformation, Mesh motion, Radial basis functions, Delaunay graph mapping, Fluidstructure interaction

Abstract. This paper presents a novel reduced radial basis function approach with exact surface reconstruction. The new approach combines two well proven mesh deformation algorithms in a three step approach. In a first pre-processing step an explicit reduction of radial basis function points is performed using a k-d tree. In the second step the classic radial basis function interpolation is used to propagate the deformation field. In a last post-processing step an exact surface reconstruction is achieved using an efficient Delaunay graph mapping approach. The new mesh deformation approach is compared to the two original approaches by investigating a 2D viscous mesh test case. The applicability of the new approach to 3D is shown via an aeroelastic relevant wing test case.

1 INTRODUCTION

The demand of mesh deformation algorithms for computational fluid dynamic (CFD) meshes arises for instance in shape optimization analyses or in fluid-structure interaction (FSI) analyses. During FSI analyses for each iteration and time step a CFD mesh deformation is required in order to follow the structural response. The interest for more detailed analyses and complex configurations in the aeronautical industry such as the analyses of complete aircrafts including slats, pylons and flaps increase the computational cost substantially for each involved algorithm. A transient FSI analysis may require several hundred of mesh deformations. Therefore, a robust and efficient technique is required.

The published algorithms can be categorized into two groups, the once that are point schemes and the once that are connectivity schemes.

Connectivity Schemes are mainly dominated by structural analogy approaches. Either the edges of the cells or the cells directly are treated as a mechanical model such as springs
or plates. A structural finite element model can be formulated and a linear sparse system needs to be solved. A classic approach is the spring analogy presented by Batina [1], which treats the cell edges as linear springs, where the stiffness of the springs is computed by the inverse of the individual edge length. This approach can lead to degenerated cells if one pair of edges snaps through the other edges of the same cell. To minimise this problem Farhat et al. [2] developed the torsional spring method. In addition to the classic approach torsional springs are added to the corners of the cells to prevent the cells from colliding. Continuum elastic approaches have been suggested by Lynch et al. [3] and improved by Stein et al. [4]. The common feature of these approaches is an increasing stiffness with decreasing cell size. This may result in a stiff system matrix which often allows only small incremental deformations. Furthermore, these methods are difficult to implement for arbitrary mesh types especially for meshes with hanging nodes.

Point schemes directly support arbitrary mesh types because an interpolation algorithm is used which propagates the boundary movements without taking the mesh connectivity into account. An efficient algorithm developed by Liu et al. [10] is based on a coarse Delaunay mesh which is generated using the boundary and far field points. The deformations are propagated with the Delaunay graph through a linear mapping using barycentric coordinates. The drawback of this approach is a strong shearing of orthogonal aligned cells near the boundary if large deformations are present. Mesh deformation based on radial basis functions (RBFs) has been studied by several authors [5, 6, 7, 8] and has shown to generate smooth mesh deformations. If an appropriate RBF is used, a good orthogonal behaviour near the structured boundary layer is achieved. The drawback of the original form of the RBF approach is the large linear dense system that needs to be solved if a large amount of source deformation points is present. RBFs with compact support presented by Wendland [9] allow the reduction of the system to a sparse one. If the radius is small, only small deformations can be incorporated inside the volume mesh which limits this reduction technique. Consequently, RBF mesh deformation is only applicable to large meshes if a reduction of source points is performed. This reduction results in an inexact surface deformation which might lead to unsatisfactory results if too many source points are eliminated. In order to minimise this effect Rendall et al. [8] proposed different Greedy algorithms to make an optimal choice of a source point reduction. Subsequent to the RBF deformation a correction function is used to account for the dropped surface points. The method is more efficient compared to the classic RBF deformation and generates good results, however during a transient FSI analysis the optimization problem has to be solved at least once if the deformations are small or more often if the deformation variation is large. In this paper we present an alternative explicit reduced RBF deformation approach with exact surface reconstruction based on the Delaunay graph mapping (DGM) approach developed by Liu et al [10]. In the following we present the numerical approach and use a 2D aeroelastic relevant test case to analyse and discuss the quality of the new algorithm compared to the original RBF and DGM approach. Last, we demonstrate the applicability of the new approach to a 3D wing test case.
2 NUMERICAL APPROACH

Our hybrid approach combines two well proven mesh deformation techniques in a three-step approach:

I. The first pre-processing step is needed to perform a reduction of deformation source points in order to achieve a small limited amount of source points for the following RBF deformation step. Different methods are possible to reduce the number of source points. In this study we suggest an explicit reduction technique using a k-d tree. A reduced number of source points are extracted from each leaf to determine a good representation of the provided source points.

II. The second RBF deformation step uses the reduced number of source points to induce the reduced deformation field to the volume mesh points, which results in an inexact surface deformation.

III. In a third post-processing step a remaining offset deformation vector between the initial surface deformation and the resulting reduced RBF deformation is computed. This offset vector is propagated to the volume mesh points by using the DGM approach in order to reconstruct the exact surface.

2.1 Source point reduction technique

An explicit source point reduction technique is developed, which is not connected to the current deformation field. The advantage of this approach is a so called black box method, which can be applied to the initial undeformed mesh at the beginning of an FSI analyses. The information of the reduced number of source points can be stored and reused in each deformation step.

A lot of different approaches are possible. For instance, one could use the connectivity of the boundary mesh and use the generated multigrid meshes of the CFD analyses code. We suggest a point based approach based on a k-d tree. This approach holds on to the requirement that an entire point based approach is desired, which makes the overall approach suitable to a more generic application field.

The present k-d tree is generated using the open source library vtk [13], which allows a broad direct access to the generated k-d tree data structure. The k-d tree algorithm is not used in the original sense that the access time and search time to locate a point is optimized. Moreover, we use the control parameters of the division of spatial block regions in order to generate as many leaf regions as desired. A small region is generated where a local clustering of surface points is located, e.g. the leading edge of a wing or a shock position on the surface. The bounding box of each leaf region is used to construct search points inside the region. These search points are only used to find a close source point (surface point) and therefore generate an explicit generated limited number of source points for the subsequent RBF deformation step.

The source point reduction technique can be subdivided in the following steps:
1. Construction of a k-d tree with minimal and maximal allowable region bounds.
2. Gather the bounding boxes of each generated leaf region.
3. Generate search points for each region using linear hexahedron finite elements.
4. Find the closest mesh point to each search point.
5. Drop the duplicate found mesh points.

The position of generated search points inside each region can be arbitrary as desired. So far we used 1 central point, 8 gauss points \((\pm1/\sqrt{3}, \pm1/\sqrt{3}, \pm1/\sqrt{3})\) or 8 corner points depending on the test case.

Two different generated coarse k-d trees are shown in figure 1. In figure 1(a) one can see the used search points in red (corner points) and the found mesh points in green of the entire boundary mesh point set (yellow triangles). Figure 1(b) shows a coarse k-d tree generated for the 3D wing test case. A finer k-d tree would consequently generate more leaf regions along the leading and trailing edge of the wing.

![Figure 1](image-url)

(a) k-d tree (blue) applied to a simple test case; mesh points (yellow); k-d tree search points (red); reduced RBF points (green).

(b) k-d tree (red) applied to the surface of the wing test case.

2.2 Radial basis function interpolation

RBF based mesh deformation is a point based interpolation approach which propagates the displacement field \(u^s_j\) of a finite set of source points \(x^s_j = [x^s_{1j}, x^s_{2j}, x^s_{3j}]\), \(j = 1, 2, ..., n_s\) using a continuous interpolation function \(c(x^t_i)\):

\[
c(x) = p(x) + \sum_{j=1}^{n_s} \beta_j \varphi \left( \|x - x^s_j\| \right) .
\]

(1)
A smooth interpolation is realized by a linear combination of basis functions \( \phi \) being radial with respect to the Euclidean distance of the source points. For an exact recovery of rigid body motions a linear polynomial \( p(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \) is added in the case of RBF based mesh deformation. The inclusion of the linear polynomial requires the fulfilment of an additional zero condition:

\[
\sum_{j}^{n_s} \beta_j q(x_j^i) = 0 ,
\]

for all polynomials \( q \) with degree \( \text{deg}(q) \leq \text{deg}(p) \). The weights of the linear polynomial \( \alpha_i \) and the RBF’s \( \beta_j \) can be evaluated by solving the linear system

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 1 & 1 & \cdots & 1 \\
0 & 0 & 0 & 0 & x_{1,1}^s & x_{1,2}^s & \cdots & x_{1,n_s}^s \\
0 & 0 & 0 & 0 & x_{2,1}^s & x_{2,2}^s & \cdots & x_{2,n_s}^s \\
0 & 0 & 0 & 0 & x_{3,1}^s & x_{3,2}^s & \cdots & x_{3,n_s}^s \\
1 & x_{1,1}^s & x_{2,1}^s & x_{3,1}^s & 0 & 0 & \cdots & 0 \\
1 & x_{1,2}^s & x_{2,2}^s & x_{3,2}^s & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \cdots & \vdots \\
1 & x_{1,n_s}^s & x_{2,n_s}^s & x_{3,n_s}^s & 0 & 0 & \cdots & 0
\end{bmatrix}
\begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\beta_1 \\
\beta_2 \end{bmatrix} =
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
\phi_{1,1}^s \\
\phi_{1,2}^s \\
\vdots \\
\phi_{n_s,1}^s
\end{bmatrix}
\]

or in matrix form \( c = Aw \) with the symmetric interpolation matrix \( A \). The RBF’s \( \phi_{ij} = \phi(\|x_i^s - x_j^s\|) \) in (3) are evaluated for the Euclidean distance of the source points. The formulation of the linear system (3) holds for a scalar displacement field. The extension to two or three dimensions is a trivial task. Therefore, \( 3 \times n_s \) weights have to be evaluated for a 3D mesh deformation. In the case of global RBF’s a dense system \((4+n_s) \times (4+n_s)\) needs to be solved. Wendland introduced in [11] compact RBF’s where sparsity of the system is controlled via a compact support radius. Smaller support radii generate a system which can be solved more efficiently. However, this counteracts towards an accuracy decrease of the deformation algorithm. Another drawback of small radii is a decrease of maximal allowable deformations in one deformation step. For the original version of the deformation approach the entire boundary points are used as source points. Therefore, the original form of the RBF based mesh deformation algorithm is only practical for a small number of source points.

The evaluated interpolation weights \( \alpha_i \) and \( \beta_j \) can be used in the same linear combination to compute the deformation of the target points \( x_i^t = [x_{1,i}^t, x_{2,i}^t, x_{3,i}^t] \), \( i = 1, 2, \ldots, n_t \):

\[
c(x_i^t) = p(x_i^t) + \sum_{j}^{n_s} \beta_j \phi(\|x_i^t - x_j^s\|) .
\]
studies that the performance of the mesh deformation algorithm depends highly on the chosen RBF.

Using an appropriate reduction technique one can limit the number of source points which limits in equal measure the numerical cost to solve the linear system.

In [6] it was shown that the final deformed mesh can be improved if a large deformation field is subdivided into multiple incremental steps. Therefore, one can conclude that the deformation algorithm is path dependent if more than one step is used for an FSI analyses.

2.3 Surface reconstruction based on Delaunay Graph Mapping

The surface reconstruction technique suggested in this paper can be seen as an additional post-processing mesh deformation step. Large deformations are propagated with a limited number of source points using the presented RBF algorithm. A limited number of source points results in an inexact surface deformation which leads to an accuracy decrease of the overall CFD analyses. In order to eliminate this effect an exact surface reconstruction is desired. One can consider the surface reconstruction as an additional mesh deformation step based on small deformations that have to be propagated inside the volume part of the mesh. We suggest the application of the DGM algorithm proposed by Liu et al. [10], which is a very efficient point based scheme. The remaining deformation to reconstruct the exact moving boundary can be easily evaluated:

\[ u_{DGM}^b = u^b - u_{RBF}^b \] (5)

The DGM mesh deformation algorithm can be divided into four steps:

1. Construction of a Delaunay graph using all boundary points.
2. Locating the volume mesh points inside the Delaunay graph.
3. Prescribing the deformation field to the Delaunay graph.
4. Relocate the volume mesh points inside the deformed Delaunay graph.

A graphical representation of the four necessary sub-steps is given for a simple test case in figure 2. If in step 3 a negative area in 2D or a negative volume in 3D is created for one of the Delaunay cells one can subdivide the deformation field and reiterate the steps one till four until the entire deformation field is propagated. However, in general this is only the case if large deformations are to be propagated with this algorithm.

Several efficient open source libraries exist which allow the construction of a Delaunay graph. For the present study we use the qhull library developed by Barber et al. [12]. A unique linear mapping of the boundary points exists for the constructed coarse mesh. In terms of barycentric coordinates the linear mapping between a field point \( x_P \) and the cell corner points \( x_A, x_B \) and \( x_C \) is evaluated by:
\[ x_P = e_1 \cdot x_A + e_2 \cdot x_B + e_3 \cdot x_C \quad \text{with} \quad e_i = \frac{S_i}{S}, \quad (6) \]

where \( e_i \) are the area ratio coefficients which are illustrated in figure 3 and can be determined using the partial areas \( S_i \) as follows:

**Figure 2**: Sub-steps of the DGM mesh deformation algorithm.

**Figure 3**: Linear mapping using barycentric coordinates.
A search algorithm is needed to associate the field points with the surrounding Delaunay cells. Inside a triangle cell the sum of the area ratio coefficients yields

\[ S = 1 \]

A range restriction \( e_i \in [0, 1] \) for \( i = 1, 2, 3 \) of the area ratio coefficients can be used to determine if a field point lies inside a cell. Either the displacement field of the Delaunay cells can be interpolated for the point \( P \) as follows

\[ u_P = e_1 \cdot u_A + e_2 \cdot u_B + e_3 \cdot u_C \]

or the new point positions of the deformed Delaunay graph can be used to evaluate the new location of a field point:

\[ x_{P,new} = e_1 \cdot x_{A,new} + e_2 \cdot x_{B,new} + e_3 \cdot x_{C,new} \]

An illustration of the linear mapping approach showing an undeformed Delaunay cell and a new deformed Delaunay cell can be seen in figure 3. The extension of this algorithm to 3D can be easily achieved. In 3D a tetrahedralization of the boundary point set needs to be constructed. Barycentric coordinates can be evaluated for tetrahedra cells likewise.

3 RESULTS

The capability of the new mesh deformation approach is investigated by analysing the mesh quality for the rotation of an airfoil. Furthermore, we demonstrate the applicability of the developed approach for a mode shape deformation applied to an aeroelastic relevant 3D test case.

3.1 Airfoil rotation test case: Viscous mesh

In this section the quality of the deformation algorithms RBF and DGM is analysed and compared to the developed reduced RBF deformation algorithm. For viscous calculations a high resolution of the boundary layer is necessary. For accuracy reasons it is desired that the orthogonality of the structured boundary layer is sustained as far as possible. In order to investigate this requirement the rotation of a NACA0015 airfoil C-grid is analysed. This mesh is shown in figure 4 and consists of 39666 points, 39300 cells, 566 far field boundary points and 166 wall boundary points.

The mesh quality is measured in terms of the skewness of the cells. A perfect orthogonal cell has a skewness of zero. The mean skewness of the present mesh is 0.071. This value
is obtained by integrating all cell skewnesses over the entire mesh domain and dividing
the value by the total mesh area. The largest skewness of the undeformed mesh is 0.175.

Using the original RBF and DGM approaches from literature as well as the newly
developed reduced RBF deformation approach, the rotation of the airfoil is investigated.
For simplicity no variation on the type of radial basis functions is analysed. We use the
global thin plate spline, which performed best of the global radial basis functions studied
in [6]. We focus on the analyses of a large reduction of source points used for the RBF
step. The rotation of 15° is computed in three equal steps.

The resulting deformed meshes of the pure DGM deformation and the reduced RBF
deformation are shown in figure 5. The DGM algorithm alone is not capable of sustaining
orthogonality at the interface as can be seen in figure 5(a). Although the developed
reduced RBF deformation makes use of a massive reduction in RBF source points a fairly
good orthogonal behaviour can be drawn from figure 5(b). In total only 10 source points
are used for the RBF deformation step. This is a source point reduction of 98.63%.

A comparison of all analysed algorithms is shown for the average and maximum skew-
ness in figure 6. The DGM approach produces a much better average skewness, but has a
drawback in the maximum skewness. The RBF approach generates the lowest maximum
skewness for each deformation step. The comparison of all studied approaches shows that
the combination of both approaches with a large reduction of RBF source points leads to
a competitive approach for this test case. Even for a reduction to a low number of source

Figure 4: Airfoil NACA0015 test case: Undeformed original mesh.
points, the quality of the deformed mesh lies in the vicinity of the pure RBF approach.

Figure 5: deformed Airfoil due to 15° rotation (parameters of reduced RBF deformation: 1 RBF and 1 Delaunay step; 4 RBF source points of the airfoil; 6 RBF source points of the farfield).

Figure 6: Mesh quality for the airfoil rotation test case (reduced RBF approach: Reduction of source points is only used for the rotating boundary; 6 far field points are used for the RBF deformation step).

3.2 Wing mode shape deformation test case: Euler mesh

In this section the AGARD 445.6 wing is deformed in order to demonstrate the applicability of the developed deformation algorithm to three dimensions. The test case consists of an unstructured euler mesh with 334861 cells and 64943 points. The number of fixed boundary points is 923 and the number of moving boundary points is 16285. The mean aspect ratio of the undeformed mesh is 1.546 and the maximum aspect ratio is 7.223. The aspect ratio of a perfect tetrahedra is exactly one.

The first mode shape is used as prescribed as a large deformation field. Parts of the final deformed volume mesh are shown in figure 7. In figure 7(a) the undeformed surface mesh and the deformed surface mesh are coloured in red and green, respectively. A close view of
the deformed volume mesh at the trailing edge of the wing tip can be seen in figure 7(b).
For this large deformation a reduction of 98.2% wing surface source points is used. The
deformation is subdivided in one reduced RBF step and three DGM reconstruction steps.
The mean aspect ratio of the deformed mesh is 1.548 and the maximum aspect ratio
is 15.609. This is still a fairly good value considering the large deformation, especially
for a total of only four mesh deformation steps, one reduced RBF step and three DGM
reconstruction steps.

Figure 7: Wing test case: Deformed wing mesh using 293 of 16285 boundary points (1 reduced RBF
step, 3 Delaunay reconstruction steps).

4 CONCLUSIONS
In this paper a new mesh deformation algorithm is developed by combining two well
proven deformation algorithms in a three-step approach. The first pre-processing step
is needed to realize a reduction of deformation source points in order to achieve a small
limited amount of source points for the following radial basis function deformation step.
In order to keep an overall point based scheme no surface connectivity is taken into
account. A surface point reduction technique based on a k-d tree is presented. In the
second step the classic radial basis function (RBF) interpolation is used to propagate the
reduced deformation field. In a last post-processing step an exact surface reconstruction
is achieved using an efficient Delaunay graph mapping (DGM) approach.

The new algorithm is analysed and compared with the classic RBF algorithm and the
DGM algorithm for a 2D viscous mesh. The new algorithm outperforms the DGM ap-
proach concerning sustainment of cell orthogonality. It is shown that the results compared
to the classic RBF algorithm are almost as good, even if a large reduction of source points
is used. Furthermore, the applicability of the new algorithm to 3D was shown with an
aeroelastic relevant test case.

Future research will focus on the analyses and comparison to other methods for more
relevant test cases and on applying this approach to a fluid-structure interaction analysis.
REFERENCES


A FINITE ELEMENT ANALYSIS OF THE PERFORMANCE OF STENTS FOR ANGIOPLASTY USING THE HYDROFORMING PROCESS

ROGÉRIO DE ARAÚJO†, SONIA A. G. OLIVEIRA†, TOBIAS A. GUIMARÃES†† AND FABIO R. F. NETO‡*

†Centro Internacional de Métodos Numéricos em Engenharia (Sala FEMEC - CIMNE) Universidade Federal de Uberlândia
Av. João Naves de Ávila, 2121, Bloco 1M, sala 219, Campus Santa Mônica, Uberlândia, Minas Gerais, Brasil
e-mail: rogerau_mat@yahoo.com.br

†† Departamento de Matemática, Universidade Federal de Goiás
Av. Dr. Lamartine Pinto de Avelar 1120, Setor Universitário
Catalão, Goiás, Brasil
email: asptobias@yahoo.com.br

Key words: biomechanics, explicit finite elements, hydroforming process, stents.

Abstract. Nowadays, stent implantation is the most used procedure to treat restenosis of coronary arteries. To prevent restenosis, a stent must absorb a large amount of plastic strain energy during its expansion to avoid the re-closure of artery wall after the implantation. According to this design criterion, it was applied the topological optimization technique to provide the best layout of a unit or cell to be used in the formation of the tubular structure of a stent. Subsequently, the topologically optimized stent cell was employed to create the three dimensional stent structure using the software Autocad/Mechanical Desktop® for the generation and treatment of this model. To test this model it was applied the software Stampack® of explicit finite elements to simulate and analyse the expansion process of optimized stent during the angioplasty by hydroforming. Although the hydroforming has been extensively employed in forming process of tubes used in the automotive industry, it was used in this work to study mechanical behavior of stents during the angioplasty. From the simulation by finite elements, it was possible to predict some design parameters of this novel optimized stent model, such as, the pressure of expansion to be applied in the balloon and the plastic strain field of the stent structure after the implant.

1 INTRODUCTION

In most cases, only the expandable balloon and catheter therapy are not able to avoid the elastic recoil or restenosis of the artery wall after angioplasty [9, 17]. In order to maintain the luminal diameter of the artery opened to receive the blood flow, cardiologists are induced to implant a metallic tubular device, known as stent, during angioplasty.
procedure. After [19], the major role of a stent is to reinforce artery wall and support arterial compression pressure after implantation.

Although restenosis in coronary arteries is immediately solved by the angioplasty, there is still a risk of its incidence in 30 to 50% of cases [9]. By considering that coronary heart disease a common reason for death, it is still necessary to search alternatives to reduce restenosis rates in coronary arteries submitted to the angioplasty process with stents. Usually, restenosis is caused either by the insufficient stiffness of a expanded stent implanted in the artery or due to lesions produced in the artery wall due to contact stress provided by the interaction between stent and artery. Nowadays, drug-eluting stents have been applied to hyperplasia and restenosis treatment. Even so, coronary revascularization problem still lies in further developments in the design of drug-eluting stents [9, 16]. Stiff stents have also been used in order to reinforce the arterial wall after balloon angioplasty [17]. Recently, it was applied the topological optimization technique as a design tool for stents geometry [8]. Since ductility of the structure was maximized, it is expected that this novel stent geometry design may be an appropriate choice for the cardiologist among the several commercial stents models [7].

Guimarães [8] developed optimal topologies designs of the stent cell plane model using boundary conditions by simulating the balloon expansion process in two dimensional space. It is interesting for the cardiologist and stent designers in some situations to study the mechanical behavior of the whole three-dimensional model of the set balloon and stent [6, 7, 16]. In this work, it is developed a methodology for the generation of a three-dimensional stent from the planar optimal topologies to analyse angioplasty procedure using the tube hydroforming simulation process. In this manufacture process, an internal pressure is applied into the die which is used to deform a metallic tube [2, 4]. Tube hydroforming has been extensively applied in the automotive components forming, but because of its similarity with the stent expansion process, it will be used in this work to simulate ballon angioplasty by finite element.

The aim of this study is to evaluate the expansion pressure of the balloon and the plastic strain field of the stent by the finite elements simulation using tube hydroforming process.

2 STENTS FOR ANGIOPLASTY

Figure 1 shows the catheter, expandable balloon and stent commonly used in the angioplasty process to unblock coronary arteries [17]. It can be seen that the catheter is the minor diameter tube and the balloon and stent are in the expanded state. When the balloon expands, the stent is subjected to the outward internal pressure. Consequently, the stent diameter is increased until its external surface to be in contact with the internal surface of the artery wall. After angioplasty, the stent should reinforce arterial wall in order to prevent restenosis or reduct artery diameter.

Figure 1: Stent, expandable balloon and catheter.
Stent structure consists of a repeating unit which is distributed along the cylindrical surface of a metallic tube with diameter between 1.5mm to 3.0mm before the angioplasty. The material used in the manufacturing of an expandable-balloon stent is usually stainless steel 316L or an alloy of cobalt, titanium, platinum or gold [17]. For these metals, when the stent expands, each repeating unit is subjected to a large displacement in the traction at vertical direction, as can be seen in the Figure 2 [1]. From this diameter expansion, each repeating unit undergoes plastic strain and the hardening of the stent material reinforces the arterial wall. Additionally, the stent geometry is also designed to be structurally stiff in order to improve the support capacity of the artery after the angioplasty.

When catheter and stent move inside the blood vessel with a small radius, they should bend before the balloon expansion. Thus, a stent should also be flexible during the angioplasty since the catheter is usually subjected to large deflections in the elastic range [15]. The stents from the first generation were exclusively designed to be stiff after the angioplasty and the flexibility was not considered [17]. However, in some situations, the angioplasty procedure could not be carried out because the set catheter, stent and balloon simply could not move inside the vessel. Since then, the stents geometries have been modified and flexible elements have been incorporated in most of commercial stents designs. For example, in the Figure 2, it is observed the presence of structures with the sinusoidal shape in the longitudinal direction placed between the structures subjected to plastic strain. These structures should improve flexibility of the stent in the longitudinal direction.

![Figure 2: Geometry of the commercial stent cell [1].](image)

3 GENERATION OF THE THREE-DIMENSIONAL STENT MODEL

3.1 Topological optimization of the stent cell

Optimization of geometry of the hardened and flexible structures from repeating unit, similar to that shown in the Figure 2, has been the starting point for the creation of the three-dimensional stent model to be studied in this work. For this purpose, it was applied the topological optimization technique to some plane structural models with the boundary conditions by simulating the expansion balloon process and the elastic deflection of stent.
cell inside the vessel. In the case of expansion balloon, it was maximized the ductility of the stent cell topology. Hence, it is expected that the stent cell topology may absorb a large amount of plastic strain energy and to reinforce the arterial wall [17]. For the stent flexible structure, its elastic strain energy was maximized improving its navigation before expansion. Figure. 3 shows optimal material distribution or topologies for both cases. A detailed discussion of the topological optimization technique formulation and its application to the stent design can be found in Guimarães [8].

3.2 Methodology for the creation of the three-dimensional stent and balloon model

The stent cell topologies shown in the Figure. 3 represent images from the relative densities matrix extracted of the optimization process. The elements of these matrix can assume only values between zero and one, where one indicates the presence of material in the topology and zero represents the absence [18]. From these images, it is possible to view only the structural aspect of the plane model of the flexible and hardened stent cell in the non-deformed state. In addition, it is also interesting to develop some procedure to study the mechanical behavior of this novel stent model and to predict some parameters in order to quantify its performance when compared with the commercial stent models.

The first step of this methodology was to extract the contours of the relative density matrix using the software Matlab®. These contours are isolines of relative density or material distribution of the optimized stent topologies shown in the Figure. 3. The function “contour”, available in Matlab®, was used to generate the position of the isolines points of each region in the stent cell plane model. The relative density used as threshold to generate isolines of relative density was also calculated automatically by the function contour.

Next, the file generated in Matlab® was imported by the software Autocad/Mechanical Desktop®. Once the contours from the topologies were irregular and non-smoothed, a procedure based on the splines had to be applied to smooth points of the stent cell isolines. This smoothing process is necessary since jagged boundaries in a structure subjected to large plastic strain cause artificial stress singularities [11]. After the definition of contour lines from the hardened and flexible topologies shown at Figure. 3, a mirror procedure was applied to both structures to provide the right hand of the stent cell. In fact, topology optimization was only applied to the half of flexible and hardened structures design space, due to longitudinal symmetry of the stent cell illustrated at Figure. 2.
The next step was the reproduction and repetition of stent cell lines in the horizontal and vertical direction in a plane. The planar stent structure, shown at Figure 4, has 3 cells or repeating units distributed in the horizontal direction and 7 cells placed on the vertical direction. Subsequently, the whole plane structure was wrapped into a cylindrical shape to generate the three-dimensional stent structure. In this way, stent geometry became a volume or a cylindrical three-dimensional solid model. A file format .igs was created from the three-dimensional stent model in the software Autocad/Mechanical Desktop® and imported in the explicit finite elements Stampack® software. Finally, the geometrical model of the balloon and their rings was created in the Stampack® software using the available drawing tools.

4  FINITE ELEMENTS MODEL OF THE STENT AND BALLOON

4.1 Simulation of the stent expansion using the tube hydroforming process

In the tube hydroforming process, a combined loading of compressive forces and internal pressure is applied by a fluid medium to obtain tubular components with different crosssections [4]. With the advancements in computer controls and high pressure hydraulic systems, this process became a viable method for mass production of automotive components, such as, crankshafts, camshafts, exhausts parts, radiator frames, front and rear axles, body parts, etc. [2]. In fact, tube hydroforming has several advantages when compared with the conventional stamping process, as for example, the consolidation of parts in only one operation and the production of components with improved structural strength and stiffness. However, this technology still needs of development and research since the most of knowledge is based on the experimentation and simulation by trial-and-error [3].

Hydroforming simulation by finite elements will be used in this work to study the mechanical behavior of the stent during the balloon expansion. This procedure was selected because the hydroforming of a tubular sheet subjected to an internal pressure with a fluid medium inside a polymer pad and the expansion process of balloon and stent are totally analogous. The unique difference is that in biomechanics, angioplasty of the balloon has dimensions of the order of millimeters and the automotive components dimensions manufactured by hydroforming are of the order of meters.

In practice, tube hydroforming is a nonlinear deformation process and the prediction of quality from analytical parameters of the manufactured component, such as, the wrinkling,
buckling and formability, are available only for some cases of the blank with simple
geometry. In this context, the finite elements method proved to be an indispensable
simulation tool to analyse these parameters [2, 3, 4].

Hydroforming is a dynamic manufacturing process dependent of the time of contact
between plastically deformed solid surfaces, therefore, most of the finite elements codes
to simulate this process use an inverse or explicit approach. The incremental or implicit
approach gives a detailed description of the whole process and the time history of stress
and strain components in any time instant during simulation may be obtained.
Nevertheless, the incremental approach requires efficient and robust contact algorithms
since the physical presence of tools and blank and all parameters of the process are
considered. Hence, the explicit approach is more computationally efficient as compared to
the incremental one. In the explicit approach, final shape and stress and strain field of the
blank and tools are computed from the initial and boundary conditions and no contact
algorithm is requested.

The total time of simulation using hydroforming process depends on the size of the
smallest finite element in the mesh. The minor is the size of the finite element, the larger
is the time increment required to compute strain and stress fields of the blank and the
tool. As the stent cell length is equal to 1.5x10^{-3} m, the total time for the finite elements
simulation can become impractical.

In this study, it will be used the similitude theory in engineering to simulate stent and
balloon expansion. The idea is to multiply dimensions of the stent and expandable
balloon by a scale factor to obtain a model for angioplasty in increased scale. In this way,
the smallest finite element dimension is increased and the critical time for simulation of
finite elements is reduced. From the results of model with the increased dimensions, it is
possible to extract the stent design parameters of the original model by means of the
theorem of pi of similitude between the models [12].

4.2 Material models of the stent and expandable balloon

Stampack® software was designed and created to simulate using finite elements the
stamping process of metallic sheets subjected to dynamic pressure loading. Some practical
examples from the stamping process have already been validated numerically by this
software, as shown in its tutorial. Material models available in the software stampack®
are exclusively dedicated to this kind of process. In stamping, the blank material models
are not usually isotropic since the rolling process provides different properties through
thickness and in the plane of the sheet. Pad used for the internal pressure application in
tube hydroforming is a rubberlike hyperelastic material subjected to large strains.
Adaptation of these material models to the stent and balloon expansion problem in
angioplasty will be discussed in the following.

Elastic-plastic material model of the stent. In this study, it will be considered only
expandable balloon stents made of stainless steel 316L. In the literature, most of papers
are related to models made of this material using a bilinear with hardening isotropic model
[6]. As the stent material is isotropic, a Von Mises model for the equivalent stress
calculation is used in the estimative of elastic-plastic stress state. In the stampack®
software, equivalent stress, \( \sigma_{eq} \), of the blank material subjected to a general stress state is
computed by considering the material anisotropy:

\[
\sigma_{eq} = \frac{\sigma_{11} (I + r_0) + \sigma_{22} (I + r_{45}) + \sigma_{12} (I + r_{90})}{r_0 (I + r_0) + r_{45} (I + r_{45}) + r_{90} (I + r_{90})} \sigma_1^{2} + \sigma_2^{2} + \sigma_{12}^{2}
\]

where \(\sigma_{11}, \sigma_{22}\) represent the normal stress components and \(\sigma_{12}\) is the shearing stress. Parameters \(r_0, r_{45}\) and \(r_{90}\) known as Lankford coefficients, determine the plastic anisotropy of the material. Subscripts 0, 45 and 90 are the material anisotropy angles measured in degrees. The larger is the magnitude of the Lankford coefficients, larger is the anisotropy between the directions. To stainless steel 316L used in the manufacturing of the stent, Lankford coefficients are equal to one since material does not have anisotropy in any direction. In this case, it can be demonstrated that the material model described in Equation (1), called Hill 48, corresponds to the Von Mises material model [10].

Evolution of yield surface defined in the Equation (1) depends on the relation between equivalent stress and plastic strain. For the bilinear with hardening isotropic material model, the equivalent stress changes linearly with strain within the plastic range. This model has been commonly used to simulate stent expansion. In the simulation of stamping process, different exponential material models have been used as the hardening law instead of the linear model. In these models, the error between experimental and theoretical strain and stress behavior is reduced. In this work, it will be used the Ludwik-Nadai model:

\[
\sigma_{eq} = K(\varepsilon_{p0} + \varepsilon_p)^n
\]

where \(K, n, \varepsilon_{p0}\) and \(\varepsilon_p\) are material parameters to be determined experimentally through uniaxial tensile test in a stainless steel 316L piece. Values of these parameters for several metallic materials used in the construction of sheets, including stainless steel 316L, are available in the software Stampack®. Stainless steel stent parameters are described on Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>1160.4 MPa</td>
</tr>
<tr>
<td>n</td>
<td>0.28</td>
</tr>
<tr>
<td>(\varepsilon_p)</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Rubberlike material model of the expandable balloon. The material of the balloon for angioplasty is a polymer with tubular shape and small thickness subjected to large elastic strains. There are several rubber type materials to be used in the manufacturing of expandable balloon. Materials for angioplasty balls of first generation were flexible and the final shape of the stent after expansion had some irregularities. Nowadays, balloon material is stiffer and stent expansion is more uniform. Polyurethane rubber is the base material used to manufacture the balloon. It can be combined with another material, such as, nylon in order to improve some mechanical properties, as for example, strength or

485
stiffness of the balloon in the expanded state.

Calculation of the stress in a rubber type material is based on the finite strain theory. Once defined the strain-energy density function, $W$, or constitutive law for the rubber, stress components are derived by differentiating $W$ with respect to the strain components [14]. The software Stampack® employs Ogden’s model to define strain-energy density:

$$W = \sum_{i=1}^{N} \frac{\mu_i}{\alpha_i} \left( \lambda_1^{\alpha_i} + \lambda_2^{\alpha_i} + \lambda_3^{\alpha_i} - 3 \right)$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are the main stretch ratios of the material, and $\mu_i$ and $\alpha_i$ are scalar parameters from the Ogden’s model. In this formulation, polyurethane rubber is considered incompressible, that is, the ratio between its deformed volume and original volume of the balloon is equal to one. The parameter $N$ defines the order of Ogden’s model. In this study, it will be used a model with three parameters ($N=3$) for the balloon polyurethane rubber. The strain energy density function parameters for the stiff rubber of balloon to be used in the stent expansion are shown on Tab. 2. A Mooney-Rivlin type function has been also applied to model hyperelastic balloon polyurethane rubber [6]. However, Ogden’s model provides a best fitting to experimental data of the rubber strain energy, particularly for large strain, when compared to the Mooney-Rivlin model [14].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>1.3</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>6.3</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>5.0</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.013</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-2.0</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

### 4.3 Design parameters of the stent and expandable balloon

After finite elements simulation, stent and balloon design parameters were extracted and analyzed in the post-processing step from Stampack® software. A complete description of the displacement, strain and stress fields of the stent and balloon in the expanded state after hydroforming simulation can be obtained from results analysis. Other parameters, such as, wrinkling, forming and stretching of the final part can also be analyzed in the post-processing stage. The analyzed parameters were the following:

**Internal pressure to be applied in the balloon.** Usually, balloon manufacturers provide a curve of internal pressure versus final diameter of the stent after expansion. Thus, cardiologists know what is the internal pressure to be applied in the expandable balloon to increase stent diameter until the diameter specified. In practice, this curve is experimentally determined by means of trial-and-error since expansion process is highly nonlinear due to large plastic strain and the geometric complexity of the stent design. In general, the curve of internal pressure versus diameter is defined analytically only for the simple blank geometry in hydroforming process [2, 3].
In our study, magnitude of internal pressure to be applied in the balloon will be estimated by running some iterations of the stent hydroforming simulation process. An initial value for internal pressure will be input in the simulation and the diameter of the expanded stent will be checked. Subsequently, internal pressure will be adjusted by trial-and-error according to final diameter observed in the balloon and stent. The maximum pressure to be applied in the expandable balloon for angioplasty is 12 atm or 1.21 Mpa [17]. An extremely high pressure can cause the rupture of balloon material or damage the arterial wall during expansion.

**Plastic strain distribution in the expanded stent.** After angioplasty, it is desired that the plastic strain fields of the stent structure are uniformly distributed. It means that the formability or ability of stent structure to absorb plastic strain energy is high. In this way, stent diameter can increase during expansion without rupture of the material. These hardened regions in the stent structure material also improve the support ability of artery wall.

5 ANALYSIS OF THE RESULTS

Figure 5 shows balloon and stent geometries before expansion process. Indeed, geometry shown at Figure. 6 represents the model in increased scale of the original balloon and stent. To obtain this amplified model, dimensions of the original geometry were multiplied by 100. Stent geometry design was created using the methodology described at section 4.1. Rings illustrated at the figure restrict expansion movements at the ends of the balloon. Stent internal diameter is equal to the balloon external diameter. Both are equal to 278mm. Lengths of the expandable balloon and stent are 4000mm and 1298mm, respectively. Constant thickness of the stent was assumed to be 10mm. The balloon has variable thickness in its ends. The ends of the balloon were modeled to be 1000mm in length and thickness linearly increases from 30 to 100mm. At the middle of the balloon thickness is constant and equal to 30mm. Figure 6 illustrates the geometry of the variable thickness from the balloon.

![Figure 5: Stent and balloon geometries before the angioplasty.](image)

In Stampack® software, the pad or expandable balloon is a volumetric solid model and the blank is considered as a shell model. As the pad is a regular geometric solid, it was meshed using structured volume hexahedral elements. On the other hand, the blank meshing was created from unstructured triangular shell elements due to the irregularity of its geometry. Another advantage of shell finite elements is the reduction of total processing time in the nonlinear finite elements analysis. Stent thickness is taken into account in the formulation of shell finite elements in its surface mesh. Figure 6 shows the stent surface mesh, the rings structured at the surface mesh and the internal surface of the
expandable balloon which will be subjected to the internal pressure. Stent meshing has 9086 nodes and balloon meshing has 48000 nodes.

Figure 6: Surface meshing from the stent and expandable stent...

Figure 7: Final shape of the stent and balloon after the expansion process.

Figure 8: Plastic strain distribution in the stent material after the angioplasty.

After some simulations, the maximum pressure to be applied in the internal surface of the balloon was selected by trial-and-error to be 0.25x10^9 Pa. Total time to simulate balloon and stent expansion was 1 hour and 28 minutes. This pressure increased the stent original diameter in about 60%. Final shape of the stent and balloon after expansion are shown at Figure. 7. Using similitude theory, it can be demonstrated that the internal pressure to expand the original model is the pressure to be applied in the increased scale model multiplied by 10^-4 [12]. Hence, the pressure to be applied in the pad of the actual model is 0.25 MPa. As this value is smaller than the maximum nominal pressure (1.21 MPa) used in the commercial balloons angioplasty, it is expected that angioplasty is a safe procedure and without damages to arterial wall.

Plastic strain distribution of the expanded stent is illustrated at Figure. 8. As can be seen, plastic strain field is concentrated in regions near to the 3 holes of the stent structure in the left edge. Hardening of these plastically deformed regions added to the stiffness of each repeating unit from the stent maintains arterial wall opened. The maximum Von
Mises plastic strain observed in the expanded stent was 58% and the strain corresponding to rupture stress for stainless steel 316L is 50%. However, due to the high ductility of the stainless steel 316L, it is expected that stent structure may expand and deform without rupture of the material.

6 CONCLUSIONS

In this study, it was developed a methodology to analyses stent and balloon three-dimensional model using hydroforming simulation by finite elements. The optimal topologies from the stent repeating unit were the starting point to create three-dimensional model. From the simulation of the model in amplified scale, finite elements processing time was reduced and the internal pressure to be applied in the balloon, as well as the plastic strain field of the stent were analyzed. This proved that the tube hydroforming simulation by explicit finite elements, extensively used in the study of automotive components hydroforming, is also a viable tool to be employed in the analysis of the mechanical behavior of stents for angioplasty.

Design parameters extracted in the post-processing stage proved that the stent and balloon geometries satisfy required specifications for the angioplasty procedure. Expansion pressure to be applied into the balloon is significantly less than the maximum nominal pressure commonly used in the practice. Consequently, the risk of rupture of the balloon and the damage of artery wall is decreased. It was also proved that plastic strain field is more distributed in the three-dimensional stent model as compared to the traditional stent cells (for example, see Figure. 2). This improves stent stiffness and capacity of support in the arterial wall reducing restenosis risk after angioplasty.

REFERENCES


FEMORAL VECTORING FOR HIP DYSPLASIA IN NEONATES

ORLANDO J. ARDILA*,†, KYLE G. SNETHEN†, EDUARDO DIVO†
FAISSAL MOSLEHY†, GEORGE T. RAB‡, ALAIN J. KASSAB†
CHARLES T. PRICE•

*,† Department of Mechanical and Aerospace Engineering
University of Central Florida
Orlando, FL 32816 USA

† Mechanical Engineering Department
Embry-Riddle Aeronautical University
Daytona Beach, FL 32114 USA

‡ Department of Orthopaedic Surgery
University of California Davis
Sacramento, CA 95817 USA

• Department of Medical Education
College of Medicine
University of Central Florida
Orlando, 32827 FL USA

Key words: Pavlik harness, hip dysplasia, dynamical analysis, passive reduction, finite element model, non-linear muscle model

Abstract. The biomechanical factors influencing the reduction of dislocated hips with the Pavlik harness in patients of developmental hip dysplasia (DDH) were studied and simulated using a three-dimensional Finite Element Method (FEM) computer model.

We identified five hip adductor muscles as key mediators in the prognosis of reduction for DDH, and determined the non-dimensional force contribution of each muscle in the direction necessary to achieve hip reduction for subluxated and fully dislocated hip joints.

Results indicate that the effects of the muscles studied are functions of the severity of DDH. For an abducted and flexed subluxated hip, the Pectineus, Adductor Brevis, proximal Adductor Magnus, and Adductor Longus muscles aid reduction, while the portions of the Adductor Magnus muscle with middle and distal femoral insertion contribute negatively. For a fully dislocated hip all muscles contribute detrimentally.

Consequently, our study points at the adductor muscles as the mediators of reduction of subluxated hips, and suggests the need for external traction to bring fully dislocated hips over the posterior acetabulum and labrum. Additionally, the reduction process of dysplastic hips was found to occur in two phases: (1) Release phase and (2) Reduction phase, and the muscles studied act distinctively in each phase. Moreover, we performed a cadaveric dissection to study the 3-dimensional orientation of the iliopsoas tendon in different hip configurations, and
found that in hip abduction and flexion this tendon is likely not an obstruction to DDH reductions.

We also report on the development of an improved three-dimensional anatomical computer model of the hip and femora of a 10-week old female infant for further study of hip dysplasia and other conditions of the hip using dynamic simulations and the Finite Element Method. This model was generated by combining CT-scans from four human subjects, as well as muscle positional data. It was segmented to encompass the distinct cartilaginous regions of infant anatomy, as well as the different regions of cortical and cancellous bone; these properties were retrieved from the literature.

This engineering computer model of an infant anatomy is being employed for (1) the development of an anatomy-based finite element and dynamics computer model for simulations of hip dysplasia reductions using novel treatment approaches, (2) the determination of a path of least resistance in reductions of hip dysplasia based on a minimum potential energy approach, (3) the study of the mechanics of hyperflexion of the hip as alternative treatment for late-presenting cases of hip dysplasia, and (4) a comprehensive investigation of the effects of femoral anteversion angle (AV) variations in reductions of hip dysplasia.

This study thus reports on an interdisciplinary effort between orthopedic surgeons and mechanical engineers to apply engineering fundamentals to solve medical problems. The results of this research are clinically relevant in pediatric orthopaedics.

1 INTRODUCTION

Hip dysplasia refers to an abnormal hip condition where misalignment, instability of the hip joint, or hip joint insufficiency without misalignment occurs, that if unsuccessfully treated, may lead to long term disability, as studies have found that as many as 76% of osteoarthritis cases are attributed to untreated hip dysplasia patients. This is in part the reason for which about 120,000 total hip replacements are performed yearly in the United States, and patients often require total hip replacement before the age of 50 [1].

This condition is typically discovered in infancy during physical examination and ultrasound [2, 3], and requires immediate treatment, for at this developmental stage ligaments are lax and the tissues surrounding the acetabulum have not yet ossified. It is thus crucial to diagnose hip dysplasia timely after birth and begin treatment immediately; The success of the treatment is inversely related to the age at which treatment is begun, with success rates remaining the highest when the treatment is begun immediately after birth or within the first month, and declining steeply after 9 months of age [2].

For infants afflicted with hip dysplasia, the Pavlik harness is the standard brace of treatment worldwide [2, 4]. This harness (Figure 1) consists of shoulder straps, a halter, anterior and posterior abduction straps, and stirrups. It is designed to maintain the hips of an infant in abduction and flexion simultaneously as this position has shown to direct the femurs to direct the femoral head toward the triradiate cartilage [4]. It is indicated for infants between one and nine months of age, with its maximum effectiveness achieved when worn shortly after birth[2], and apart from rare exceptions, it is appropriate for most children diagnosed with the condition[2, 5].

It is however a fact that in certain circumstances the Pavlik harness fails to achieve
reduction of the dislocated hips, and if unrecognized early, this failure may bring adverse consequences for children, possibly leading to permanent disability later in life as a consequence of hip osteoarthritis, delayed acetabular development, failure to stretch the hip adductors, femoral nerve palsy, or inferior (obturator) dislocation among others [5, 6].

More alarming perhaps is the fact that often physicians cannot indicate whether hip dislocations are reducible during initial examination [5], because a conclusive parameter that indicates whether reduction will or will not occur does not exist. Successful treatment currently is therefore a matter of chance, and “the construction, application, and use of the Pavlik harness is guided by a few principles and practical techniques” and “with commercial production of the Pavlik harness, some of the principles of its construction have been violated” resulting in an inadequate harness [5], with inevitable surgical interventions as the end result.

It is known that 1/20 full-term babies have some type of hip instability, and 2-3 out of 1000 babies will require treatment [7]. Hence with over 4.13 million births in the United States for the year 2009, it is safe to predict that 8-12,000 babies will require treatment for hip dysplasia, of which nearly 15% will fail treatment [2, 5].

With these pressing figures we embarked in a study that used engineering fundamentals to determine the physical laws governing the operation of the Pavlik harness, in order to bring to light conclusive determinants to its mechanism of action, and in this way allow physicians to make better decisions during treatment, thereby decreasing the incidence of disability at cause of unsuccessful treatment of hip dysplasia.

As background, Suzuki and Iwasaki found that hip reductions with the Pavlik harness are due to passive mechanical factors, with no active movement involved, and that reductions occur only during muscle relaxation in deep sleep [8, 9].

Additionally Sten-Knudsen studied muscle fibers, Hill studied toad Sartorius whole muscles, and Magid studied frog semitendinosus muscles and all three studies agree that tension in response to elongation of muscles occurs in an exponential manner [10-13]. Furthermore, these studies and our clinical observations form the basis for our study.

This study thus uses computational methods to elucidate the dynamics of hip dysplasia reductions and to characterize the obstacles to the successful treatment of hip dysplasia. Furthermore, because no method has yet been found to reduce advanced degrees of hip
dysplasia non-surgically, this study serves as a baseline to extend the frontiers of current knowledge and aim for the development of case-specific methods based on factual determinants obtained in Computed Tomography (CT) scan exam results, to actively vector the femoral head to its proper concentric position in the acetabulum using both active mechanical means, and/or passive methods that employ possible paths of minimum potential energy that can be used to relocate the femur to its physiological position.

Expanding on these possibilities, we present the results of a study of the biomechanics of different severities of hip dysplasia, conducted using a simplified dynamics model. Additionally, we expand with a description of the four additional, simultaneous studies that are being carried out in our laboratory; namely (1) the development of a complete finite element and dynamics computer model for simulations of hip dysplasia reductions using novel treatment approaches, (2) the determination of a path of least resistance in reductions of hip dysplasia based on minimum potential energy, (3) the study of the mechanics of hyperflexion of the hip as alternative treatment for late-presenting cases of hip dysplasia, and (4) a comprehensive investigation of the effects of femoral anteversion angle (AV) variations in reductions of hip dysplasia.

This study was based on a thorough literary review in both the technical and clinical areas relevant to hip dysplasia, as well as the professional expertise and active participation of a group of orthopaedic surgeons.

2 METHODS

2.1 Development of a Simplified Dynamics Computer Model

We constructed a simplified three-dimensional (3D) dynamical computer model of an infant hip utilizing SolidWorks (Dassault Systèmes Simulia Corp., Providence, RI, USA). This model consists of the anatomical features that correspond to the hip bone, right femur, and five muscles of relevance to the biomechanics of DDH (Pectineus, Adductor Longus, Adductor Brevis, Adductor Magnus, Gracilis). Because the geometry of the Adductor Magnus can be regarded as a triangular sheet, we divided it into three components (Adductor Minimus, Middle, and Posterior) [14]. Our computational model thus results in seven distinct muscle entities. This model is capable of simulating spontaneous reductions by the Pavlik harness, as it was observed that these reductions occur passively during muscle relaxation in deep sleep [8, 9]. To create this model we employed the medical segmentation software Mimics (Materialise Inc, Plymouth, MI) to measure the anatomical features of the hip of a 6-month old female infant, and used these measurements to generate the 3D model of a simplified hip and leg (Figure 2). We constrained the leg to the degrees of freedom required to simulate a dislocated, and a reduced hip, in abduction and flexion. Furthermore, we restricted translation of the leg in the Y-direction, and restricted rotation about the Z-axis to account for the motion restrictions imposed by the Pavlik harness.

The simulated hip-leg assembly model is driven by gravitational loads and constrained to respond in the spatial envelope determined by the Pavlik Harness. It is supported by adductor muscles which were modeled following the constitutive model for passive muscle tension-elongation response introduced by Magid and Hill [10-12]. Their model was calibrated to fit the requirements of our model. The final model is presented in Equation 1.
In Equation 1, \( T_{\text{calib}} \) is the tension in the muscles, \( C \) is a calibration constant obtained, \( A \) is the cross-sectional area of the muscles, \( E_0 \) is an initial elastic modulus, \( \alpha \) is an empirical constant, and \( \lambda \) is muscle stretch.

\[
T_{\text{calib}} = CA \frac{E_0}{\alpha} \left( e^{\alpha(\lambda-1)} - 1 \right)
\]  

Figure 2: Three-dimensional dynamic computer model for simulations of hip dysplasia reductions. a) Hip and right leg assembly viewed laterally (topmost) and axially (middle). b) Hip and right leg assembly viewed axially, displaying modeled musculature.

This model was then used to study two severities of DDH:

(1) Graf III: subluxated hip in which the center of the femoral head lies on the posterior rim of the acetabulum.

(2) Graf IV: fully dislocated hip in which the femoral head is located posterior to the acetabulum.

Initial conditions for (1) and (2) were imposed by dislocating the femoral head and placing it in the locations corresponding to each severity. The differential equations of motion defined by Newton’s second law were then solved using the GSTIFF integrator embedded in the SolidWorks software package until equilibrium was reached. We defined successful reductions as conditions in which the femoral head slid into the acetabulum, and concentric reduction was maintained throughout the solution until reaching equilibrium. Furthermore, we
calculated the directional cosines of muscle tension necessary to affect reduction for each DDH severity studied.

2.2 Development of an anatomy-based FEM and dynamics computer model

For additional studies of conditions of the hip, we developed an anatomy-based FEM and dynamics computer model consisting of the hip and femora of a 10-week old female infant. For its development we combined CT-scan data and muscle positional data from four different human subjects:

1. 10-week old female infant
2. 14-year old female
3. 38-year old male (Visible Human Project, The National Library of Medicine)
4. Adult Male of unknown age [14].

Various important regions of the infant hip are composed of cartilage, and thus are difficult to visualize in CT-scans. For this reason we performed a 3-dimensional reconstruction of the CT-scans of a 14-year old female (2), and scaled it anisotropically to match the anatomical proportions of the 10-week old infant (1). Our scaling was performed by superposition of anatomical landmarks, namely anterior superior and posterior superior iliac spines, and acetabuli. Upon scaling the pubic and ischial rami of the scaled 14 year old female hip was found to trace a wider arc while all other landmarks matched well. This arc was manually modified to closely match the trace of the infant hip arc; we attribute the observed variation to the possible widening of female hips at puberty. Our resulting scaling factors were 0.35, 0.32, and 0.32 in the X, Y, and Z directions respectively.

To generate the right femur, CT-data from the Visible Human Project (3) was scaled anisotropically to match the femur size of the 10-week old female (1), according Standards in Pediatric Orthopedics [15]. We used femoral head diameter and length between epiphyseal plates as scaling criteria. This procedure yielded scaling factors of and 0.22, 0.25 and 0.23 in the X, Y and Z directions. Making use of symmetry, the right femur was mirrored to create a model for the left femur.

Locations of origins and insertions of muscles were assigned by scaling adult male data (4) [14] isotropically to match the proportions of the 6-month old infant, using the distances between acetabular centers as scaling parameters. The resulting scaling factor was 0.39. The scaled muscle origin and insertion points matched the expected anatomical landmarks accurately, except in the X-direction at the insertions. In this location some muscles were found to lie slightly off the linea aspera. These muscles were manually adjusted to match the linea aspera, and this procedure is not expected to cause significant error in the results as the moment variation due to the modified moment arms is not believed to be significant.

3 RESULTS

3.1 Dynamic Simulation of reductions of DDH, grades Graf III and Graf IV.

A successful dynamic reduction simulation was carried out indicating that Graf III subluxations can be reduced by the Pavlik harness. For further insight we considered (a) the magnitudes and (b) directions of muscle tensions. We found that the Gracilis, Adductor
Middle, and Adductor Posterior contribute negatively to the reduction with use of the Pavlik harness (Figure 3, Table 1), and that the percent constructive contribution of the muscles studied increases in direct proportion with abduction angle.

![Figure 3: Percent contribution of muscle tension towards reduction vs. abduction angle - Graf III.](image)

Table 1 - Percent directional contributions of muscle tensions in the direction of reduction - Graf III.

<table>
<thead>
<tr>
<th>Muscle</th>
<th>Abduction Range (degrees, measured from sagittal)</th>
<th>Contribution order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pectineus</td>
<td>42.7 52.2 60 70 80</td>
<td>1</td>
</tr>
<tr>
<td>Add Longus</td>
<td>-3.0% 10.1% 19.7% 31.6% 42.8%</td>
<td>4</td>
</tr>
<tr>
<td>Add Brevis</td>
<td>10.6% 20.4% 29.1% 38.9% 48.2%</td>
<td>2</td>
</tr>
<tr>
<td>Add Minimus</td>
<td>4.7% 15.7% 23.9% 34.2% 44.0%</td>
<td>3</td>
</tr>
<tr>
<td>Add Middle</td>
<td>-32.6% -19.3% -9.1% 4.4% 17.8%</td>
<td>5</td>
</tr>
<tr>
<td>Add Posterior</td>
<td>-44.6% -30.2% -19.0% -3.9% 11.3%</td>
<td>6</td>
</tr>
<tr>
<td>Gracilis</td>
<td>-50.32% -36.60% -25.18% -10.60% 4.00%</td>
<td>7</td>
</tr>
</tbody>
</table>

Simulation results of hip dysplasia of grade Graf IV indicate that reductions of hip dysplasia of this severity are unlikely to occur by Pavlik harness treatment. In this severity, and while the infant wears the Pavlik Harness, the tensions of all muscles contribute negatively to the direction of the motion necessary for reduction.
Figure 4: Percent contribution of muscle tension towards reduction vs. Abduction angle - Graf IV.

Table 2 - Percent contribution of muscle tensions in the direction of reduction - Graf IV.

<table>
<thead>
<tr>
<th>Percent directional contribution of tensions - Graf IV</th>
<th>Abduction angle</th>
<th>Contribution order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>52.3°</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70°</td>
<td></td>
</tr>
<tr>
<td>Pectineus</td>
<td>-42.9%</td>
<td>-29.7%</td>
</tr>
<tr>
<td>Add Longus</td>
<td>-64.8%</td>
<td>-52.2%</td>
</tr>
<tr>
<td>Addr Brevis</td>
<td>-53.5%</td>
<td>-42.1%</td>
</tr>
<tr>
<td>Add Minimus</td>
<td>-57.3%</td>
<td>-45.7%</td>
</tr>
<tr>
<td>Add Minimus</td>
<td>-77.7%</td>
<td>-67.9%</td>
</tr>
<tr>
<td>Add Posterior</td>
<td>-86.5%</td>
<td>-77.3%</td>
</tr>
<tr>
<td>Gracilis</td>
<td>-91.3%</td>
<td>-84.0%</td>
</tr>
</tbody>
</table>

3.2 Resultant anatomy-based computer model

Converging and scaling CT-scan and muscle positional data from four human subjects resulted in a 3-dimensional anatomical computer model of an infant hip and femora corresponding to the size of a 10-week old female infant. This model was segmented to account for the different regions of cortical and cancellous bone, as well as for the distinct regions of cartilage observed in infant anatomy. Properties of the different bone tissues and cartilage were obtained from the literature and assigned to the model.
Figure 5: Three dimensional anatomy-based model of human hip and femor

The method developed to create this infant anatomical computer model may be used to generate Computer Aided Design models of various different components of the human anatomy with different dimensional requirements, by combining various sets of data into a single model. The resulting models may then be used to study various conditions in orthopaedics.

The 10-week old female anatomical model that we generated is being used in our laboratory to carry out four parallel investigations:

1. Development of a finite element and dynamics computer model for simulations of hip dysplasia reductions: This model includes adductor muscles that follow physiological geometries to account for the moment arms that simplified and straight-line muscle models cannot model properly.

2. Determination of a path of least resistance for reductions of hip dysplasia: With this model (Figure 6) we make use of the anatomical geometry around the acetabulum, gravitational potential energy, and strain energy in the muscles, to find a path of minimum potential energy which can be used to vector the femoral head to the acetabulum.

3. Study the mechanics involved in hyperflexion of the hip: This study (Figure 7) aims to determine the mechanics of hyperflexion as alternative treatment for late-presenting cases of hip dysplasia.

4. Effects of femoral anteversion angle (AV) variations in reductions of hip dysplasia: We performed artificial derotation osteotomies to a femur (Figure 8) to evaluate the mechanical implications of AV variations in the treatment of hip dysplasia. We aim to determine if femoral anteversion may be used as determining factor for reducible or irreducible hips.
Figure 6: MATLAB rendering of femur and acetabulum for minimum potential energy analysis. Color gradient represents depth in Z direction.

Figure 7: Development of model to evaluate hyperflexion of the hip as alternative treatment approach.


4 DISCUSSION & CONCLUSIONS

We developed a simplified model of the anatomy of the human hip, and carried out simulations of reductions of hip dysplasia for the Graf III and Graf IV severities.

Our study aims to prevent treatment failures with the Pavlik harness, and we found that in hip subluxations (Graf III) the tension that develops in most muscles contributes to successful treatment; while in complete hip dislocations (Graf IV) the tension in most muscles contributes detrimentally to treatment. This implies that a treatment different from the Pavlik harness must be employed in greater severities of DDH.

Additionally we developed a method by which CT-scan data belonging to different human subjects may be combined into a single model, and scaled to meet different dimensional requirements. We employed this method to combine the CT-data and muscle positional data from four different human subjects, including adults, to engineer a comprehensive anatomy-based FEM and dynamics model of the lower anatomy of a 10-week old female infant. We are using this model for the study of the treatment for hip dysplasia.

Our findings are clinically relevant and directly applicable in the field of orthopaedics.

5 ACKNOWLEDGMENTS

This study was supported in part by the US National Science Foundation under grant number CBET-1160179, Orlando Health, and the International Hip Dysplasia Institute. The authors gratefully acknowledge the contribution of Mechanical Engineering students Sergio Gomez, and Justin Kingsley.
6 REFERENCES


A NEW VARIATIONAL FRAMEWORK FOR LARGE STRAIN PIEZOELECTRIC HYPERELASTIC MATERIALS

J. BONET*, R. ORTIGOSA* AND A. J. GIL*

*Civil & Computational Engineering Centre, College of Engineering, Swansea University, Swansea SA2 8PP, UK

Key words: Energy harvesting; Piezoelectricity; Polyconvexity; Large deformations; Finite Element Method

Abstract. In this paper, a novel nonlinear variational formulation is presented for the numerical modelling of piezo-hyperelastic materials. Following energy principles, a new family of anisotropic extended internal energy density functionals is introduced, dependent upon the deformation gradient tensor and the Lagrangian electric displacement field vector. The requirement to obtain solutions to well defined boundary value problems leads to the definition of energy density functionals borrowing concepts from polyconvex elasticity. Material characterisation of the constitutive models is then carried out by means of experimental matching in the linearised regime (i.e. small strains and small electric field). The resulting variational formulation is discretised in space with the help of the Finite Element Method, where the resulting system of nonlinear algebraic equations is solved via the Newton-Raphson method after consistent linearisation. Finally, a series of numerical examples are presented in order to assess the capabilities of the new formulation.

1 INTRODUCTION

The earliest piezoelectric materials to be discovered, i.e. crystals, have shown limited applicability due to their high stiffness and brittleness. The recent advent of piezoelectric polymers has meant a turning point in the development of piezoelectricity. The circumvention of the drawbacks associated with their crystal predecessors has broadened considerably their applications as actuators, power generators and energy harvesters.

Piezoelectric polymers have traditionally been used as smart actuators in microelectromechanical systems. However, their ability to emulate the functioning of biological muscles as well as their large strain capabilities, have recently triggered the emergence of new exciting applications, such as artificial muscles. A more recent application within the field of smart actuators can be found in space microwave antennas. These devices experience shape deviations due to pre-stress and thermal expansion. These deviations are
corrected by controlling an applied electric potential on a piezoelectric patch to maintain the desired shape of the antenna reflector. Figure 1 illustrates two application examples.

![Image 1](image1.png)

**Figure 1:** (a) Electroactive actuated robotic arm. (b) Shape control of space antenna.

The large strain and piezoelectric capabilities associated to piezoelectric polymers confer them with attractive properties in the field of power generation and energy harvesting. There is currently a growing need for this kind of applications. For instance, piezoelectric eels (see Figure 2(a)), which are used as part of submarine devices in long endurance military missions. The energy-harvesting eel is designed to extract energy from the wake of a bluff body in an ocean current. The basic configuration is a leading bluff body trailed by a thin flexible piezoelectric eel. The bluff body generates vortices which excite a flapping motion of the eel. The eel deformation results in strain of the piezoelectric membrane, which in turn generates a voltage across the material. Other current applications include small piezoelectric devices that when attached to the insole of a shoe are able to transform mechanical motion into electrical power (see Figure 2(b)).

![Image 2](image2.png)

**Figure 2:** (a) Piezoelectric eel. (b) Walking energy harvesting device.
Electromechanical interactions in materials are not only due to piezoelectricity. Another important phenomenon which has found a vast number of applications within the field of electromechanical actuators is electrostriction. This phenomenon results from a quadratic relation between stresses and electric field. The application of electric fields on purely electrostrictive materials would lead to a deformation on the material. However, unlike piezoelectricity, this effect cannot be reversed. It is worth emphasising that piezoelectricity and electrostriction are not mutually exclusive.

The existing framework for the numerical simulation of piezoelectric materials requires an enhancement as a result of the development of these new polymers, capable of undergoing large deformations. Unlike crystals, the classical linearised theory can no longer be applied for a reliable computer simulation. In this paper, a nonlinear variational formulation for piezo-hyperelastic materials is introduced with the help of the internal energy density $U$ constructed on the basis of the right Cauchy-Green deformation tensor $C$ and the Lagrangian electric displacement field vector $D_0$.

2 ANISOTROPIC STRUCTURE OF PIEZOELECTRIC MATERIALS.

The proposed internal energy density $U$ must lead to the definition of a constitutive model which needs to capture the material response in an accurate manner. Accordingly, $U$ must be compliant with the physical and mathematical constraints inherent to these materials (e.g. anisotropy). This will be taken into consideration by means of a hybrid approach which combines the isotropic extension surface concept introduced in reference [2] and a more recent methodology proposed in reference [3].

The first approach, i.e. the isotropic extension concept, introduces the so-called isotropic extension surface $\phi$, which includes the set of vectors and tensor valued functions which are preserved under the action of the material symmetry group which characterises the corresponding anisotropy of the material, that is,

$$\phi = \phi (C, D_0).$$  

As a result, the creation of anisotropic invariants can alternatively be achieved by formulating isotropic invariants which take into consideration the isotropic extension surface $\phi$. Thus, the extended internal energy density functional $U$ is created in terms of isotropic invariants of the following arguments,

$$U = \hat{U} (C, D_0, \phi (C, D_0)).$$  

An alternative approach for the creation of anisotropic invariants was recently proposed in [3]. In this reference, a new metric tensor $G = HH^T$ is introduced, where $H$ represents a linear tangent map established between the Cartesian base triad \(\{e_1, e_2, e_3\}\) and the crystallographic base system \(\{a_1, a_2, a_3\}\), the latter related to the associated Bravais lattice (see Figure 3).
3 CONSTITUTIVE EQUATIONS IN ELECTRO-MECHANICS.

The derivation of the constitutive equations for Electro-Mechanics requires a revision of the First and Second laws of Thermodynamics. Following an integral material description, the First law of Thermodynamics gives the variation of internal energy as [1],

\[
\frac{d}{dt} \int_{\Omega_0} U (F, D_0, \eta_0) \, d\Omega_0 = \int_{\Omega_0} P : \dot{F} \, d\Omega_0 + \int_{\Omega_0} E_0 \cdot \dot{D}_0 \, d\Omega_0 - \int_{\partial\Omega_0} \nabla_0 \cdot q_0 \, d\Omega_0 + \int_{\Omega_0} R_0 \, d\Omega_0,
\]

where \( F \) is the deformation gradient tensor, \( D_0 \) is the Lagrangian electric displacement field, \( \eta_0 \) is the Lagrangian entropy, \( P \) is the first Piola-Kirchhoff stress tensor, \( E_0 \) is the Lagrangian electric field, \( \theta \) is the temperature, \( q_0 \) is the Lagrangian Fourier-Stokes heat flux vector field and \( R_0 \) is the Lagrangian heat supply field per unit volume.

The Clausius-Duhem form of the Second law of Thermo-Electro-Mechanics can be written as [1],

\[
\frac{d}{dt} \int_{\Omega_0} \eta_0 \, d\Omega_0 \geq \int_{\Omega_0} \frac{R_0}{\theta} \, d\Omega_0 - \int_{\Omega_0} \nabla_0 \cdot \frac{q_0}{\theta} \, d\Omega_0.
\]

Combination of both First and Second laws leads to,

\[
\int_{\Omega_0} \left( P - \frac{\partial U}{\partial F} \right) : \dot{F} \, d\Omega_0 + \int_{\Omega_0} \left( E_0 - \frac{\partial U}{\partial D_0} \right) \cdot \dot{D}_0 \, d\Omega_0 + \int_{\Omega_0} \left( \theta - \frac{\partial U}{\partial \eta_0} \right) \dot{\eta}_0 \, d\Omega_0 - \int_{\Omega_0} \frac{1}{\theta} \nabla_0 \theta \cdot q_0 \, d\Omega_0 \geq 0.
\]
Constitutive equations based on the internal energy $U$ can be derived from the combination of the First and Second law of Thermo-Electro-Dynamics [1],

$$
P = \frac{\partial U(F, D_0, \eta_0)}{\partial F}, \quad E_0 = \frac{\partial U(F, D_0, \eta_0)}{\partial D_0}, \quad \theta = \frac{\partial U(F, D_0, \eta_0)}{\partial \eta_0} - \frac{1}{\theta} \nabla_0 \theta \cdot q_0 \geq 0, \quad (6)$$

where $q_0$ can be related to the material gradient of the temperature $\theta$ as,

$$
q_0 = -k \nabla_0 \theta, \quad (7)
$$

where $k$ is the thermal conductivity tensor. Positive definiteness of $k$ ensures that $-\frac{1}{\theta} \nabla_0 \theta \cdot q_0 \geq 0$. For reversible Electro-Mechanics, the constitutive equations can be simplified as,

$$
P = \frac{\partial U(F, D_0)}{\partial F}, \quad E_0 = \frac{\partial U(F, D_0)}{\partial D_0}. \quad (8)$$

Different constitutive laws could be proposed by introducing alternative internal energy densities $U$. In addition, the electric enthalpy $H(F, E_0)$ can be introduced by means of the Legendre transform,

$$
H(F, E_0) = U(F, D_0) - E_0 \cdot D_0. \quad (9)
$$

For the electric enthalpy, analogous constitutive laws to equations (8) can be written as,

$$
P = \frac{\partial H(F, E_0)}{\partial F}, \quad D_0 = -\frac{\partial H(F, E_0)}{\partial E_0}. \quad (10)
$$

In the absence of external electric or mechanical loads, equilibrium configurations are reached when the internal energy $U(F, D_0)$ attains a minimum. It is possible to show that for the same equilibrium configuration, the complementary electric enthalpy $H(F, E_0)$ attains a saddle point. In other words, in a neighbourhood of the equilibrium configuration, the internal energy $U(F, D_0)$ would be a convex function of $F$ and $D_0$ whereas the electric enthalpy $H(F, E_0)$ would be a saddle function, i.e convex in $F$ and concave in $D_0$. Figure 3 shows the nature of both $U(F, D_0)$ and $H(F, E_0)$ in a neighbourhood of the equilibrium configuration for a one-dimensional case.
4 Mathematical requirements: Polyconvexity.

The properties of the constitutive models used for the numerical simulation of piezoelectric polymers must also guarantee existence of solutions to boundary value problems in large deformation regimes. Convexity of the internal energy functional ensures existence. However, this condition is too stringent as it precludes the appearance of different equilibrium configurations with equal potential energy, i.e. buckling. Moreover, it can also violate material frame indifference. A necessary and almost sufficient condition for the existence of minimisers to a variational formulation is quasiconvexity. However, this is an integral condition whose verification is a cumbersome task. As a result, the conditions preferred in this work which guarantee the existence of solutions to well defined boundary value problems, are polyconvexity and coercivity [4].

Borrowing concepts from nonlinear elasticity, an energy density functional is said to be polyconvex in the deformation gradient tensor $F$ if it can be written as a convex function of the following arguments,

$$U = \hat{U} (F, \text{Cof} F, \det F).$$

An intuitive and physical understanding of polyconvexity can be gained from its geometric interpretation. The arguments appearing in equation (11) govern the transformations by means of which line, area and volume material elements are mapped to their spatial counterparts (see Figure 5).

There is an extensive work done in the creation of polyconvex energy functionals in the field of nonlinear elasticity. However, the coupled nature of the electromechanical problem requires a generalisation of this concept. As a result, this paper presents a methodology for the development/implementation of polyconvex energy functionals on the basis of the deformation gradient tensor $F$ and the material electric displacement field $D_0$ [5]. According to this methodology, polyconvexity is enforced by means of a convex function.
of the following arguments,

\[ U = \hat{U}(F, \text{Cof} F, \det F, D_0). \]  

(12)

5 Constitutive models for piezo-hyperelastic materials.

From the physical point of view, a possible decomposition of the internal energy density is carried out in terms of the three constitutive parts involved in the formulation, i.e. mechanical, piezoelectric and electrostrictive,

\[ U(D_0, C) = U_m(C) + U_e(D_0, C) + U_p(D_0, C), \]  

(13)

where \( U_m, U_p \) and \( U_e \) account for the mechanical, piezoelectric and electrostrictive components of the stored internal energy functional \( U \), respectively. Naturally, the mechanical component of the energy functional can also be split into its isotropic and anisotropic contributions,

\[ U_m = U_{iso,m}(C) + U_{aniso,m}(C). \]  

(14)

The isotropic contribution of the mechanical component \( U_{iso,m}(C) \) can be chosen from any of the existing polyconvex strain energy functionals, i.e. Mooney-Rivlin. The definition of the three remaining components (see equations (13)-(14)) completes the characterisation of the material model. Anisotropic mechanical energy functionals built on the basis of polyconvex invariants can be found in [3]. This author proposes the follow-
ing expression for $U_{\text{aniso,m}}$ in terms of an additive decomposition of a series of families $j$ characterising the anisotropy of the material,

$$
U_{\text{aniso,m}} = \sum_j \mu_j \left( J_{4j}^{\alpha_j+1} + J_{5j}^{\beta_j+1} + \frac{\text{tr} \left( \mathbf{G}_m^j \right) III - \gamma_j}{\gamma_j} \right),
$$

(15)

with,

$$
J_{4j} = \text{tr} \left( \mathbf{CG}_j^m \right); \quad J_{5j} = \text{tr} \left( \mathbf{C} \text{Cof} \mathbf{G}_j^m \right); \quad III = \det \mathbf{C},
$$

(16)

where $\mathbf{G}_j^m$ is a possible metric tensor characterising the anisotropy of the material, $\alpha_j, \beta_j$ and $\gamma_j$ are dimensionless parameters and $\mu_j$ is a mechanical material parameter.

For the material model to be complete, it is necessary to account for the remaining constitutive counterparts, i.e. piezoelectric and electrostrictive.

In this paper, a novel definition for the respective components of the energy functional on the basis of polyconvex invariants is introduced. A possible electrostrictive component $U_e$ would be,

$$
U_e = \sum_j \omega_j \left[ p_{e,j} \cdot \mathbf{C} p_{e,j} + I_2^e \left( p_{e,j} \cdot p_{e,j} \right)^2 \right] + f_e (\mathbf{C}) + g_e (\mathbf{D}_0),
$$

(17)

with,

$$
p_{e,j} = m_j \mathbf{G}_j^e \mathbf{D}_0 \quad \mathbf{D}_0 = \frac{D_0}{\sqrt{\mu_{\text{ref}}} \varepsilon_0},
$$

(18)

where $\mathbf{G}_j^e$ is a possible metric tensor characterising the anisotropy of the material, $m_j$ is a dimensionless parameter, $\omega_j$ is an electrostrictive material parameter, $\varepsilon_0$ is the electric permittivity in vacuum and $\mu_{\text{ref}}$ is a reference mechanical material parameter. A possible contribution for the piezoelectric component $U_p$ would be,

$$
U_p = \sum_{j,k} \lambda_{jk} \left[ p_{p,jk} \cdot \mathbf{C} p_{p,jk} + I_2^e \left( p_{p,jk} \cdot p_{p,jk} \right)^2 \right] + f_p (\mathbf{C}) + g_p (\mathbf{D}_0),
$$

(19)

with,

$$
p_{p,jk} = s_j \mathbf{G}_j^p \mathbf{D}_0 + \mathbf{N}_k,
$$

(20)
where $G^p_j$ is a possible metric tensor characterising the anisotropy of the material, $s_j$ is a dimensionless parameter, $N_k$ accounts for the structural vectors included in the isotropic extension surface of the material considered and $\lambda_{jk}$ is a piezoelectric material parameter. It is important to emphasise that the definitions of the mechanical $U_m$ and electrostrictive $U_e$ components are based upon the methodology proposed in [3]. However, the piezoelectric component has been formulated by means of a combination of this method and the isotropic extension concept in [2].

The functions $f_e$, $f_p$, $g_e$ and $g_p$ are chosen so that no stress or electric field is obtained in the origin,

$$S|_{D_0=0,C=I} = 2 \frac{\partial U_p}{\partial C} \bigg|_{D_0=0,C=I} + 2 \frac{\partial U_e}{\partial C} \bigg|_{D_0=0,C=I} = 0$$

$$E|_{D_0=0,C=I} = \frac{\partial U_p}{\partial D_0} \bigg|_{D_0=0,C=I} + \frac{\partial U_e}{\partial D_0} \bigg|_{D_0=0,C=I} = 0. \quad (21)$$

The complete characterisation of the material model defined through the internal energy density requires the definition of the material properties $\alpha_j$, $\beta_j$, $\gamma_j$, $\mu_j$, $\omega_j$, $m_j$, $\lambda_{jk}$, $s_j$ and the metric tensors $G^m_j$, $G^e_j$ and $G^p_j$. This is achieved by performing a match in the origin, i.e. $C = I$, $E_0 = 0$, between the constitutive tensors derived from the proposed energy functionals and the experimental tensors available in the linearised regime (i.e. small strains and small electric field),

$$C|_{C=I,E_0=0} = c; \quad P|_{C=I,E_0=0} = p; \quad A|_{C=I,E_0=0} = \epsilon; \quad (22)$$

where $C$, $P$ and $A$ are the elastic, piezoelectric, and dielectric material tensors, respectively, derived from the energy functional as follows,

$$A_{ij} (C, E_0) = \left[ \frac{\partial^2 U (C, D_0)}{\partial D_{0j} \partial D_{00}} \right]_{ij}^{-1}$$

$$P_{ijk} (C, E_0) = -2A_{mj} \frac{\partial^2 U (C, D_0)}{\partial C_{jk} \partial D_{0m}}$$

$$C_{ijkl} (C, E_0) = 4 \frac{\partial^2 U (C, D_0)}{\partial C_{ij} \partial C_{kl}} + 2 \frac{\partial^2 U (C, D_0)}{\partial C_{ij} \partial D_{0m}} P_{mkl}. \quad (23)$$

From the previous match, the necessary material parameters can be obtained either by identification or minimisation.

6 Numerical results.

The resulting variational formulation is discretised in space with the help of the Finite Element Method, where the resulting system of nonlinear algebraic equations is solved
via the Newton-Raphson method after consistent linearisation. In this section, a series of numerical examples are presented in order to demonstrate the robustness and applicability of the formulation.

Figure 6 shows an example in which the application of an external electric field produces a deformation in a composite shell. Figure 7 depicts another example of actuator application. The application of an external electric field on a composite material produces deformations and changes in shape. In this case, the different anisotropic orientation of both layers of materials leads to an out of plane deformation. This effect enables these materials to be designed to carry out a specific mechanically demanded task.

REFERENCES


Figure 6: Composite circular shell subject to electric potential gradient across the thickness.
Figure 7: Composite rectangular shell subject to electric potential gradient across the thickness.
MATHEMATICAL STRUCTURE OF THE TRANSPORT EQUATIONS FOR COUPLED 2D-3D ELECTRON GASSES IN A MOSFET.

V. D. Camiola*, V. Romano*

*Dipartimento di Matematica e Informatica
Università degli Studi di Catania
Viale A. Doria 6, 9505 Catania, Italy
e-mail: camiola@dmi.unict.it
e-mail: romano@dmi.unict.it, web site: http://www.dmi.unict.it/~romano

Key words: quantum transport; semiconductor; hydrodynamical models; 2DEG-3DEG coupling

Abstract. In a previous paper [1] we have studied the coexistence of coupled 2DEG and 3DEG in the proximity of a silicon-oxide interface in a MOSFET devising a hydro-dynamical model obtained by taking the moment of the kinetic transport equation and by resorting to the maximum entropy principle for the closure relations. Here we classify the model from the point of view of PDEs by showing that it is hyperbolic in the relevant physical region of density, energy, velocity and energy fluxes in each subband and bulk electrons.

1 Introduction

In [1] we have presented a subband transport model for the description of charge transport in a MOSFET. Under the gate oxide, in the channel of the device, there is a quantization in the transversal direction forming a 2D electron gas but far from such a region electrons are 3D. Therefore one has to include the coexistence of both 2D and 3D electron gas inside the channel and only 3D electrons in the remaining part.

Starting from the Boltzmann equations, corresponding moment equations have been written and the closure problem, typical of such a kind of balance equations, has been solved by resorting to the maximum entropy principle. The obtained complete model is constituted by a set of balance equations for average density, velocity, energy and energy flux in each subband and for bulk electrons, coupled with the Schrödinger-Poisson system.

A crucial point has been how to take into account the transition of electrons from the 3DEG to the 2DEG. We have solved the problem inspired by the procedure used in [2]. If an electron belonging to the 2DEG gains an energy above a threshold value after a
scattering, it is considered as 3D and vice versa if a 3D electron gets an energy below a threshold value after an emission process it is considered into the 2DEG.

Here we show how the range of variation of the threshold energy influences the mathematical structure of the balance equations.

2 Confinement effects in nanoscale MOSFET

In a MOSFET (fig.1-left), in the proximity of the silicon-oxide (Si/SiO$_2$) interface a two dimensional electron gas (2DEG) is created with a discrete energy spectrum along the z-direction (fig.1-right). Above a fixed energy level, the energy spectrum is continuous and a three dimensional electron gas (3DEG) coexists with the 2DEG. To describe the whole system, we define a spatial quantum region $R_Q$ where the 2DEG is confined, associated with an energy quantum region $R_E$ in the wave-vector space. Outside $R_Q$ electrons are only belonging to the 3DEG.

In the quasi-static approximation, the 2D-charges in $R_Q$ are described by the steady wave function
\[ \psi_{\nu}(k, r) = \phi_{\nu}(\mathbf{r}_\parallel, z) e^{i k_\parallel \cdot \mathbf{r}_\parallel} \]
with $k_\parallel = (k_x, k_y)$ and \( \mathbf{r}_\parallel = (x, y) \) denoting the longitudinal components of the wave-vector $k$ and the position vector $r$, respectively, and $A$ symbolizing the area of the $xy$ cross-section of $R_Q$.

\( \phi_{\nu}(\mathbf{r}_\parallel, z) \) is called envelope function and, under the scaling where the ratio between transversal and longitudinal characteristic lengths is small, it is solution of the Schrödinger equation
\[
-\frac{\hbar^2}{2m^*} \frac{d^2}{dz^2} \phi_{\nu}(\mathbf{r}_\parallel, z) + q(V_C + V) \phi_{\nu}(\mathbf{r}_\parallel, z) = \varepsilon_{\nu} \phi_{\nu}(\mathbf{r}_\parallel, z) \]
in the effective mass approximation. \( \hbar \) is the reduced Planck constant, \( m^* \) is the effective electron mass, \( V_C \) is the confining potential and \( V \) is the self-consistent electrostatic potential which solves the Poisson equation. Note that $\mathbf{r}_\parallel$ enters as a parameters.

Under the assumption that the confining potential gives rise to an infinite barrier at the oxide-silicon interface (\( z=0 \)) and that a fictitious boundary is posed at \( z = t_Q \), we solve the Schrödinger equation only inside the \( R_Q \) region by setting $\phi = 0$ at $z = 0$ and $z = t_Q$, the boundary of $R_Q$. The problem is a self-adjoint problem posed on a limited domain. So one finds a countable set of normalized eigen-pairs (subbands) \( (\phi_{\nu}, \varepsilon_{\nu}) \).

In each subband the energy \( E_{\nu} \) is the sum of a transversal contribution $\varepsilon_{\nu}(\mathbf{r}_\parallel)$ and a

Figure 1: Simulated MOSFET (left). Energy spectrum along the transversal direction (right).
longitudinal (kinetic) contribution \( \varepsilon_\parallel = \frac{\hbar^2}{2m^*} (k_x^2 + k_y^2) \), that is \( E_\nu(r_\parallel, k_\parallel) = \varepsilon_\nu(r_\parallel) + \varepsilon_\parallel(k_\parallel) \) and the corresponding longitudinal velocity is \( v_\parallel = \frac{\hbar}{m^*} k_\parallel. \)

We assume that above a threshold energy \( E_T \) electrons are 3D and therefore only the subbands with \( E_\nu < E_T \) are retained. We will denote by \( \nu_T \) the threshold subband index. Then \( k_\parallel = (k_x, k_y) \in B_3^\nu \), where \( B_3^\nu = \{(k_x, k_y) \in \mathbb{R}^2 : 0 \leq \frac{\hbar^2}{2m^*} (k_x^2 + k_y^2) \leq E_T - \varepsilon_\nu \} \) is the selected Brillouin zone for 2D electrons in the \( \nu \)-th subband.

The Brillouin zone for the 3D electrons is \( B_3^\nu = \{(k_x, k_y, k_z) \in \mathbb{R}^3 : \mathcal{E}(k) \geq E_T \} \) where we are assuming a Kane dispersion relation \( \mathcal{E}(k) [1 + \alpha \mathcal{E}(k)] = \frac{\hbar^2 k^2}{2m^*} \) in order to take into account the effect of nonparabolicity at high energies.

Under the assumption that the channel length is no shorter than a few tenths of nanometers, the transport of the carriers is assumed to be well described by semiclassical Boltzmann equations. 2D electrons in each subband are considered as different populations and for each subband it is introduced a distribution function \( f_\nu(x_\parallel, k_\parallel, t) \) obeying the Boltzmann equation

\[
\frac{\partial f_\nu(x_\parallel, k_\parallel, t)}{\partial t} + v_\parallel \cdot \nabla_r f_\nu(x_\parallel, k_\parallel, t) - \frac{q}{\hbar} \mathbf{E}^{\nu ff}_r \cdot \nabla_k f_\nu(x_\parallel, k_\parallel, t) = \mathcal{C}^{2D}_\nu, \quad \mathbf{k} \in B_2^\nu
\]  

where \( \mathbf{E}^{\nu ff}_r = \frac{1}{q} \nabla_r \phi_\nu(r_\parallel) \) and \( \mathcal{C}^{2D}_\nu \) describes the scattering with phonons, including the mechanisms pushing 2D electrons into the 3DEG.

The 3DEG in the region \( \mathcal{R}_Q \) is described by the Boltzmann equation

\[
\frac{\partial f(x, k, t)}{\partial t} + v \cdot \nabla_r f(x, k, t) - \frac{q}{\hbar} \mathbf{E} \cdot \nabla_k f(x, k, t) = \mathcal{C}^{3D}, \quad \mathbf{k} \in B_3^\nu
\]

where \( v = \frac{1}{\hbar} \nabla_k \mathcal{E} = \frac{1}{m^*} \frac{\hbar k}{1 + 2\alpha \varepsilon} \) is the electron group velocity. \( \mathcal{C}^{3D} \) represents the scattering of 3D electrons with phonons, including the mechanisms pushing 3D electrons into the 2DEG. \( \mathbf{E} = -\nabla \phi \) is the self-consistent electric field which is related to the electron distributions function through Poisson’s equation \( \nabla (\varepsilon \nabla \phi) = -q (n_d(r) - n_T(r)) \), with \( \varepsilon \) the relative permittivity, \( n_d(r) \) the doping concentration and \( n_T(r) \) the total charge density given by \( n_T(r, t) = n(r, t) + \sum_{\nu=1}^{\nu_T} \rho_\nu(z, t) |\phi_\nu(z, t)|^2 \) with \( n(r, t) = \int_{B_3^\nu} f(x, k, t) d^2k \) the density of the bulk electrons, and \( \rho_\nu(r_\parallel, t) = \int_{B_2^\nu} f_\nu(r_\parallel, k_\parallel, t) d^2k_\parallel \) the areal density of electrons in the \( \nu \)-th subband.

All the main scattering processes have been considered, that is those due to acoustic phonons in the elastic approximation and non polar optical phonons, that cause the particles exchange between 2DEG and 3DEG. The reader is referred to [1] for the details.

3 The moment system and its closure by the MEP

We now write a system of moment equations deduced from Boltzmann transport equations under suitable closure relations. Let us define the generic moment associated with electrons in the subband \( \nu \) with respect to a weight function \( a(k_\parallel) \) as

\[
M_\alpha(r_\parallel, t) = \int a(k_\parallel) f(r_\parallel, k_\parallel, t) d^2k_\parallel.
\]
In particular we take as basic moments for the 2DEG the following ones

\[ \rho^\nu(r_\parallel, t) = \int_{B_\parallel^2} f_\nu(r_\parallel, k_\parallel, t) d^2k_\parallel \]

areal density  \( \rho^\nu (r_\parallel, t) \)

longitudinal mean velocity  \( V^\nu (r_\parallel, t) \)

longitudinal mean energy  \( W^\nu (r_\parallel, t) \)

longitudinal mean energy flux  \( S^\nu (r_\parallel, t) \)

The corresponding moment system is obtained by multiplying the Boltzmann equation by the weight functions entering into the definition of the fundamental moments and by integrating with respect to \( k_\parallel \). Explicitly we get

\[
\frac{\partial \rho}{\partial t} + \nabla_r \cdot (\rho \nu V_\nu) = \rho \nu C_{\rho \nu} + \rho \nu C_{\rho \nu}^{(no)} + \rho \nu C_{\rho \nu}^{(no),3D}
\]

\[
\frac{\partial (\rho \nu V_\nu)}{\partial t} + \nabla_r \cdot (\rho \nu F^{(0)\nu}) = \rho \nu C_{\nu V_\nu}^{(ac)} + \rho \nu C_{\nu V_\nu}^{(no)} + \rho \nu C_{\nu V_\nu}^{(no),3D}
\]

\[
\frac{\partial (\rho \nu W_\nu)}{\partial t} + \nabla_r \cdot (\rho \nu S_\nu) = \rho \nu C_{W \nu}^{(ac)} + \rho \nu C_{W \nu}^{(no)} + \rho \nu C_{W \nu}^{(no),3D}
\]

\[
\frac{\partial (\rho \nu S_\nu)}{\partial t} + \nabla_r \cdot (\rho \nu G^{(1)\nu}) = \rho \nu C_{S \nu}^{(ac)} + \rho \nu C_{S \nu}^{(no)} + \rho \nu C_{S \nu}^{(no),3D}
\]

where

\[
\left( \begin{array}{c}
F^{(0)\nu} \\
F^{(1)\nu}
\end{array} \right) = \frac{1}{\rho^\nu} \int_{B_\parallel^2} \left( \frac{1}{\varepsilon} \right) v_\parallel \otimes v_\parallel f_\nu(r_\parallel, k_\parallel, t) d^2k_\parallel,
\]

\[
\left( \begin{array}{c}
G^{(0)\nu} \\
G^{(1)\nu}
\end{array} \right) = -\frac{1}{\rho^\nu} \int_{B_\parallel^2} \left( \frac{1}{\hbar^2} \right) v_\parallel \otimes v_\parallel \nabla_k f_\nu(r_\parallel, k_\parallel, t) d^2k_\parallel,
\]

\[
\left( \begin{array}{c}
C_{\rho \nu} \\
C_{W \nu}
\end{array} \right) = \frac{1}{\rho^\nu} \int_{B_\parallel^2} \left( \frac{1}{\varepsilon} \right) \left[ S_{\mu \nu}(k_\parallel, k_\parallel) f_\nu - S_{\nu \mu}(k_\parallel, k_\parallel) f_\nu \right] d^2k_\parallel d^2k_\parallel,
\]

\[
\left( \begin{array}{c}
C_{\nu V_\nu} \\
C_{S \nu}
\end{array} \right) = \frac{1}{\rho^\nu} \int_{B_\parallel^2} \left( \frac{1}{\varepsilon} \right) \left[ S_{\mu \nu}(k_\parallel, k_\parallel) f_\nu - S_{\nu \mu}(k_\parallel, k_\parallel) f_\nu \right] d^2k_\parallel d^2k_\parallel,
\]

\[
L_0 = -\frac{1}{\hbar \rho^\nu} \nabla_{r_\parallel} \varepsilon f_\nu \int_{B_\parallel^2} \nabla_k f_\nu d^2k_\parallel,
\]

It is worth to underline here that the two last drift terms \( (L_0^\nu \text{ and } L_0^\nu) \) are due to the anisotropy of the distribution function. Usually they do not appear in the constitutive
equations of the carriers transport because the boundary of the first Brillouin zone is moved to infinity or the distribution function is symmetric on this boundary.

In a similar way, the basic moments we take for 3D electrons are the following ones

\[
\begin{align*}
n(r, t) &= \int_{B_3^*} f(r, k, t) d^3k \\
\mathbf{V}(r, t) &= \frac{1}{n(r, t)} \int_{B_3^*} \mathbf{v}(k) f(r, k, t) d^3k \\
W(r, t) &= \frac{1}{n(r, t)} \int_{B_3^*} \varepsilon(k) f(r, k, t) d^3k
\end{align*}
\]

and the corresponding moments system reads

\[
\begin{align*}
\frac{\partial n}{\partial t} + \nabla_r \left( n(r, t) \mathbf{v}(r, t) \right) &= nC_n^{(ac)} + nC_n^{(na)} + nC_n^{(na),2D} \\
\frac{\partial (n\mathbf{v})}{\partial t} + \nabla_r (n\mathbf{F}^{(0)}) + q\mathbf{E}(n\mathbf{G}^{(0)}) &= nC_V^{(ac)} + nC_V^{(na)} + nC_V^{(na),2D} \\
\frac{\partial (nW)}{\partial t} + \nabla_r (n\mathbf{S}) + q\mathbf{E}(n\mathbf{V}) &= nC_W^{(ac)} + nC_W^{(na)} + nC_W^{(na),2D} \\
\frac{\partial (n\mathbf{S})}{\partial t} + \nabla_r (n\mathbf{F}^{(1)}) + q\mathbf{E}(n\mathbf{G}^{(1)}) &= nC_S^{(ac)} + nC_S^{(na)} + nC_S^{(na),2D}
\end{align*}
\]

with analogous definition to the previous ones for \( \mathbf{F}^{(0)}, \mathbf{F}^{(1)}, \mathbf{G}^{(0)}, \mathbf{G}^{(1)} \) and the production terms.

The above written moment systems are not closed because there are more unknowns than equations. Therefore, constitutive relations in terms of the fundamentals variables are needed for extra fluxes and production terms. The maximum entropy principle leads to a systematic way for obtaining constitutive relations on the basis of information theory and has been widely used for semiconductor modeling (see for example [3, 4, 5, 6]).

According to the MEP, if a given number of moments of \( f \nu A \)

\[
M_{\nu A}(r_{||}, k_{||}, t), \quad A = 1, ..., N, \text{ and } \nu = 1, 2, ...
\]

are known along with a given number of moments of \( f \)

\[
M_{\nu B}(r, k, t) \quad B = 1, 2, ..., N
\]

the distribution functions \( f(r, k, t), f_{\nu}(r_{||}, k_{||}, t), \nu = 1, 2, ..., \) can be estimated by the extremal \((f ME P, f_1^{ME P}, f_2^{ME P}, ...)\) of the entropy functional under the constraints

\[
\begin{align*}
\int_{B_3^*} a_A(k_{||}) f_{ME P}(r_{||}, k_{||}, t) d\mathbf{k} &= M_{\nu A}(r_{||}, t) \quad A = 1, ..., N, \nu = 1, 2, ... \\
\int_{B_3^*} b_B(k) f^{ME P}(r, k, t) d^3k &= M_{\nu B}(r, t) \quad B = 1, 2, ..., N
\end{align*}
\]
Following the approach in [3, 4, 5], we assume the following definition

**Definition 1** We define the entropies of the two subsystems, 2DEG and 3DEG, as

\[
S_{2D} = -k_B \sum_{\nu=1}^{+\infty} |\phi_\nu(z,t)|^2 \int_{B_3^*} \left( f_\nu \log \frac{f_\nu}{y} - f_\nu \right) d^2k_\|, \quad y = \frac{2}{(2\pi)^2},
\]

\[
S_{3D} = -k_B \int_{B_3^*} f(k) \left[ \log f(k) - 1 \right] d^3k.
\]

The total entropy is of course \( S = S_{2D} + S_{3D} \). Therefore, according to MEP and our choice of the basic moments in the case we are dealing with, \( f \) and the \( f_\nu \)'s are estimated with the distributions \( f^{\text{MEP}} \) and \( f^{\nu\text{MEP}} \)'s that solve the problem:

maximize \( S \) under the constraints

\[
\int_{B_3^*} f^{\text{MEP}}(r, k, t)d^3k = n(r, t), \quad \int_{B_3^*} v(k)f^{\text{MEP}}(r, k, t)d^3k = n(r, t)V(r, t) \quad (11)
\]

\[
\int_{B_3^*} \varepsilon(r, k, t)d^3k = n(r, t)W(r, t), \quad \int_{B_3^*} \varepsilon f^{\text{MEP}}(r, k, t)d^3k = n(r, t)S(r, t) \quad (12)
\]

and for \( \nu = 1, 2, \ldots \)

\[
\int_{B_2^*} f_\nu(r_\|, k_\|, t)d^2k_\| = \rho^{\nu}(r_\|, t), \quad \int_{B_2^*} v_\|f_\nu(r_\|, k_\|, t)d^2k_\| = \rho^{\nu}(r_\|, t)V^{\nu}(r_\|, t) \quad (13)
\]

\[
\int_{B_2^*} \varepsilon_\|f_\nu(r_\|, k_\|, t)d^2k_\| = \rho^{\nu}(r_\|, t)W^{\nu}(r_\|, t), \quad (14)
\]

\[
\int_{B_2^*} \varepsilon_\|v_\|f_\nu(r_\|, k_\|, t)d^2k_\| = \rho^{\nu}(r_\|, t)S^{\nu}(r_\|, t) \quad (15)
\]

One has

\[
f^{\nu\text{MEP}} = \exp \left[ - \left( \lambda^{\nu} + \lambda_{W}^{\nu} \cdot v_\| + \lambda_{S}^{\nu} \cdot v_\|\varepsilon_\| \right) \right] \quad \nu = 1, 2, \ldots \quad (16)
\]

\[
f^{\text{MEP}} = \exp \left[ - (\lambda + \lambda v \cdot v + (\lambda_{W} + \lambda_{S} v)\varepsilon) \right] \quad (17)
\]

Now, following the same approach as in [4, 5], we assume a small anisotropy of the distribution functions and expand them up to first order with respect to the lagrangian multipliers relative to velocity and energy-flux

\[
f^{\nu\text{MEP}} \approx \exp \left( -\lambda^{\nu} - \lambda_{W}^{\nu}\varepsilon_\| \right) \left[ 1 - \left( \lambda_{W}^{\nu} \cdot v_\| + \lambda_{S}^{\nu} \cdot v_\|\varepsilon_\| \right) \right], \quad (18)
\]

\[
f^{\text{MEP}} \approx \exp(-\lambda^{\nu} - \lambda_{W}^{\nu}\varepsilon_\| \left[ 1 - (\lambda_{W}^{\nu} \cdot v + \lambda_{S}^{\nu} \cdot v\varepsilon) \right] \quad (19)
\]

Inserting the above-written expansions into the constraints (11)-(15), it is possible to get analytical explicit expressions of the lagrangian multipliers and in turn to get the closure relations for the moment system.
4 Closure relations for the 2DEG

The Lagrange’s multipliers of the subbands into the 2DEG are

\[
\begin{align*}
\rho^\nu &= \frac{2\pi m^*}{\hbar^2} e^{-\lambda^\nu} I_{0}^\nu, \\
W^\nu &= \frac{1}{\lambda_W^\nu} \left[ 1 + \frac{\lambda_W^\nu (E_T - \varepsilon^\nu)}{1 - e^{\lambda_W^\nu (E_T - \varepsilon^\nu)}} \right], \\
\lambda_V^\nu &= b_{11}(W^\nu) V^\nu + b_{12}(W^\nu) S^\nu, \\
\lambda_S^\nu &= b_{21}(W^\nu) V^\nu + b_{22}(W^\nu) S^\nu
\end{align*}
\]

(20)

where

\[
\begin{align*}
b_{11} &= -\frac{m^* T_0^\nu}{\Delta^\nu} T_3^\nu, \\
b_{22} &= -\frac{m^* T_0^\nu}{\Delta^\nu} T_1^\nu, \\
b_{12} &= \frac{m^* T_0^\nu}{\Delta^\nu} T_2^\nu, \\
\Delta^\nu &= T_1^\nu T_3^\nu - (T_2^\nu)^2
\end{align*}
\]

with

\[
\begin{align*}
T_0^\nu &= 1 - e^{-\lambda_W^\nu (E_T - \varepsilon^\nu)}, \\
T_n^\nu &= (-1)^n \frac{d^n}{d(\lambda_W^\nu)^n} T_0^\nu, \quad n = 0, 1, 2, \ldots
\end{align*}
\]

Once the lagrangian multipliers have been obtained, we can calculate the extra fluxes and the drift and production terms. The fluxes read

\[
\begin{align*}
F^{(0)}^\nu &= \frac{T_1^\nu}{m^* T_0^\nu} I, \\
F^{(1)}^\nu &= \frac{T_2^\nu}{m^* T_0^\nu} I
\end{align*}
\]

(22)

for the other expressions the reader is referred to [1]. A crucial point is that there exists [1] the critical values \( W^\nu = \frac{E_T}{2} - \varepsilon^\nu \) such that \( \lambda_W^\nu (W^\nu) = 0 \). Therefore the allowed values of the longitudinal average energy are \( k_B T_L < W^\nu < W^\nu_c \).

5 Closure relations for the 3DEG

In the same manner as for the 2DEG, explicit formulas for the closure relations of the 3DEG part of the moment system are obtained. The lagrangian multipliers are given by the following relationships

\[
\begin{align*}
n &= \frac{4\pi m^*}{\hbar^3} e^{-\lambda^B} \mathcal{I}(E_T, \lambda_W^B), \\
W &= -\frac{d}{d \lambda_W^B} \ln \mathcal{I}(E_T, \lambda_W^B) \\
\lambda^B_V &= B_{11} V + B_{12} S, \\
\lambda^B_S &= B_{21} V + B_{22} S
\end{align*}
\]

(23)

(24)

with

\[
\begin{align*}
B_{11} &= -\frac{3m^*}{2\Delta} \mathcal{I}(E_T, \lambda_W^B), \\
B_{12} &= B_{21} = \frac{3m^*}{2\Delta} \mathcal{I}(E_T, \lambda_W^B), \\
\Delta &= \mathcal{L}^{(2)}(E_T, \lambda_W^B) - (\mathcal{L}^{(1)}(E_T, \lambda_W^B))^2 \\
\mathcal{I}(x, \beta) &= \int_x^{+\infty} (1 + 2\alpha \varepsilon) \sqrt{\varepsilon(1 + \alpha \varepsilon)} e^{-\beta \varepsilon} d\varepsilon \\
\mathcal{L}^{(0)}(x, \beta) &= \int_x^{+\infty} e^{-\beta \varepsilon} [\varepsilon(1 + \alpha \varepsilon)]^{3/2} d\varepsilon, \quad \mathcal{L}^{(n)}(x, \beta) = (-1)^n \frac{d^n}{d(\beta)^n} \mathcal{L}^{(0)}(x, \beta)
\end{align*}
\]
Once $\lambda_W$ has been obtained via a numerical inversion, the extra fluxes and production terms can be obtained. Here we show only the flux expressions (see [1] for others details).

\[ F^{(0)} = \frac{2}{3m^*T(E_T, \lambda_W^B)} C^{(0)}(E_T, \lambda_W^B), \quad F^{(1)} = \frac{2}{3m^*T(E_T, \lambda_W^B)} C^{(1)}(E_T, \lambda_W^B) \]

At variance with the 2DEG, no upper limit for the energy density arises.

6 Mathematical structure of the moment system closed with MEP

We want to give a strong numerical evidence that the moment system of the subbands and bulk electrons augmented with the MEP closure relations forms a quasilinear hyperbolic system in the time direction in the physically relevant range of $W^\nu$. This preliminary analysis is crucial for the development of appropriate numerical schemes.

Since the differential part of each subband and of the 3DEG is decoupled in the moment system, we can limit our analysis to the study of a single subband and the 3DEG. Let us consider the quasilinear system of PDEs

\[ \frac{\partial}{\partial t} F^{(0)}(U) + \sum_{i=1}^{2} \frac{\partial}{\partial x^i} F^{(i)}(U) = P(U, x, t), \tag{25} \]

with $U(x, t)$ vector field belonging to a connected open set $\Omega \subset \mathbb{R}^m$, $\forall t > 0$ and $\forall x$ belonging to a domain $D \subseteq \mathbb{R}^k$ with $k = 2$ for th 2DEG or $k = 3$ for the 3DEG, and $F^{(\beta)} : \Omega \mapsto \mathbb{R}^m$, $\beta = 0, \ldots, k$ sufficiently smooth functions. Defining the Jacobian matrices

\[ A^{(\beta)} = \nabla_U F^{(\beta)}, \quad \beta = 0, \ldots, k, \]

we recall that the system (25) is said to be hyperbolic in the $t$-direction if $\det(A^{(0)}(U)) \neq 0$ and the eigenvalue problem

\[ \det \left( \sum_{i=1}^{2} n_i A^{(i)}(U) - \lambda A^{(0)}(U) \right) = 0 \tag{26} \]

has real eigenvalues and the eigenvectors span $\mathbb{R}^m$ for all unit vectors $n = (n_1, \ldots, n_k)$ of $\mathbb{R}^k$. Will first treat the case of a generic subband and then the 3DEG case.

6.1 Iperbolicity of the generic subband subsystem

In the case under consideration, by omitting the subband index, we have

\[ U = \begin{pmatrix} \rho \\ V^1 \\ V^2 \\ W \\ S^1 \\ S^2 \end{pmatrix}, \quad F^{(0)} = \rho \begin{pmatrix} 1 \\ V^1 \\ V^2 \\ W \\ S^1 \\ S^2 \end{pmatrix}, \quad F^{(1)} = \rho \begin{pmatrix} V^1 \\ F^{(0)} \\ 0 \\ S^1 \\ F^{(1)} \\ 0 \end{pmatrix}, \quad F^{(2)} = \rho \begin{pmatrix} V^2 \\ 0 \\ S^2 \\ 0 \\ F^{(1)} \end{pmatrix}, \]

522
and the Jacobian matrices are given by

\[
A^{(0)} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix},
A^{(n)} = \sum_{i=1}^{2} n_i A^{(i)} = \begin{pmatrix}
n \cdot V & n_1 \rho & n_2 \rho & 0 & 0 & 0 \\
n_1 F^{(0)} & 0 & 0 & n_1 \rho (F^{(0)})' & 0 & 0 \\
n_2 F^{(0)} & 0 & 0 & n_2 \rho (F^{(0)})' & 0 & 0 \\
n_1 F^{(1)} & 0 & 0 & n_1 \rho (F^{(1)})' & 0 & 0 \\
n_2 F^{(1)} & 0 & 0 & n_2 \rho (F^{(1)})' & 0 & 0
\end{pmatrix},
\]

where the prime denotes partial derivation with respect to \(W\).

The equation

\[
\det (A^{(n)} - \lambda A^{(0)}) = 0
\]
gives the eigenvalues

\[
\lambda_{1,2} = 0, \quad \text{with multiplicity 2}
\]

\[
\lambda_{3,4,5,6} = \pm \sqrt{a(W) \pm \sqrt{a(W)^2 - 4b(W)}}
\]

where

\[
a(W) = F^{(0)} + (F^{(1)})' - W (F^{(0)})', \quad b(W) = F^{(0)} (F^{(1)})' - (F^{(0)})' F^{(1)}.
\]

In Fig. 2 the eigenvalues \(\lambda_{3,4,5,6}\) are plotted against the longitudinal mean energy \(W\) for several values of \(E_T - \varepsilon_\nu\) in the range belonging to the realizability region. Since the four eigenvalues \(\lambda_{3,4,5,6}\) are real and distinct, each of them has a corresponding eigenspace of dimension one.

Concerning the eigenvalue \(\lambda = 0\), we observe that whatever \(n\) we take the first and fourth rows of \(A^{(n)}\) are linearly independent, the second and third rows are proportional and similarly the last two rows since \(\rho > 0\) and \(n_1\) and \(n_2\) cannot be both zero. We observe that

\[
\det \begin{pmatrix}
F^{(0)} \\
F^{(1)}
\end{pmatrix} = \rho b(W).
\]

The fact that the eigenvalues \(\lambda_{3,4,5,6}\) are real implies that \(b(W) > 0\) and therefore the rank of \(A^{(n)}\) is four which means that the eigenspace associated to \(\lambda = 0\) has dimension two, leading to the hyperbolicity of the system (25) in the physical region \(\rho > 0\) and \(k_B T_L < W^\nu < W^\nu_c\).

In the one dimensional case one has only the eigenvalues \(\lambda_{3,4,5,6}\) and by similar computations the hyperbolicity is again guaranteed.

**Remark 1** *In the particular case \(E_T \to \infty\) we explicitly have

\[
\lambda_{3,4,5,6} = \pm \sqrt{2 \pm \sqrt{2}} \frac{W}{m^*}
\]

which are real and distinct provided \(W > 0\) according to [4].
Figure 2: Plot of $\lambda_{3,4,5,6}$ versus the longitudinal mean energy for $E_T - \varepsilon_\nu = 0.125$ (upper left), 0.15 (upper right), 02 (bottom left), 03 (bottom right) eV.

6.2 Hyperbolicity of the 3DEG subsystem

In a similar way we have for the 3DEG

$$U = \begin{pmatrix} \rho \\ V^1 \\ V^2 \\ V^3 \\ W \\ S^1 \\ S^2 \\ S^3 \end{pmatrix}, \quad \mathcal{F}^{(0)} = \begin{pmatrix} 1 \\ V^1 \\ V^2 \\ V^3 \\ W \\ S^1 \\ S^2 \\ S^3 \end{pmatrix}, \quad \mathcal{F}^{(1)} = \begin{pmatrix} V^1 \\ F^{(0)} \\ 0 \\ 0 \\ S^1 \\ F^{(1)} \\ 0 \\ 0 \end{pmatrix},$$

$$\mathcal{F}^{(2)} = \begin{pmatrix} V^2 \\ 0 \\ F^{(0)} \\ 0 \\ S^2 \\ 0 \\ F^{(1)} \\ 0 \end{pmatrix}, \quad \mathcal{F}^{(3)} = \begin{pmatrix} V^3 \\ 0 \\ 0 \\ F^{(0)} \\ S^3 \\ 0 \\ 0 \\ F^{(1)} \end{pmatrix}.$$
and the Jacobian matrices are given by

\[ A^{(0)} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
V^1 & n & 0 & 0 & 0 & 0 & 0 \\
V^2 & 0 & n & 0 & 0 & 0 & 0 \\
V^3 & 0 & 0 & n & 0 & 0 & 0 \\
W & 0 & 0 & 0 & n & 0 & 0 \\
S^1 & 0 & 0 & 0 & 0 & n & 0 \\
S^2 & 0 & 0 & 0 & 0 & 0 & n \\
S^3 & 0 & 0 & 0 & 0 & 0 & n
\end{pmatrix}, \]

\[ A^{(n)} = \sum_{i=1}^{2} n_i A^{(i)} = \begin{pmatrix}
\mathbf{n} \cdot \mathbf{V} & n_1 n & n_2 \rho & n_3 n & 0 & 0 & 0 \\
n_1 F^{(0)} & 0 & 0 & 0 & n_1 n (F^{(0)})' & 0 & 0 \\
n_2 F^{(0)} & 0 & 0 & 0 & n_2 n (F^{(0)})' & 0 & 0 \\
n_3 F^{(0)} & 0 & 0 & 0 & n_3 n (F^{(0)})' & 0 & 0 \\
\mathbf{n} \cdot \mathbf{S} & 0 & 0 & 0 & 0 & n_1 n & n_2 n & n_3 n \\
n_1 F^{(1)} & 0 & 0 & 0 & n_1 n (F^{(1)})' & 0 & 0 \\
n_2 F^{(1)} & 0 & 0 & 0 & n_2 n (F^{(1)})' & 0 & 0 \\
n_3 F^{(1)} & 0 & 0 & 0 & n_3 n (F^{(1)})' & 0 & 0
\end{pmatrix}, \]

where the prime denotes partial derivation respect to \( W \). The equation

\[ \det (A^{(n)} - \lambda A^{(0)}) = 0 \]

gives the eigenvalues

\[ \lambda_{1,2,3,4} = 0, \quad \text{with multiplicity 4} \quad (29) \]

\[ \lambda_{5,6,7,8} = \pm \sqrt{a(W) \pm \sqrt{a(W)^2 - 4b(W)}} \quad (30) \]

where, assuming the same notation of the previous case,

\[ a(W) = F^{(0)} + (F^{(1)})' - W (F^{(0)})', \quad b(W) = F^{(0)} (F^{(1)})' - (F^{(0)})' F^{(1)}. \]

In Fig. 3 the eigenvalues \( \lambda_{5,6,7,8} \) are plotted against the longitudinal mean energy \( W \) for several values of \( E_T \). Since the four eigenvalues \( \lambda_{5,6,7,8} \) are real and distinct, each of them has a corresponding eigenspace of dimension one.

Concerning the eigenvalue \( \lambda = 0 \), we use arguments similar to that used previously and, assuming \( n_1 = 1, n_2 = n_3 = 0 \), observe that

\[ \det \begin{pmatrix}
\mathbf{n} \cdot \mathbf{V} & n_1 n & 0 & 0 \\
n_1 F^{(0)} & 0 & n_1 n (F^{(0)})' & 0 \\
n_2 F^{(0)} & 0 & 0 & n_1 n \\
n_3 F^{(1)} & 0 & n_1 n (F^{(1)})' & 0
\end{pmatrix} = n^3 b(W). \]

The fact that the eigenvalues \( \lambda_{5,6,7,8} \) are real implies that \( b(W) > 0 \) and therefore the rank of \( A^{(n)} \) is four which means that the eigenspace associated to \( \lambda = 0 \) has dimension four, leading to the hyperbolicity of the system (25) in the physical region \( n > 0 \).
REFERENCES


THE COUPLED PROBLEM AMONG DISPLACEMENT, THERMO AND ELECTROMAGNETIC FIELDS AND APPLICATIONS IN ELECTRONIC EQUIPMENTS

B. Y. DUAN and P. Y. LIAN

Key Laboratory of Electronic Equipment Structure Design of Ministry of Education, China Research Institute on Mechatronics, Xidian University, Xi’an, 710071, China

Key words: Multi-field-coupling Problem, High-density Packaging System, Information Transferring Technology, Model Simplification, Integrated Optimization Design.

Abstract. With the rapid development of electronic equipment, coupling problem of the displacement field, electromagnetic field and temperature field is increasingly prominent. This paper focuses on the high-density packaging system, which is a kind of typical electromechanical-thermal coupling equipment. To begin with, the mathematical model is established and the electrical performance is expressed as a function of structural design variables and temperature. Then, the simplification method of the electromagnetic model and the transfer technology of grid information among different physical models are discussed. Last, the mathematical model for the optimization design based on the electro-mechanical-thermal coupling model is constructed, and satisfactory results have been obtained in the application of the optimization design model to the design of a high-density cabinet.

1 INTRODUCTION

Microwave electronic equipment (reflector antenna, planar slotted antenna, active phased array radar, high-density packaging system, etc.) is widely used in various fields such as land, sea, air and space, and it is a typical combination of mechanical and electrical systems. The mechanical structure is not only the carrier and security of the electrical performance, but also restricts the improvement of the performance, showing a strong coupling characteristic between the electricity and mechanic[1]. With the development of electronic devices toward high frequency and high gain, high density and miniaturization, and fast response and high pointing precision, the performance is not only dependent on the levels of the designs of the various disciplines, but also much more on the interdisciplinary intersection and mergence, showing the multi-field coupling characteristics[2,3].

Multi-field coupling problem exists widely in the engineering practice, and has been in-depth researched in many areas and numerous valuable results have been obtained, such as the coupling contact problem researched by using the boundary element method[4,5], the coupling among different electromagnetic fields in the electronic equipment[6,7], the fluid-structure interaction problems studied with the grid matching method[8]. But for the electronic equipment, especially for the high-density packaging system, the research is still less. With the development of modern electronic devices toward miniaturization, high density, high

Project supported by the National Natural Science Foundation of China (Grant No. 51035006)
frequency direction, on the one hand, the higher frequency makes electromagnetic compatibility problems of electronic devices more prominent: not only itself needs to resist against external electromagnetic environmental, but also cannot interfere the other electronic equipment nearby\cite{9,10}. On the other hand, high density will make the cooling difficult and the high temperature would affect the performance of electronic devices. Besides, while electronic equipment is being miniaturized, the requirement of structural strength and stiffness should also be ensured and the devices must be able to work properly in the variety of conditions and harsh environment. To this end, it is necessary to in-depth research the problem from the perspective of the multi-field coupling problem and the overall system, and further to propose the field coupling theory and perform the multidisciplinary design optimization.

For the high-density packaging system, one typical kind of electro-mechanical-thermal coupling equipment, this paper establishes the multi-field coupling model and discusses the theory and method of the simplification of electromagnetic model and the transfer of grid information among different physical models. Then, based on the coupling model, the integrated optimization design model is presented. Results show that both the multi-field-coupled model and integrated optimization design proposed here are correct and valid.

2 COUPLED MODEL AMONG ELECTROMAGNETIC, TEMPERATURE, AND STRUCTURE FIELDS

For the high-density packaging cabinet, the electromagnetic (EM) shielding effectiveness is one key performance. Assume that there are \( M \) electronic devices in its box and \( e_i \) is the electric field intensity generated by the \( i \)th device. \( P \) is of the distance of \( d \) from the center of the box. Let the field intensity amplitude at \( P \) be \( \left| \sum_{i=1}^{M} E_i(e_i) \right| \) with box and \( \left| \sum_{i=1}^{M} E_i^0(e_i) \right| \) without box. Then the EM shielding effectiveness (SE) is

\[
SE = 20 \log \left( \left| \sum_{i=1}^{M} E_i(e_i) \right| / \left| \sum_{i=1}^{M} E_i^{0}(e_i) \right| \right)
\]

In engineering practice, there are a lot of factors that affect SE. Assume that the cabinet has \( n \) leak paths and SE\(_i\) is the shielding effectiveness of the \( i \)th leak, then the SE considering all the factors is\cite{1,11}

\[
SE_{db} = -20 \log \left( \sum_{i=1}^{n} 10^{-SE_i/20} \right)
\]

The physical meaning of formula (2) is that SE of the cabinet is mainly determined by the leakage factor with the lowest SE. But there are several shortages. Firstly, when considering only one single factor, the analytical precision is limited and the influencing factors not able to be expressed by a formula, say, the gap closed by conductive rubber, is difficult to be considered. The second shortage is that the coupling characteristic of the electromagnetic field, structural displacement and temperature of electronic equipment is ignored, for example, the cabinet would be deformed under loads and the characteristic of electromagnetic radiation would be changed under different temperature. Next, the model of high-density packaging cabinet will be derived from the angle of the multi-field coupling analysis.
2.1 The influence of structural parameters and deformation

Various holes and seams on the cabinet’s surface are partial main factors that affect SE. Structural parameter $\beta$ involves the wall thickness, stiffeners, the size and position of internal partitions, vents, and seams, the position of the internal electromagnetic devices, and so on. The external electric field is a function of $\beta$. Under the external loads, the cabinet is deformed resulting in the change of the induced current on the surface acting as the electromagnetic boundary and then SE will be to some extent affected. Although the effect is small, it is still well worth being researched in theoretical study. In addition, structural deformation will also lead to the position change of the internal devices and then affect the leakage of electric field. Considering structural deformation $\delta$ is the function of $\beta$, the leakage electric field can be written as

$$\sum_{i=1}^{M} E_i(e_i, \delta(\beta))$$  \hspace{1cm} (3)

And the SE can be described as

$$SE = 20\log\left(\frac{\sum_{i=1}^{M} E_i^0(e_i)}{\sum_{i=1}^{M} E_i(e_i, \delta(\beta))}\right)$$  \hspace{1cm} (4)

2.2 The influence of temperature

While the cabinet working, the internal devices generate heat and their performance will be affected by the changing temperature to a certain extent. So the electric field strength generated by the electromagnetic devices can be considered to be a function of the temperature $T$. Generally, the influence relationship will be described by a curve or chart that can be used by looking up in according to the temperature. Besides, the influence of temperature on the structural deformation can be written as $\delta(\beta, T)$, so the leakage electric field can also be expressed as

$$\sum_{i=1}^{M} E_i(e_i(T), \delta(\beta, T))$$  \hspace{1cm} (5)

Then SE is

$$SE = 20\log\left(\frac{\sum_{i=1}^{M} E_i^0(e_i)}{\sum_{i=1}^{M} E_i(e_i(T), \delta(\beta, T))}\right)$$  \hspace{1cm} (6)

2.3 The influence of contact gap

In the cabinet with high SE, the contact gap is generally closed by such materials as conductive gasket to ensure electrical continuity of the contact surface in order to reduce the leakage. For the gaps closed by conductive gaskets, it can be assumed that they can be replaced by some equivalent conductive materials with the same leakage electric fields. According to the transfer impedance of the contact gap or conductive rubber obtained by
testing and their structural parameters, the electromagnetic parameters, such as conductivity, can be derived. By now, the geometric model of the cabinet can be established in the simulation software and then the field distribution near the cabinet can also be calculated after the material parameters and excitation source being set.

Here suppose that \( W_1 \) and \( W_2 \) represent the height of the gap and filling material respectively. The transfer impedance is \( Z \) and the area of the gap is \( S = \text{length} \times \text{width} \). Then the conductivity (\( \sigma_T \)) of filling material is

\[
\sigma_T = \sqrt{W_1 W_2 / SZ}
\]

Then the leakage electric field can be calculated with the material parameters above. For the conductive rubber, the compressed height \( h \) can also be used directly, the conductivity is

\[
\sigma_T = h / SZ
\]

The leakage electric field of the cabinet with contact gaps is the function of transfer impedance \( Z \) and simultaneously \( Z \) is closely related to the frequency \( \text{freq} \), so the leakage electric field can be expressed as

\[
\sum_{i=1}^{M} E_i(e_i(T), \delta(\beta,T), Z(\text{freq}))
\]

Eventually, SE can be further rewritten as \([12, 13]\),

\[
SE = 20 \log \left( \frac{\sum_{i=1}^{M} E_i^0(e_i)}{\sum_{i=1}^{M} E_i(e_i(T), Z(\text{freq}), \delta(\beta,T))} \right)
\]

The above formula is just the electro-mechanical-thermal coupling model of the high-density packaging cabinet. It reflects the interrelationship among structure field, electromagnetic field and temperature field, and the electrical performance is expressed as a function of structural design variables (e.g., structural size, shape, topology, and type), which makes it possible to perform the electro-mechanical-thermal integrated optimization design.

3 SOLUTION OF MULTI-FIELD COUPLING MODEL

Sequential coupling analysis method is an iterative method solving each field in turn within a time step and transferring coupling information among different fields, and it is applicable to low non-linear coupling problems \([14, 15]\). Because of its advantage of separately modeling, easily using respective simulation tools, software reuse and modularizing, the method is widely used in solving coupling problems. Of course, this method is used to solve the multi-field coupling model in this paper.

3.1 Simplification of electromagnetic model

Electromagnetic simulation needs to create a geometric model of the electronic equipment. However the actual model is very complicated with a lot of fine structures, such as various bosses, grooves, counterbores, and so on. If the simulation model is established in accordance with the actual model, it will not only dramatically increase the computational workload, but
also it is not necessary in engineering. For this reason, this paper presents a method for the simplification of the tiny structures based on perturbation theory.

Generally, the SE would be worst at the resonant frequency. So the simplified model should be able to reflect the actual resonant frequency and ensure its offset within the permissible error range. Because simplification mainly affects the offset of the resonant frequency, if let the offset \(|\Delta \omega / \omega_0| < \varepsilon\), the volume \(\Delta V\) of the fine structure can be ignored only when \(\Delta V < \varepsilon V_0 / f(x,y)\). Where, \(f(x,y)\) is a proportionality constant related to the position of the perturbation and the resonance mode excited in the cavity and would act as a dominant role when the rectangular cabinet works at the frequency lower than 1GHz. Engineering experience shows that the general range for the threshold \(\varepsilon\) is \(0 < \varepsilon < 0.01\). Let \(\varepsilon = 0.01\), it means that the change of the resonant frequency after the perturbation is less than 1% of the original resonant frequency. Now the volume of the tiny structure that can be ignored is \(\Delta V \leq \varepsilon V_0 / f(x,y)\), where, \(V_0\) is the original chamber volume.

### 3.2 Transfer of grid information

The transfer of grid information among different physical fields is a key problem in the coupling analysis. Generally, because of the different forms and precision of grids of different fields, the grid mismatch at the common interface of the different physical fields occurs and grid overlap and gap will also occur at the same time [2]. Fig.1 shows the mismatch in a two-dimensional case, where, \(\Gamma\) is the common smooth interface, \(\Gamma_A\) and \(\Gamma_B\) the discrete boundaries of different fields. In the coupling analysis, the information transfer relationship of \(\Gamma_A\) and \(\Gamma_B\) must be established through the method of mathematical physics.

![Figure 1: Schematic diagram of the grid mismatch](image)

Electronic equipment usually involves the displacement field, electromagnetic field and temperature field, it’s necessary to transfer coupling information among physical fields. The most common way is to transfer the structure deformation to the electromagnetic analysis module and temperature analysis module, and then to carry out the calculation of temperature
and electromagnetic field. In this paper, PRO/E is used for parametric modeling, ANSYS for structural analysis, FEKO for electromagnetic analysis and ICEPAK for temperature analysis.

In the structural finite element analysis, shell element mainly includes triangular element and quadrilateral element, and solid element involves tetrahedral cell and hexahedron element. It is known that the software ICEPAK used for temperature calculation can recognize the model assembled by triangular planes, the node coordinates and cell information of the outer surface of the deformed structure grid can be used to produce triangular planes and then generate the corresponding grid files that can be identified by FEKO and ICEPAK software respectively, ultimately achieving the transfer of grid information [13], as shown in Fig.2.

To begin with, based on the grid files proposed above, the thermal analysis model and electromagnetic analysis model can be produced. Then, temperature analysis will be performed. According to the temperature analysis results, the powers of the electromagnetic devices in the model can be determined, and the electric parameters of contact gap or conductive rubber are calculated at the same time. Lastly, the electromagnetic analysis can be able to be carried out with the electromagnetic analysis model, and thus the multi-field coupling analysis is realized.

![Figure 2: Procedure diagram of the transfer of deformation information](image)

4 MATHEMATICAL DESCRIPTION OF INTEGRATED OPTIMIZATION DESIGN

Structural strength, ventilation and cooling, and electromagnetic compatibility are the three main aspects in the design of high-density packaging cabinets, and there are contradictions among each other. The first contradiction exists between the mass and strength. The requirement for structural strength is very strict to ensure that it can work properly under various conditions of shock and vibration, but the working environment requires small size and light weight, especially for the airborne and missile-borne equipment. For another kind of electric equipment such as radar and antenna, the contradiction between the mass and stiffness is the major one. Second one exists between the ventilation and electromagnetic shielding effectiveness. Larger hole or seam is conducive to cooling but not to SE, and high temperature would affect the performance of electronic devices. In order to meet all the requirements at same time, it is necessary to in-depth research the problem from the angle of the multi-field coupling problem and the overall system, and to perform the multidisciplinary integrated optimization design.

The establishment and solution of the multi-field coupling model make the integrated optimization design possible, thus the requirements can be met simultaneously for structural stiffness and strength, ventilation and cooling, and electromagnetic compatibility. Based on formula (10), the mathematical description of the electro-mechanical-thermal coupling
optimization model of high-density packaging system can be expressed as follows \cite{12,13,16},

$$
\begin{align*}
\text{find } & \quad \beta = (\beta_1, \beta_2, \ldots, \beta_{N_d})^T \\
\text{min } & \quad W(\beta) \\
\text{s.t. } & \quad -SE(\beta) + SE^0 \leq 0 \\
& \quad -f_{\text{eigen}} + f_{\text{eigen}}^0 \leq 0 \\
& \quad T_{j_{\text{max}}}(\beta) - T_{j_{\text{max}}^0} \leq 0, \quad j = 1, 2, \ldots, N_u \\
& \quad \sigma_i - [\sigma] \leq 0, \quad i = 1, 2, \ldots, N_s \\
& \quad \beta_{k_{\text{min}}} \leq \beta_k \leq \beta_{k_{\text{max}}}, \quad k = 1, 2, \ldots, N_d
\end{align*}
$$

(11)

Where, $W(\beta)$ is the structural weight, $SE(\beta)$ and $SE^0$ the actual value and minimum allowable value of SE, $f_{\text{eigen}}$ and $f_{\text{eigen}}^0$ the actual value and minimum allowable value of natural frequency, $T_{j_{\text{max}}}(\beta)$ and $T_{j_{\text{max}}^0}$ the actual value and maximum allowable value of the temperature at the $j$th point, $\beta_{k_{\text{max}}}$ and $\beta_{k_{\text{min}}}$ the upper and lower limits of variable $\beta_k$, $N_d$, $N_s$, and $N_u$ the total numbers of the design variables, temperature constraints and stress constraints respectively.

Here the weight is used as the objective function, besides, SE or both can also be used as the objective function according to the actual situation. Investigating the above programming problem leads us to the following points. The coupling optimization model is a highly nonlinear programming problem, since the objective and constraints are all highly nonlinear functions of the design variables. So the optimization algorithm and efficiency are two main topics needed to be further researched.

5 ENGINEERING APPLICATION AND DISCUSSION

In this section, the optimization design of an airborne electronic equipment cabinet is performed with the multi-field coupling optimization model \cite{13,17}. The box structure is shown in Fig.3. The length, width and height of the aluminum box are 575mm, 482mm, and 532mm respectively. The box is divided into two parts inside: the upper part is installed with 12 PCB planes and two powers, while the lower part is ventilation duct with an inclined wind shield. The upper part of the box front panel has two groups of radiating holes and its lower part has two fans. Besides, three fans are installed in the upper part of the back panel. The box is installed in plane, so there is a requirement for the fundamental frequency and weight. Thus, the goal of this paper is to reduce the weight, and the highest temperature inside the cabinet, SE, the first natural frequency and maximum stress all act as constrains, the size and position of the holes, the wall thickness of the cabinet, and wind shield length are design variables.

There are 8 design variables which can be divided into three kinds. The first category is the position and size of the holes, including the aperture length L and width W, distance d1 from hole to left side plate, distance d2 from hole to the power supply, distance d3 between holes and distance d4 between hole columns. The second category is the wall thickness D1 and the third is the length D2 of the wind deflector. Thus design variables can be expressed as follows,

$$
\beta = (L, W, d1, d2, d3, d4, D1, D2)^T
$$

(12)
**Figure 3**: Schematic diagram of an airborne cabinet

**Table 1**: Coupling optimization results

<table>
<thead>
<tr>
<th>Item</th>
<th>Name</th>
<th>Initial Value</th>
<th>Optimal Value</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>D1</td>
<td>4.5mm</td>
<td>3.75mm</td>
<td>2mm</td>
<td>6mm</td>
</tr>
<tr>
<td>Variable</td>
<td>D2</td>
<td>300mm</td>
<td>50mm</td>
<td>50mm</td>
<td>400mm</td>
</tr>
<tr>
<td>Variable</td>
<td>L</td>
<td>10mm</td>
<td>20mm</td>
<td>5mm</td>
<td>30mm</td>
</tr>
<tr>
<td>Variable</td>
<td>W</td>
<td>10mm</td>
<td>5mm</td>
<td>5mm</td>
<td>50mm</td>
</tr>
<tr>
<td>Variable</td>
<td>d1</td>
<td>20mm</td>
<td>50mm</td>
<td>5mm</td>
<td>80mm</td>
</tr>
<tr>
<td>Variable</td>
<td>d2</td>
<td>20mm</td>
<td>20mm</td>
<td>5mm</td>
<td>80mm</td>
</tr>
<tr>
<td>Variable</td>
<td>d3</td>
<td>15mm</td>
<td>15mm</td>
<td>5mm</td>
<td>50mm</td>
</tr>
<tr>
<td>Variable</td>
<td>d4</td>
<td>15mm</td>
<td>15mm</td>
<td>5mm</td>
<td>30mm</td>
</tr>
<tr>
<td>Object</td>
<td>Weight</td>
<td>72.25/Kg</td>
<td>67.71/Kg</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Constraint</td>
<td>Max stress</td>
<td>81.9/MPa</td>
<td>118/MPa</td>
<td>/</td>
<td>150/MPa</td>
</tr>
<tr>
<td>Constraint</td>
<td>Natural frequency</td>
<td>73.80/Hz</td>
<td>73.69/Hz</td>
<td>70/Hz</td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>SE</td>
<td>28.80/dB</td>
<td>42.07/db</td>
<td>35/db</td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>Temperature of power1</td>
<td>71.9/°C</td>
<td>65.33/°C</td>
<td>75/°C</td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>Temperature of power2</td>
<td>73.98/°C</td>
<td>70.52/°C</td>
<td>75/°C</td>
<td></td>
</tr>
</tbody>
</table>
The optimization goal is to reduce the weight $W$, 

$$ \min \ W(\beta) $$  

(13)

The external load on the center position of the upper cover board is a downward force of 100N, and the first natural frequency and maximum stress act as structural constraints,

$$ -f + [f_e] \leq 0 $$  

(14)

$$ \sigma_{\text{max}} - [\sigma_e] \leq 0 $$  

(15)

The constraint for the temperature is the maximum temperature inside the cabinet. Through preliminary analysis, the maximum temperature occurs at the power supply, so the temperatures of the two power supplies both act as constraints. Here it is noted that the power dissipation is 160W for two supplies,

$$ T_{\text{max}} \leq T_e \ (i = 1, 2) $$  

(16)

The frequency here is 300MHz and SE acts as a constraint,

$$ -\text{SE}(\beta) \leq -\text{SE}_e $$  

(17)

In the above formulas, $f_e$, $\sigma_e$, $T_e$ and $\text{SE}_e$ are the allowable minimum natural frequency, maximum stress, maximum temperature and minimum SE respectively. According to the features of the electronic equipment and the optimization model, the Hooke-Jeeves method is utilized here, and the optimizing results are denoted in Table 1.

The table 1 shows that the optimized weight decreases by 6.3% mainly because the wall becomes thinner and the wind shield becomes shorter, and the maximum stress increases but is still within the constraint. At the same time, the first-order natural frequency is hardly influenced. The radiating holes change from cube to rectangle making the electromagnetic leakage less, thus increasing the EM shielding effectiveness. Last the highest temperature is lowered. Clearly, only through once optimization, the overall performance is improved obviously.

6 CONCLUSION

Electronic equipment is a multi-disciplinary system, and the electromagnetic shielding of complex electronic equipment is actually a multi-field coupling problem in theory.

Firstly, starting from the angle of the multi-field coupling problem and interdisciplinary intersection and mergence, the multi-field coupling model of the high-density packaging system have been estblished. Then, the theory and method of the simplification of complex electromagnetic model based on perturbation theory and the transfer of grid information are discussed. Lastly, the multi-field coupling optimization model is presented and applied to the design of a high-density cabinet, showing that through the multidisciplinary optimization design, the overall performance of the electronic equipment is improved obviously.

However, it should be noted that multi-field coupling problem is a complex and very difficulty issue. Many problems, such as modeling of circuit boards, thermoelectric strong coupling problem, optimization algorithms and efficiency, and application of surrogate models, need to be researched in depth. Besides, the research and application of
multidisciplinary optimization methods should be strengthened in the future study.

In brief, deep research in the electromechanical coupling theory and method is of great significance to the design of high-performance electronic equipment, and thorough research needs to be done further.

REFERENCES


THERMOELECTRIC SIMULATION OF ELECTRIC MACHINES WITH PERMANENT MAGNETS

A. Bermúdez, D. Gómez, F. Pena, M. Piñeiro and P. Salgado

Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Spain

e-mail: alfredo.bermudez@usc.es, mdlores.gomez@usc.es, fran.pena@usc.es, martina.pineiro@usc.es, mpilar.salgado@usc.es

Key words: transient magnetic, nonlinear partial differential equations, finite element methods, voltage drops, permanent magnets, periodic solutions

Abstract. The objective of this work is to describe some numerical tools developed to perform the thermoelectric simulation of electric machines. From the electromagnetic point of view, we will focus on the computation of nonlinear 2D transient magnetic fields where the data concerning the electric current sources involve potential drops excitations. From the thermal point of view, once the electromagnetic losses are known, we will show an application of a Galerkin lumped parameter method (GLPM) to simulate the thermal behavior of an electric motor. The proposed methods are applied to the simulation of a permanent magnet synchronous electric motor.

1 Introduction

One of the limitations in designing electric machines is that temperature of their different components has to remain below some prescribed thresholds. The temperature of the machine depends on the electromagnetic losses, which are the source term in the energy equation. Thus, to create new high-performance electric machines an accurate numerical simulation of their electromagnetic and thermal behavior is needed.

In order to minimize the electromagnetic losses, the magnetic cores of electrical machines are laminated media consisting of a large number of stacked steel sheets, which are orthogonal to the direction of the currents traversing the coils. This geometry allows us to compute the electromagnetic fields in a plane transversal to the device by assuming that the magnetic flux lies in that plane, and then determining the losses a posteriori (see [4]). This methodology is very interesting because reduces the complexity of dealing with the laminated structure of the machine (see [4]). We will use the axial component of the magnetic vector potential as the main unknown of the mathematical model and we will describe the way of providing different kinds of current sources to the system. In particular, if the sources are given in terms of voltage drops we will develop a numerical method...
to compute periodic solutions by determining a suitable initial current intensity which avoids large simulations to reach the steady state. The problem is numerically solved by using an implicit time discretization scheme combined with a finite element method for space approximation.

For the thermal modeling of an electric motor, we adopted an alternative to the classical finite element (FE) method, since the numerical simulation of machines composed of a large amount of pieces can be computationally demanding. The so-called lumped parameter (LP) models use a simplification of the original problem requiring the design of a network that properly represents the physical behavior of the problem [5, 7]. With such network, an approximate problem is established where the spatially distributed variables are changed by a set of scalar unknowns.

Inspired in the LP methods, we propose in [3] a new family of methods called Galerkin lumped parameter (GLP) methods. These methods are inspired by the LP methods and the techniques used in the reduced basis methods. They consist of using Galerkin approximations of a weak formulation of the original problem in the small finite-dimensional space spanned by a special basis well adapted to the physics of the problem. GLP methods allow us to solve the problem in two steps: in the first step, a basis adapted to a decomposition of the computational domain is calculated; in the second one, the global solution is calculated by solving a small ODE. Another advantage is that the basis is independent of some magnitudes (sources, for instance), allowing us to solve several different cases with the same basis.

2 Mathematical modelling

In this section we will describe the electromagnetic and thermal models.

2.1 Electromagnetic model

We state a 2D transient magnetic problem which arises in the mathematical modelling of laminated magnetic media in the presence of permanent magnets.

Let us assume that the current sources $\mathbf{J}$ have non-null component only in the $z$ space direction and that this component does not depend on $z$, i.e., $\mathbf{J} = J_z(x, y, t)\mathbf{e}_z$. We also assume that the laminated core is invariant along the $z$-direction and that, in the field equations, we neglect the effects of eddy currents in this direction because the steel shells are electrically isolated. In this case, the core can be considered as a homogeneous medium and it is easy to see that the magnetic field $\mathbf{H}$, and then the magnetic induction, $\mathbf{B}$, have only components on the $xy$-plane and both are independent of $z$. Thus, for a given current density $\mathbf{J}$, the 2D transient magnetic problem in the $xy$-plane transversal to the device reads:

$$\text{curl}\mathbf{H} = \mathbf{J},$$

$$\text{div}\mathbf{B} = 0.$$
Notice that, $\mathbf{B} \cdot \mathbf{n}$ and, in absence of surface currents, $\mathbf{H} \times \mathbf{n}$ are continuous through the interface of different media.

In order to apply a standard finite element method, let us consider a bounded domain $\Omega$ composed by several connected conductors, permanent magnets, a ferromagnetic core and the air around. Let us denote by $\Omega_i$, $i = 1, \cdots, N$ the conductors in $\Omega$ representing the cross section of the coils and by $\Omega_i$, $i = N + 1, \cdots, N + M$ the permanent magnets in $\Omega$. We also denote by $\Omega_{N+M+1}$ the complementary domain occupied by the air and the ferromagnetic core, i.e., $\Omega_{N+M+1} = \Omega \setminus \cup_{i=1}^{N+M} \Omega_i$. We will suppose that all of the conductors are stranded conductors, which makes it possible to assume that the current density is uniformly distributed and expressed in terms of the total current across each conductor $\Omega_i$. Actually, for each conductor $\Omega_i$, we will see that the source can be given in terms of either the current or the potential drop per unit length in the $z$-direction. Then, we must solve the following system of equations:

1. $\text{curl } \mathbf{H} = \frac{I_i(t)}{\text{meas}(\Omega_i)} \mathbf{e}_z$ in $\Omega_i$, $i = 1, \cdots, N$, (3)
2. $\text{curl } \mathbf{H} = 0$ in $\Omega_i$, $i = N + 1, \cdots, N + M$, (4)
3. $\text{curl } \mathbf{H} = 0$ in $\Omega_{N+M+1}$, (5)
4. $\text{div } \mathbf{B} = 0$ in $\Omega$. (6)

This model is completed with the constitutive law relating the magnetic field to the flux density. In particular, we will assume a linear behavior for the air, $\mathbf{B} = \mu \mathbf{H}$, while the coils and the laminated media may have a nonlinear behavior, $\mathbf{B} = \mu(|\mathbf{H}|)\mathbf{H}$. On the other hand, permanent magnets will be modelled by the linear constitutive law: $\mathbf{B} = \mu \mathbf{H} + \mathbf{B}'$, where $\mathbf{B}'$ is the so-called remanent flux density which is assumed to have the form $\mathbf{B}' = B'_x(x, y, t)\mathbf{e}_x + B'_y(x, y, t)\mathbf{e}_y$. Notice that $\mathbf{B}'$ may depend on time due to the orientation of the permanent magnets change with an eventual motion of the machine.

Next, we will introduce a magnetic vector potential to solve the two-dimensional model. Since $\mathbf{B}$ is divergence free, there exists a so-called magnetic vector potential $\mathbf{A}$ such that $\mathbf{B} = \text{curl } \mathbf{A}$. Under the assumptions above, we can choose a magnetic vector potential of the form $\mathbf{A} = A_z(x, y, t)\mathbf{e}_z$ (see, for instance, [6]). Thus, in terms of $\mathbf{A}$, the transient magnetic model reads:

1. $\text{curl}(\nu_i(\text{curl } \mathbf{A})) \text{ curl } \mathbf{A} = \mathbf{J}$ in $\Omega_i$, $i = 1, \cdots, N$, (7)
2. $\text{curl}(\nu_i \text{ curl } \mathbf{A}) = \text{curl}(\nu_i \mathbf{B}_r)$ in $\Omega_i$, $i = N + 1, \cdots, N + M$, (8)
3. $\text{curl}(\nu_{N+M+1}(\text{curl } \mathbf{A})) \text{ curl } \mathbf{A} = 0$ in $\Omega_{N+M+1}$, (9)
4. $[\nu_i \text{ curl } \mathbf{A} \times \mathbf{n}] = \nu_i \mathbf{B}' \times \mathbf{n}$, on $\partial \Omega_i$, $i = N + 1, \cdots, N + M$. (10)

where $[.]$ denotes the jump across $\partial \Omega_i$, $\mathbf{n}_i$ is the outward unit normal vector to $\partial \Omega_i$ and $\nu_i$ denotes the magnetic reluctivity of $\Omega_i$. In the air $\nu_{N+M+1} = 1/\mu_0$ (where $\mu_0$ denotes the magnetic permeability of the empty space), while in the ferromagnetic material $\nu_{N+M+1}$
A. Bermúdez, D. Gómez, F. Pena, M. Piñeiro and P. Salgado

is a nonlinear function of \(|B| = |\text{curl}\, A|\). Notice that inside the permanent magnets, the term \(\text{curl}(\nu_i B^r)\) has the same effect as an equivalent current density inside the permanent magnet. In general, both \(\nu_i\) and \(B^r\) are constant in the magnet so the right-hand side of (8) is null, but, since \(\nu_i B^r\) is only non-null in the magnet, its tangential component has a jump discontinuity across the surface of the magnet similar to a surface current density-like of value \(\nu_i B^r \times \mathbf{n}\) (see (10)). This interface condition is implicitly included in the weak formulation of the problem to be used for finite element approximation.

Next, we will describe how to impose different kinds of transient sources in the coils. Let \(\sigma_i\) be the electrical conductivity of conductor \(\Omega_i\), \(i = 1, \cdots, N\). From the assumptions on \(J\) and the Ohm’s law, \(J = \sigma_i \mathbf{E}\), we deduce that, in each conductor \(\Omega_i\), the electric field \(\mathbf{E}\) has to be of the form \(\mathbf{E} = E_z(x, y, t)\mathbf{e}_z\). On the other hand, from Faraday’s law, a scalar potential \(V\) must exist such that

\[
\frac{\partial A}{\partial t} + \mathbf{E} = -\text{grad}\, V. \tag{11}
\]

Taking into account the form of \(\mathbf{A}\) and \(\mathbf{E}\), we deduce from this equality that

\[
\frac{\partial V}{\partial z} = C_i(t) \quad \text{in } \Omega_i, \ i = 1, \cdots, N,
\]

where function \(C_i(t)\) represents the potential drop per unit length in direction \(z\), in conductor \(\Omega_i\). Hence, from the previous equation and (11) one deduces

\[
\sigma_i \frac{\partial A_z}{\partial t} + \sigma_i E_z = -\sigma_i C_i(t) \quad \text{in } \Omega_i, \ i = 1, \cdots, N. \tag{12}
\]

Taking into account the previous discussion, we will assume that, for each conductor \(\Omega_i\), either the potential drop \(C_i(t)\) or the current \(I_i(t)\) is given. In particular, let us suppose there are \(N_C\) conductors of the first type and \(N - N_C\) of the second one.

On the boundary \(\partial \Omega\) of \(\Omega\), we will consider, for simplicity, a homogeneous Dirichlet boundary condition, \(\mathbf{A} = \mathbf{0}\), which means that \(\mathbf{B} \cdot \mathbf{n} = 0\) on \(\partial \Omega\). This assumption is satisfied in the physical example to be considered below. It also holds, in general, if the computational domain is taken large enough. Another classic boundary condition in magnetostatics is \(\mathbf{H} \times \mathbf{n} = 0\). In this case, \(1/\mu \text{curl}\, A \times \mathbf{n} = 0\) and further development would be done without any difficulty.

Thus, from Ohm’s law, by integrating equation (12) on each \(\Omega_i\), the problem to be solved becomes:

**Problem 2.1** Given functions \(C_i(t), \ i = 1, \cdots, N_C, I_i(t), \ i = N_C + 1, \cdots, N\), and initial currents \(I_i^0, \ i = 1, \cdots, N_C\), find a field \(A = A_z(x, y, t)\mathbf{e}_z\) and currents \(I_i(t), \ i = 1, \cdots, N\).
\[ \alpha \] is more involved because the model becomes a system of degenerate parabolic nonlinear hence time appears as a parameter. However, the case with potential drop excitations reduces to solving a nonlinear magnetostatics problem at each time in some interval, and

\[ \int_{\Omega_i} \sigma_i A_z(x, y, t) \, dxdy + I_i(t) = -C_i(t)\alpha_i^{-1}, \quad i = 1, \ldots, N, \]  

(18)

\[ I_i(0) = I_i^0, \quad i = 1, \ldots, N. \]  

(19)

where \( \alpha_i \) denotes the resistance of the \( i \)-th conductor per unit length in the \( z \) direction, that is, \( \alpha_i := \left( \int_{\Omega_i} \sigma_i \, dxdy \right)^{-1} \). We notice that the case where the currents are given reduces to solve a nonlinear magnetostatics problem at each time in some interval, and hence time appears as a parameter. However, the case with potential drop excitations is more involved because the model becomes a system of degenerate parabolic nonlinear partial differential equations.

We notice that, in (13), the currents for \( i = N_C + 1, \ldots, N \) are given, but those for \( i = 1, \ldots, N_C \) are unknown. In order to compute the latter we have added equations (18) and (19) to the system. From the computational point of view, it is better to eliminate the unknowns \( I_i(t), \quad i = 1, \ldots, N_C \) from the system. For this purpose, we first obtain \( I_i(t) \) from (18) and then replace it in (13) for \( i = 1, \ldots, N_C \). Then Problem 2.1 states:

**Problem 2.2** Given functions \( C_i(t), \quad i = 1, \ldots, N_C, \quad I_i(t), \quad i = N_C + 1, \ldots, N, \) and initial currents \( I_i^0, \quad i = 1, \ldots, N_C \), find a field \( \mathbf{A} = A_z(x, y, t) \mathbf{e}_z \) such that

\[ \frac{1}{\text{meas}(\Omega_i)} \frac{d}{dt} \int_{\Omega_i} \sigma_i A_z(x, y, t) \, dxdy \, \mathbf{e}_z + \text{curl}(\nu_i(\text{curl A})) \text{curl A} = \frac{I_i(t)}{\text{meas}(\Omega_i)} \mathbf{e}_z, \quad i = 1, \ldots, N, \]  

(20)

\[ \text{curl}(\nu_i(\text{curl A})) \text{curl A} = \frac{I_i(t)}{\text{meas}(\Omega_i)} \mathbf{e}_z, \quad i = N_C + 1, \ldots, N, \]  

(21)

\[ \text{curl}(\nu_i \text{ curl A}) = \text{curl}(\nu_i \mathbf{B}^r) \quad \text{in } \Omega_i, \quad i = N + 1, \ldots, N + M, \]  

(22)

\[ \text{curl}(\nu_{N+M+1}(\text{curl A})) \text{curl A} = 0 \quad \text{in } \Omega_{N+M+1}, \]  

(23)

\[ [\nu_i \text{ curl A} \times \mathbf{n}] = \nu_i \mathbf{B}^r \times \mathbf{n}, \quad \text{on } \partial \Omega_i, \quad i = N + 1, \ldots, N + M, \]  

(24)

\[ \mathbf{A} = 0, \quad \text{on } \partial \Omega, \]  

(25)

\[ I_i(0) = I_i^0, \quad i = 1, \ldots, N_C. \]  

(26)
In many applications, there exist two indices \( i_1 \) and \( i_2 \), \( 1 \leq i_1, i_2 \leq N_C \), such that \( I_{i_1}(t) = -I_{i_2}(t) = I(t) \). In this case, the number of unknown currents in system (20)-(26) is \( N_C - 1 \) and, accordingly, we cannot prescribe each potential drop \( C_{ij}(t), j = 1, 2 \) arbitrarily, but only the difference of potential drops: \( V(t) := C_{i1}(t) - C_{i2}(t) \). The previous model can be modified in an easy way to deal with this case (see further details in [1]).

The numerical solution of the electromagnetic Problem 2.2 is done by a backward Euler scheme for time discretization combined with standard continuous piecewise linear finite elements for space discretization. At each time step, we have to solve a nonlinear problem for which we propose an iterative algorithm known as Bermúdez-Moreno algorithm (see [2]).

### 2.1.1 Computing periodic solutions in the electromagnetic model

If \( N_C = 0 \), the nonlinear boundary-value problem (13)–(19) has a periodic solution when the given currents \( I_1(t), \ldots, I_N(t) \) are periodic functions of period \( T \). However, the problem of computing periodic solutions is more involved when there are conductors for which we know the potential drops \( C_i(t) \) instead of the currents, i.e., if \( N_C \neq 0 \). In this case we will assume that the given potential drops are periodic with the same period \( T \) and null average, that is,

\[
\int_0^T C_i(t)\,dt = 0, \quad i = 1, \ldots, N_C.
\]

We will also assume that the given currents \( I_i(t), i = N_C + 1, \ldots, N \) are periodic functions with common period \( T \). In this case, we have developed a numerical procedure allowing to determine suitable initial currents which avoid large simulations to reach the steady state. We will summarize here the main ideas and refer the reader to [1] for a complete development.

For \( t \in [0, T] \), let us denote by \( F_t = (F_{t,1}, \ldots, F_{t,N_C}) \) the mapping from \( \mathbb{R}^{N_C} \) into itself such that, to the vector of currents \( \vec{I} = (I_1, \ldots, I_{N_C}) \in \mathbb{R}^{N_C} \) associates the numbers

\[
F_{t,i}(\vec{I}) = \alpha_i \int_{\Omega_i} \sigma_i A_z(x, y, t) \,dx\,dy, \quad i = 1, \ldots, N_C.
\]

We notice that computing \( F_t(\vec{I}) \) requires to solve a nonlinear magnetostatics problem at each time \( t \), in order to determine field \( A_z(x, y, t) \). By using this mapping, equations (18) can be rewritten as

\[
\sum_{j=1}^{N_C} (D F_t(\vec{I}))_{ij} \frac{dI_j(t)}{dt} + \alpha_i I_i(t) = -C_i(t), \quad i = 1, \ldots, N_C,
\]  

(27)
where \((DF_t(\vec{I}))_{ij}\) denotes the \(ij\)-th element of the Jacobian matrix \(DF_t(\vec{I})\) of \(F_t\) at point \(\vec{I}\). Let us assume the following hypothesis:

\[
\frac{\alpha_i T}{\min_{t,\vec{I}}|DF_t(\vec{I})|_{ii}} \ll 1, \ i = 1, \ldots, N_C.
\]

In this case, the term involving \(\alpha_i\) can be neglected in (27). Thus, by using algebraic operations (see [1]) we deduce that to compute an initial condition leading to a periodic solution from the initial time, we can solve the system of equations:

\[
F_0(\vec{I}_0) = \frac{1}{T} \left( \int_0^T F_t(0) dt + \int_0^T (T - s) \vec{C}(s) ds \right).
\]

(28)

Notice that, in order to solve (28), it is first necessary to compute the term \(F_t(0)\) by solving a magnetostatics problem for each value of \(t \in [0,T]\). Once this term has been computed, the nonlinear system (28) has the important feature that only involves the magnetostatics problem for time \(t = 0\).

2.2 Thermal model

Given the volumetric heating, calculated with the electromagnetetic model explained in the previous section, we can tackle the calculation of the temperature in the electric motor. We suppose that domain \(\Omega\) is divide in several subdomains, connected through surfaces called \emph{ports}. Following the GLP method explained in [3], we pose the heat equation for the temperature \(\theta\):

\[
\rho c \frac{\partial \theta}{\partial t} - \text{div}(k \text{grad} \theta) = f \quad \text{in} \quad \Omega \times [0,T],
\]

(29)

\[
\theta(x, t) = \theta^P_l(x, t) \quad \text{on} \quad \Gamma^P_l, \ l = 1, \ldots, n_P,
\]

(30)

\[
-k \frac{\partial \theta}{\partial n}(x, t) = \alpha_l(\theta(x, t) - \theta^C_l(x, t)) \quad \text{on} \quad \Gamma^C_l, \ l = 1, \ldots, n_C,
\]

(31)

\[
k \frac{\partial \theta}{\partial n} = 0 \quad \text{on} \quad \Gamma^A,
\]

(32)

\[
\theta(x, 0) = \theta_0(x) \quad \text{in} \quad \Omega,
\]

(33)

where \(\rho\) is the density, \(c\) is the specific heat, \(k\) is the thermal conductivity, \(f\) is the volumetric heating, \(\alpha_l\) are the convective heat transfer coefficients, \(\theta^C_l\) are the convective temperatures and \(\theta_0\) is the initial temperature.

Here, the boundary \(\Gamma\) of the domain is divided into three parts: \(\Gamma^P = \bigcup_{l=1}^n \Gamma^P_l\) are the \emph{ports}, i.e., the surfaces of the domain connecting two or more sub-domains. We assume that each \(\Gamma^P_l\) is a connected component of \(\Gamma^P\). \(\Gamma^C = \bigcup_{l=1}^n \Gamma^C_l\) are the \emph{convective boundaries}, where convective heat transfer conditions are applied. Finally, \(\Gamma^A\), are the thermally isolated (or adiabatic) surfaces.
We consider a domain decomposition in order to construct an adapted basis in each sub-domain: the domain $\Omega$ splits into several sub-domains $\Omega_i$, $i = 1, \ldots, N$, connected among them through boundaries called ports. Thus, in the boundary of each sub-domain $\Omega_i$, called $\Gamma_i$, we distinguish three parts: the ports, $\Gamma_i^P = \bigcup_{j=1}^{n_i^P} \Gamma_{ij}^P$, consisting of all the boundaries between sub-domains; the convective boundaries, $\Gamma_i^C = \bigcup_{j=1}^{n_i^C} \Gamma_{ij}^C$ and the the isolated boundary, $\Gamma_i^A$.

The basis for the $i$-th sub-domain consists of $n_i^P + n_i^C$ elements, to be called $\varphi_{ij}^P$: $j = 1, \ldots, n_i^P$, and $\varphi_{ij}^C$: $j = 1, \ldots, n_i^C$ which are defined as the unique solutions to the following stationary boundary-value problems:

For $i = 1, \ldots, N$ and $j = 1, \ldots, n_i^P$ find $\varphi_{ij}^P \in H^1(\Omega_i)$ satisfying,

\begin{align}
- \text{div}(k \text{grad } \varphi_{ij}^P) &= 0 \quad \text{in } \Omega_i, \\
\varphi_{ij}^P(x) &= \delta_{jl} \quad \text{on } \Gamma_{il}^P, \ l = 1, \ldots, n_i^P, \\
k \frac{\partial \varphi_{ij}^P}{\partial n} + \alpha_l \varphi_{ij}^P &= 0 \quad \text{on } \Gamma_{il}^C, \ l = 1, \ldots, n_i^C, \\
k \frac{\partial \varphi_{ij}^P}{\partial n} &= 0 \quad \text{on } \Gamma_i^A.
\end{align}

For $i = 1, \ldots, N$ and $j = 1, \ldots, n_i^C$ find $\varphi_{ij}^C \in H^1(\Omega_i)$ satisfying,

\begin{align}
- \text{div}(k \text{grad } \varphi_{ij}^C) &= 0 \quad \text{in } \Omega_i, \\
\varphi_{ij}^C(x) &= 0 \quad \text{on } \Gamma_{il}^P, \ l = 1, \ldots, n_i^P, \\
k \frac{\partial \varphi_{ij}^C}{\partial n} + \alpha_l (\varphi_{ij}^C - \delta_{jl}) &= 0 \quad \text{on } \Gamma_{il}^C, \ l = 1, \ldots, n_i^C, \\
k \frac{\partial \varphi_{ij}^C}{\partial n} &= 0 \quad \text{on } \Gamma_i^A.
\end{align}

We notice that each problem is established in one single sub-domain $\Omega_i$, it is independent on time, and the stiffness matrix is the same for all the problems so it can be assembled and factorized only once. Now, the global basis can be constructed using the solutions of them. It consists of two types of functions:

1. the elements $w_l^P$ that coincide with $\varphi_{ij}^P$ in $\Omega_i$ when the $l$-th global port is the $j$-the local port of $\Omega_i$, and that are zero otherwise;

2. the elements $w_l^C$ that coincide with $\varphi_{ij}^C$ in $\Omega_i$ when the $l$-th global convective boundary is the $j$-th local convective boundary of $\Omega_i$, and that are zero otherwise.

Let us call $\mathcal{V} \subset H^1(\Omega)$ the linear space spanned by the above set of $n^P + n^C$ functions. The lumped parameter model is defined as the Galerkin approximation of the weak formulation of problem (29)–(33) corresponding to this basis:
For \( t \in [0, T] \), find \( \tilde{\theta}(\cdot, t) \in \mathcal{V} \) satisfying

\[
\int_{\Omega} \rho_c \frac{\partial \tilde{\theta}}{\partial t} \tilde{\psi} \, dx + \int_{\Omega} k \text{grad} \tilde{\theta} \cdot \text{grad} \tilde{\psi} \, dx + \sum_{l=1}^{n_C} \int_{\Gamma_l^C} \alpha_l \tilde{\theta} \tilde{\psi} \, d\Gamma \\
= \int_{\Omega} f \tilde{\psi} \, dx + \sum_{l=1}^{n_C} \int_{\Gamma_l^C} \alpha_l \tilde{\theta}_l^C \tilde{\psi} \, d\Gamma \quad \forall \tilde{\psi} \in \mathcal{V}
\]  

(42)

where \( \tilde{\theta}_0 \) denotes a projection of the initial condition \( \theta_0 \) on the space \( \mathcal{V} \).

Now, we seek a solution that, at each time \( t \), belongs to the space spanned by the reduced basis:

\[
\tilde{\theta}(x, t) = \sum_{l=1}^{n_P} \theta_l^P(t) w_l^P(x) + \sum_{l=1}^{n_C} \theta_l^C(t) w_l^C(x),
\]

Here, coefficients \( \theta_l^P \) and \( \theta_l^C \) denote the temperature at the ports and the convective boundaries, respectively. In order to determine those coefficients, we replace \( \tilde{\theta}(x, t) \) in equations (42)–(43), obtaining an ordinary differential system of dimension \( n_P + n_C \). We note that the dimension of this system is much smaller than the classical finite element method.

3 Numerical results

In the first part of this section we will show some numerical results obtained with a Fortran code implementing the numerical methods described for the electromagnetic problem.

Figure 1-left shows the cross-section of a permanent magnet synchronous electric motor having 16 magnet poles and three phase windings, each of them composed of 16 coils with 31 turns. This motor has buried permanent magnets, entirely enclosed in the solid rotor structure. We consider that both rotor and stator have nonlinear magnetic cores, laminated in the direction of the current, which allows us to solve a 2D problem in the cross section of the device. Furthermore, we notice that we can solve the model in an eighth of the geometry by imposing an evenly periodic boundary condition due to the configuration of coils and magnets.

The motor is driven by a uniformly distributed sinusoidal three phase current. The permanent magnets are assumed to have a remanent flux density of 1.26 T oriented parallel to the radial direction at each magnet center (i.e., parallel to its edges). On the other hand, rotor and stator are composed by a nonlinear material described by the curve \( B(H) \) depicted in Figure 1-right, while all the other materials are considered to be linear. We notice that the high relative permeability of ferromagnetic core with respect to the air surrounding the device would ensure that most of the flux will remain inside the stator,
and therefore flux lines are constrained to follow the stator boundary. This leads us to impose a Dirichlet homogeneous condition on the exterior boundary of the device.

Figure 2-left shows the modulus of the computed flux density distribution and Figure 2-right shows the corresponding vector field. From the values of the magnetic flux density at each point, we have computed the losses in each material by using a posteriori estimation techniques (see, for instance, [4]). These losses have been introduced as sources in the thermal model by assuming that are invariant in the z-direction.

For the heat equation, the GLP method has been implemented in a computer program by using Matlab. The numerical algorithm consists of two parts, that can be executed independently:

1. in the first one, functions $\varphi_{ij}^P$ and $\varphi_{ij}^C$ are calculated as solution of systems (34)–(37) and (38)–(41);

2. in the second one, function $\tilde{\theta}$ is calculated as solution of an ordinary differential system of equations equivalent to (42)–(43).

For the second case, we consider an electric motor designed by the University of Mondragon and the Orona company. In this example we consider a motor at room temperature that starts at $t = 0$ and works for 2 hours. In order to check the program, electric losses were measured by the University of Mondragon and they are given as data to the program. The program also admits electric losses calculated with an external code. The motor is decomposed into 34 pieces, each of them represents a sub-domain in the GLP method. When the temperature is constant at ports, the total number of basis functions in GLP method is only 70 and the relative error in $L^2(0,T)\cap \Omega$ respecto to the FE solution is close to 2% (see [3]).
In order to reduce the error derived from the fact that temperature is taken to be constant at the ports, a “nodal” version GLP method was programmed. In this case, a port is each single node between two pieces. Now, the meshes of the different sub-domains must be conforming on the common interfaces and the number of basis functions is related to the number of nodes on the ports, not to the number of ports or convective surfaces. In Figure 3 a cross-section of the motor can be observed. Temperature was calculated with the nodal version.

In this example, the nodal version presents a relative error with respect to the standard
FE method ten times smaller than the GLP method. Still, step 2 of the algorithm is up to 2.5 times faster than the standard FE method.

REFERENCES


AN ENHANCED IMMERSED STRUCTURAL POTENTIAL METHOD (ISPM) FOR THE SIMULATION OF FLUID-STRUCTURE INTERACTION PROBLEMS

ANTONIO J. GIL*, AURELIO ARRANZ CARREÑO*, JAVIER BONET* AND CLARE WOOD*

*Civil and Computational Engineering Centre (C^2EC)
College of Engineering, Swansea University
Singleton Park, SA2 8PP, Swansea, United Kingdom
e-mail: a.j.gil@swansea.ac.uk, web page:
http://www.swan.ac.uk/staff/academic/engineering/gilantonio/

Key words: Fluid-Structure Interaction, Immersed methods, ISPM

Abstract. Immersed methods are widely used nowadays for the computational simulation of Fluid-Structure Interaction problems. In this paper, the Immersed Structural Potential Method (ISPM) is coupled with a Runge-Kutta-Chebyshev Projection method in order to increase the overall computational efficiency of the methodology. Application of the framework to large three-dimensional problems is carried out. A series of numerical examples will be presented in order to demonstrate the robustness and flexibility of the proposed methodology.

1 INTRODUCTION

Fluid-Structure Interaction (FSI) has been a major topic of research interest for the past few decades. From the spatial discretisation point of view, two general approaches can be established, namely boundary fitted methods and immersed methods. Boundary fitted methods rely on the coupling of the velocity field at the interface between the two phases describing the problem (i.e. fluid and solid). This coupling can be carried out either in a monolithic manner [1] or in a partitioned staggered manner [2, 3], always requiring the consistent update of the interface and the use of an Arbitrary Lagrangian Eulerian approach to describe the fluid phase.

On the contrary, immersed methods are designed to embed the solid phase within the fluid phase, enabling the calculation of the FSI effect on a stationary fluid grid which can be analysed in a purely Eulerian fashion [4, 5, 6, 7]. From a methodological point of view, immersed methods are part of the so-called Fictitious Domain (FD) philosophy, introduced by Glowinski in [8] for the resolution of boundary value problems in complex geometrical settings.
Three different FD approaches can be identified. First, non-body force based schemes, where the presence of an immersed solid is established by enforcing strongly that the velocity of the surrounding fluid matches that of the immersed structure at the interface (i.e. Dirichlet Boundary Conditions (BCs) on the fluid); in turn, the movement of the immersed solid is obtained after the equation of motion of the solid is solved subjected to the force field imposed by the surrounding fluid domain (i.e. Neumann BCs on the solid). Second, body force based Distributed Lagrange Multiplier (DLM) methods, where the no-slip velocity constraint at the interface between the physics is imposed as an equation for the Lagrange multiplier defined on the solid boundary. Third, body force based non-DLM methods, where the no-slip velocity constraint is imposed strongly on the solid (i.e. Dirichlet BCs on the solid) and a FSI body force is then evaluated and applied to the surrounding fluid (i.e. Neumann BCs on the fluid). Methods such as the Immersed Boundary Method (IBM) [4], the Extended Immersed Boundary Method (EIBM) [9] and the Immersed Finite Element Method [10] belong to this group.

An alternative immersed methodology, named Immersed Structural Potential Method (ISPM), was introduced in [6]. In the ISPM, the kinematics of the solid is completely defined by the underlying fluid domain through suitable interpolating functions and the solid is modelled by means of a deviatoric energy potential which is accurately evaluated at a cloud of integration points. The methodology has been improved in [11] by introducing smoothed spline-based kernels for the definition of the interpolating functions and high order quadrature rules to ensure the accurate description of the solid phase. In this paper, a further improvement is introduced in the form of a Runge-Kutta Chebyshev Projection (RKCP-ISPM) time integration scheme, leading to a very efficient fully parallelised framework that allows for the simulation of large-scale three-dimensional problems.

2 FLUID GOVERNING EQUATIONS

Let us consider the motion of a continuum defined by means of a mapping \( \phi \) established between a reference or material configuration \( X \in \Omega_0 \subset \mathbb{R}^d \) and a spatial or current configuration \( x \in \Omega \subset \mathbb{R}^d \) at time \( t \), namely \( x(t) = \phi(X,t) \), where \( d = \{2, 3\} \) represents the number of space dimensions. The deformation gradient tensor \( F \) is defined as the material gradient of the spatial position as

\[
F = \nabla_0 x = \frac{\partial x}{\partial X}, \quad J = \det F
\]

where \( J \) is the Jacobian of the transformation. In addition, the velocity field \( u = [u_1 \ldots u_d]^T \) of the continuum is computed as \( u(X,t) = \frac{\partial x}{\partial t} \). In the case of an incompressible Newtonian viscous fluid, the conservation of linear momentum for an arbitrary spatial volume \( \Omega \) of boundary \( \partial \Omega \) with outward normal \( n \) is expressed in integral form as

\[
\int_\Omega \frac{\partial}{\partial t} (\rho u) \, dv + \int_{\partial \Omega} (\rho u \otimes u + pI - \mu \nabla u) \cdot n \, da - \int_\Omega g \, dv = 0.
\]
where $I$ is the identity tensor, $g$ denotes an external volume force field per unit of spatial volume, $p$ is the scalar pressure field, $\mu$ is the viscosity and $\rho$ the density of the fluid.

Within the framework of low order Finite Volume schemes and particularising for the case of a Cartesian staggered mesh, let $\Omega_{u^A_i}, i \in \{1 \ldots d\}$ be the control volume associated with the Cartesian component of the trial and test velocity components $u^A_i$ and $\delta u^A_i$, respectively, with an arrangement similar to that of a Marker And Cell (MAC) grid. Here, $A_i$ denotes the fluid control volume face perpendicular to the $i$-th Cartesian axis and $u^A_i$ and $\delta u^A_i$ the corresponding normal trial and test face velocity fields. The Finite Volume semi-discrete weak form of the linear momentum equation reads

$$\int_{\Omega_{u^A_i}} \delta u^A_i \frac{\partial}{\partial t} (e_i \cdot \rho u) \, dv + \int_{\delta \Omega_{u^A_i}} \delta u^A_i e_i \cdot F_n \, da - \int_{\Omega_{u^A_i}} \delta u^A_i e_i \cdot g \, dv = 0, \quad i \in \{1 \ldots d\}$$

(3)

where $\{e_i\}, i \in \{1 \ldots d\}$ is the Cartesian unit basis and $F$ defines the numerical interface flux, namely,

$$F = \rho u \otimes u + p I - \mu \nabla u.$$  

(4)

Equation (3) is solved along with the divergence free velocity constraint as part of a fractional step algorithm [12].

3 IMMERSED SOLID GOVERNING EQUATIONS

Let us consider an incompressible deformable solid with density $\rho_s$ fully immersed within the surrounding incompressible viscous fluid. The solid can be modelled as a deviatoric Helmholtz’s free energy density functional $\hat{\Psi}_s$ whose spatial gradient defines a FSI force field which is regarded as an external source term by the background viscous fluid [12].

For spatial semi-discretisation purposes, the solid domain is modelled in a Lagrangian manner as a collection of integration points $a_p$ immersed within the fluid, moving from an initial position $X^{a_p}$ to the spatial position $x^{a_p}$ at time instant $t$, through the deformation gradient tensor $F$ defined by the motion of the surrounding continuum (i.e. non-slip condition). With the purpose of distinguishing the surrounding fluid phase from the immersed solid phase, a superindex $(.)^s$ will be employed wherever necessary when referring to the latter.

The kinematics of the immersed solid is defined by the underlying fluid domain. Thus, the velocity of the deformable immersed solid can be obtained after suitable definition of an interpolation operator which enables to transfer information from the background Eulerian fluid to the Lagrangian solid. Specifically, the velocity field $u^s$ at any integration point $a_p$ currently at $x^{a_p}$ can be evaluated as follows,

$$I(u)(x^{a_p}) := u^{a_p} = \left[ u_1^{a_p} \ldots u_d^{a_p} \right]^T, \quad u_1^{a_p} = \sum_{A_i} u_i^{A_i} \varphi^{A_i}(x^{a_p})$$

(5)
where
\[ \varphi^{A_i}(x^s) = \varphi(x^s - x^{A_i}) \] (6)
are interpolating functions centred at fluid control volume faces \( A_i \), defined by the spatial position \( x^{A_i} \), mid-point of the respective fluid control volume face. Similarly, the virtual velocity field vector \( \delta u^s \) is evaluated at a structure integration point \( a_p \) with a consistent interpolating methodology (7), to ensure conservation of the overall scheme
\[ \delta u^{a_p} = \left[ \delta u_1^{a_p} \ldots \delta u_d^{a_p} \right]^T, \quad \delta u_i^{a_p} = \sum_{A_i} \delta u^{A_i} \varphi^{A_i}(x^{a_p}). \] (7)

The internal virtual work formulated in the case of the immersed solid domain is defined as the directional derivative of the Helmholtz’s free energy functional with respect to a virtual velocity field vector as follows,
\[ \delta W_{int}^s(\phi, \delta u^s) = \int_{\Omega_0^s} \tau^s : \nabla \delta u^s \, dV \simeq \sum_{a_p} W_i^{a_p} \tau_i^{s,a_p} : \nabla \delta u^{a_p} \] (8)
where \( \tau^s \) is the deviatoric Kirchhoff stress tensor. The evaluation of the above formula (8) requires the computation of the spatial gradient of the virtual velocity at integration point \( a_p \), which can be easily obtained by making use of the spatial gradient of the interpolating functions defined above,
\[ \nabla \delta u^{a_p} = \left[ \sum_{A_1} \nabla \varphi^{A_1}(x^{a_p}) \delta u^{A_1} \ldots \sum_{A_d} \nabla \varphi^{A_d}(x^{a_p}) \delta u^{A_d} \right]^T. \] (9)

After re-writing the Kirchhoff stress tensor in the form \( \tau^s = [\tau_i^s \ldots \tau_d^s] \), equation (8) can be reformulated in a discrete manner as
\[ \delta W_{int}^s(\phi, \delta u^s) = \sum_{A_1} \delta u_{A_1} f_{A_1} + \ldots + \sum_{A_d} \delta u_{A_d} f_{A_d} \] (10)
where
\[ f_{A_i} = \int_{\Omega_0^s} \tau_i^s : \nabla \varphi^{A_i}(x^s) \, dV \simeq \sum_{a_p} W_i^{a_p} \tau_i^{s,a_p} : \nabla \varphi^{A_i}(x^{a_p}), \quad i \in \{1 \ldots d\} \] (11)

In order to guarantee conservation of the scheme it transpires that
\[ g_{A_i} := \int_{\Omega_{u,A_i}} e_i \cdot g \, dv = \frac{f_{A_i}}{|\Omega_{u,A_i}|} \] (12)
where \( g_{A_i} \) represents the FSI force per unit volume \( |\Omega_{u,A_i}| \) which must be applied at the fluid control volume face \( A_i \). A key ingredient for the evaluation of the stress tensor \( \tau^s \) is
the deformation gradient tensor $F$ at any location within the immersed continuum. First, the spatial velocity gradient tensor $l$ is evaluated at every integration point $a_p$, for which the following interpolation operator is used

$$\nabla I(u)(x^{a_p}) := l^{a_p} = \nabla u^{a_p} = \left[ \sum_{A_1} \nabla \phi^{A_1}(x^{a_p}) u^{A_1} \ldots \sum_{A_d} \nabla \phi^{A_d}(x^{a_p}) u^{A_d} \right]^T \tag{13}$$

where the spatial gradient of the interpolating functions $\{\nabla \phi^{A_x}, \nabla \phi^{A_y}\}$ can be explicitly computed. Second, a time integration scheme is proposed for the tensor system of kinematic differential equations

$$\dot{F} = lF = (d + w)F \tag{14}$$

where $d$ and $w$ are the strain rate tensor and the vorticity tensor, respectively. An iterative explicit time integration scheme can be employed with the purpose of obtaining $F$ at time instant $n + 1$ and iteration $k + 1$ as

$$F_n^{n+1}_k = e^{\Delta t d_n^{k+1}} F^n.$$ \tag{15}

A suitable $d$-$w$ structure preserving time integration scheme is presented in [12] in order to ensure the volume conservation of the immersed solid phase. In addition, the introduction of smoothed spline-based kernels for the evaluation of the interpolating functions $\phi^{A_i}$ leads to improved shear stresses in the immersed solid phase [11]. From the close-ups in Figure 1 it can be noted how the combination of the ISPM with the new kernel ‘k15’ removes most of the oscillations present in $F$, which are present in both the EIBM and ISPM while using existing Peskin’s kernels [4]. Furthermore, the use of high order quadrature rules used to discretise the immersed solid phase ensures the accurate description of the solid phase, preventing the so-called numerical leakage [11].

4 \hspace{0.1cm} EXPLICIT RUNGE-KUTTA-CHEBYSHEV-PROJECTION METHOD

Runge-Kutta-Chebyshev (RKC) methods are a subclass of explicit Runge-Kutta (RK) methods, where instead of trying to maximise the order of the method for a given number of stages, as it is customary for typical RK methods, only the first two stages are used to ensure second order accuracy, and the remaining stages are used to maximise the region of absolute stability of the method, while keeping an explicit scheme with low storage requirements. In the case of RKC methods, this is achieved by careful choice of the constants so that the stability polynomials can be expressed in terms of Chebyshev polynomials. The well known three-term recurrence equation that these polynomials satisfy translates also to three-term recurrence relations for stage updates.

This choice of time integrator has been favoured for the ISPM framework as the second order accuracy in time is in agreement with the errors incurred by the spatial semi-discretisation, but also because it minimises the number of intermediate vectors that
have to be stored while providing extra stability, a highly desirable feature for fluid-
structure interaction problems given their multiple sources of non-linearity. The original
RKC method was introduced by Van der Houwen [13] and extended to the use with
projection methods (RKCP) by Zheng and Petzold [14] for incompressible flow with no
solid interaction. In this section, a combination of the RKCP with the ISPM is presented
(RKCP-ISPM).

In order to ease the exposition, it will be convenient to recast the problem in the
notation that is classical in the context of ODE integrators. To this purpose, we rewrite
the governing equations in the form
\[
\dot{Y} = f(t, Y); \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} u \\ F \\ x \end{bmatrix} \tag{16}
\]
where
\[
f(t, Y) = \begin{bmatrix} -(Y_1 \cdot \nabla)Y_1 + \mu \Delta Y_1 - \frac{1}{\rho} \nabla p + \frac{1}{\rho} g(Y, t) \\ \nabla I(Y_1)Y_2 \\ I(Y_1)(Y_3, t) \end{bmatrix} \tag{17}
\]
Note how in the above equations, the first term corresponds to the incompressible Navier-
Stokes equations, whereas the second accounts for the integration of the deformation
gradient using the interpolated velocity gradient and the third for the integration of the
position of the quadrature nodes using the interpolated velocity field.

With the above notation, we can pose the following s-stage explicit RKCP integration
scheme coupled with the ISPM. In the formulas that follow, \( Y_{j,k} \) denotes the value of the
k-th variable of the system for stage j. For the initial stage \( (j = 0) \) we have
\[
\begin{cases}
Y_{0,1} = u^n, \\
Y_{0,2} = F^n, \\
Y_{0,3} = x^{n,n}
\end{cases} \tag{18}
\]
The first stage advances in time as follows,
\[
\begin{align*}
Y_{1,2} &= e^{\mu \Delta t}Q(\nabla I(Y_{0,1})(Y_{0,3})Y_{0,2}, \\
Y_{1,3} &= x^{n,n} + \mu_1 \Delta t I(Y_{0,1})(Y_{0,3}), \\
\tilde{F}_{0,1} &= -(u^n \cdot \nabla h)u^n + \mu \Delta_h u^n - \frac{1}{\rho} \nabla_h p_n + f^0_h(u^n, x^{n,n}, t, F^n), \\
Y_{1,1} &= u^n + \mu_1 \Delta t \tilde{F}_{0,1}, \\
Y_{1,1} &= P(Y_{1,1})
\end{align*} \tag{19}
\]
where in the above \( \nabla_h \) and \( \Delta_h \) denote the corresponding discretised versions of the gradient
and Laplacian operators, and the interaction force term is defined as
\[
f^0_{i,h} = \int_{\Omega_0} \tau^i_{h}(Y_{1,2}) \cdot \nabla \varphi^i_h(Y_{0,3}) \, dV + (\rho_s - \rho) \frac{I(Y_{0,1})(Y_{0,3}) - I(u^{n-1})(x^{n-1})}{\Delta t} \tag{20}
\]
Note that in the first equation in (19) the deformation gradient is integrated to obtain a deformation gradient for the stage to evaluate the forces arising from (20). The second equation in (19) integrates the position of the quadrature nodes, and the remaining equations apply the fractional step method to solve the incompressible Navier-Stokes equations.

For stage \( j \) we have analogously

\[
\begin{cases}
Y_{j,2} = e^{c_j \Delta t} Q(\nabla I(Y_{j-1,1})(Y_{j-1,3})) Y_{0,2}, \\
Y_{1,3} = (1 - \mu_j - \nu_j) A + \mu_j Y_{j-1,3} + \nu_j Y_{j-2,3} + \tilde{\mu}_j \Delta t I(Y_{j-1,1})(Y_{j-1,3}) + \tilde{\gamma}_j \Delta t I(Y_{0,1})(Y_{0,3}), \\
\tilde{F}_{j-1,1} = -(Y_{j-1,1} \cdot \nabla h) Y_{j-1,1} + \frac{\rho}{\gamma} \tilde{\chi} Y_{j-1,1} - \frac{1}{\rho} \nabla h p_n + f_{h}^{j-1}(Y_{j-1,1}, Y_{j-1,3}, t + c_{j-1} \Delta t, Y_{j,2}), \\
Y_{j,1} = (1 - \mu_j - \nu_j) Y_{0,1} + \mu_j Y_{j-1,1} + \nu_j Y_{j-2,1} + \tilde{\mu}_j \Delta t \tilde{F}_{j-1,1} + \tilde{\gamma}_j \Delta t \tilde{F}_{0,1}, \\
Y_{j,1} = \mathcal{P}(Y_{j,1}^\ast)
\end{cases}
\]

where

\[
f_{l,h}^{j-1} = \int_{\Omega} \tau_i^s(Y_{j,2}) \cdot \nabla \varphi^h(Y_{j-1,3}) \, dV + (\rho_s - \rho) \frac{I(Y_{j-1,1})(Y_{j-1,3}) - I(Y_{0,1})(Y_{0,3})}{c_{j-1} \Delta t}
\]

In order to complete the fractional step method, the operator \( \mathcal{P} \) is defined as

\[
\mathcal{P}(\mathcal{u}^\ast) = \mathcal{u}^\ast - \nabla \phi
\]

where \( \phi \) is the solution of following Poisson’s problem

\[
\begin{align*}
\Delta \phi &= \nabla \cdot \mathcal{u}^\ast, \quad \text{in } \Omega \\
\nabla \phi \cdot n &= 0, \quad \text{on } \partial \Omega
\end{align*}
\]

And in order to ensure that the integration scheme used for component \( Y_2 \) preserves the incompressibility constraint, the operator \( Q \) is defined as the projection onto traceless tensors as follows

\[
Q(l) = l - \frac{1}{d} (I \otimes I) : l
\]

for \( d = 3 \) and by

\[
Q(l) = \begin{bmatrix}
\frac{1}{2}(l_{11} - l_{22}) & l_{12} & 0 \\
l_{21} & \frac{1}{2}(l_{22} - l_{11}) & 0 \\
0 & 0 & 0
\end{bmatrix}
\]

for \( d = 2 \), being \( d \) the number of space dimensions. The scalar constants in the stage recurrence relations are defined by

\[
\begin{align*}
c_0 &= 0, \quad c_1 = c_2, \quad c_2 = \tilde{\mu}_1, \quad c_j = \frac{T_j^\omega(T_j^\omega)}{T_j^\omega}, \quad j = 3, \ldots, s \\
b_0 &= \frac{1}{2} \omega_0, \quad b_1 = \frac{1}{\omega_0}, \quad b_j = \frac{T_j^\omega}{T_j^\omega} \frac{1}{T_j^\omega}, \quad j = 2, \ldots, s \\
\mu_j &= 2b_0 b_j / b_j - 1, \quad \tilde{\mu}_j = \mu_j \omega_1 / \omega_0, \quad \tilde{\gamma}_j = (T_j - 1)(\omega_0) b_j - 1 \mu_j \omega_1 / \omega_0, \quad j = 1, \ldots, s \\
\nu_j &= -b_j / b_j - 2, \quad j = 2, \ldots, s, \quad \omega_0 = 1 + \frac{\varepsilon}{4\pi}, \quad \omega_1 = \frac{T_2^\omega}{T_2^\omega}
\end{align*}
\]
so that the corresponding stability polynomial is

\[ P_j(z) = a_j + b_j T_j(\omega_0 + \omega_1 z), \quad \text{where} \quad a_j = -\frac{\tilde{\gamma}_j}{\tilde{\mu}_j} \]

(28)

In all the above \( T_j \) denotes the \( j \)-th Chebyshev polynomial of the first kind and the damping parameter \( \varepsilon \) is chosen as 2/13. The choice for \( b_0 \) and \( b_1 \) in the formulas above is carried out as recommended in [15] for advection-diffusion-reaction problems. It is worth mentioning that this approach is not a mere application of the RKCP method to fluid-structure interaction using ISPM. The component \( Y_2 \) of the system vector is not integrated using the RKC scheme, but updated using the explicit structure-preserving scheme introduced in [12]. The use of this integrator ensures the satisfaction of the incompressibility constraint in the solid exactly.

5 NUMERICAL RESULTS

In this section we present a large-scale numerical example solved using the above presented methodology. We consider a prismatic channel \( \Omega = [0,8] \times [0,2] \times [0,2] \) filled with an incompressible fluid of density \( \rho = 1 \) and viscosity \( \mu = 5 \cdot 10^{-3} \) and five flexible membranes \( \Omega_s^i, \ i = 1, \ldots, 5 \) of density \( \rho = 1 \) and dimensions \( 0.2 \times 1.5 \times 1 \) uniformly spaced and centred along the channel (separated by a distance of 1). The membranes obey a Neo-hookean material model with shear modulus \( G = 3.846 \cdot 10^4 \). All boundaries are set to a non-slip boundary condition, except for the short ends, where a Poiseuille pulsatile flow of amplitude \( A = 5(\sin(2\pi t) + 1.1) \) is applied (left end) and outflow boundary conditions (right end). Note that due to the boundary condition, the flow always goes from left to right, with different positive velocity. The fluid is discretised using a staggered Finite Volume scheme with \( 256 \times 64 \times 64 \) control volumes and adaptive time-step using the RKCP presented in the previous section (\( s = 10 \)). The solid membranes are integrated using a composite Gaussian quadrature rule of degree 6 in each direction with \( 2 \times 11 \times 8 \) subintervals (giving a total of 38016 integration points per membrane).

As it can be observed in Figures 2 and 3, a complex flow pattern develops around the membranes and causes them to deform. The always right-flowing fluid exerts a force on the first membrane, which stays deformed during the whole cycle. The rest of the membranes oscillate depending on the formation of vortices downstream as the flow decelerates (see amplitude function).

The above example has been solved using a hybrid Matlab/Fortran implementation where the computation of forces is parallelised using OpenMP. The total number of fluid variables is close to 3.2 million, and the solid discretisation is short of 0.2 million integration points. The runtime on a i7-2760QM processor is around 20 hours per amplitude cycle.
Figure 1: Deformation gradient tensor components $F_{12}$ and $F_{22}$ for the leading edge of an immersed membrane (as seen with respect to the material coordinates) for a model problem [11]. Thick lines represent solutions obtained with the ISPM and an improved kernel ‘k15’ for different mesh refinements (runs 2, 3 and 4 in increasing order of refinement). Thin lines represent corresponding solutions obtained with the EIBM and Peskin’s kernel (‘k2’). The solution obtained with the ISPM and Peskin’s kernel and the finest mesh has also been added to allow for comparison with the original ISPM. Note that equal line colours correspond to the same discretisation level.
Figure 2: Streamlines, deformation and flow slices at $t = 0.28s$. Contour colors correspond to horizontal component of fluid/solid velocity. Red spherical markers are passive flow tracers.

Figure 3: Streamlines, deformation and flow slices at $t = 2.44s$. Contour colors correspond to horizontal component of fluid/solid velocity. Red spherical markers are passive flow tracers.
REFERENCES


MINIMIZATION OF THE PHENOMENON OF VIBRATIONS INDUCED BY VORTICES USING SURROGATE BASED OPTIMIZATION

ULISSES P. FARIAS FILHO*, ALESSANDRO R. E. ANTUNES† SILVANA M. A. BASTOS‡ AND PAULO R. M. LYRA§

* Civil Engineering Department
Federal University of Pernambuco (UFPE)
Recife, Brazil.
e-mail: uliporto@hotmail.com

† Technology, Center
UFPE
Caruaru, Brazil.
e-mail: areantunes@yahoo.com.br

‡ Civil Engineering Department,
UFPE
Recife, Brazil.
e-mail: smb@ufpe.br

§ Mechanical Engineering Department,
UFPE
e-mail: prmlyra@padmec.org

Key words: Fluid Structure Interaction, ALE FEM Formulation, VIV, Optimization, Surrogate Models

Abstract. In this paper we briefly describe the finite element procedure developed to simulate the two dimensional fluid-structure interaction of a rigid circular cylinder, immerse in an incompressible viscous fluid flow which is coupled with optimization techniques in order to minimize the amplitude of the cylinder displacements. Due to the high computational cost associated with the numerical simulations, surrogate models are built using kriging based data fitting scheme. The results obtained when the whole methodology is applied to the minimization of Vortex Induced Vibration (VIV) on Fluid Structure Interaction (FSI) problems are presented and discussed.

1 INTRODUCTION

Several applications, in different engineering fields, are subjected to vibration as a result of flow induced phenomena. Such behavior can compromise the integrity of the structure or make it uncomfortable for human use. The analysis of these problems, involve the study of a coupled fluid-
structure interaction model and can be done using computational modeling.

This work uses a stabilized like Petrov-Galerkin “ALE” finite element formulation with Euler timeintegration for the fluid-dynamics analysis [1]. This scheme represents an SUPG-like algorithm (Streamline Upwind Petrov-Galerkin [2] with the Fractional Step Method to stabilize the pressure field. For the structural analysis it uses a simple lumped model with three degrees-of-freedom and the Newmark Method [3]. The fluid-structural coupling is solved through interfacing and implemented and tested in a segregated approach, using an algorithm to control errors due to the existing time delay between the fluid and structural analysis [4].

In order to minimize VIV, here we employ two techniques. The first one is an acoustic signal, simulated as an increase of the boundary layer linear momentum, in order to control the flow field parameters, specifically the vortex shedding frequency. The second technique applied is the positioning of a plate behind the cylinder.

In order to minimize the transversal vibrations of the cylinder, the parameters involved in the cases studied were investigated by the application of optimization techniques. As optimization techniques commonly involves several calls of the numerical simulator, which may turn the optimization task into a very time consuming process, surrogate models using kriging based data fitting are employed in substitution to the coupled fluid-structure numerical simulations. The optimization algorithm of choice is the Sequential Quadratic Programming (SOP). This will be embedded here in an interactive procedure, named Sequential Approximate Optimization (SAO) [5]. A trust region based method is used to update the design variable space for each local (sub problem) optimization solution. As will be shown, the optimization process resulted in a reduction of up to 85% of the vibrations amplitude.

2 NUMERICAL FORMULATION

The coupling between fluid and structure fields is characterized by displacements of some of the boundaries of the domain. The regions close to these moving boundaries are more naturally discretized with a Lagrangean approach. The fluid regions away from the moving boundaries, however, are more naturally treated with a conventional Eulerian formulation, with a fixed reference frame. In this work, we use an Arbitrary Lagrangean Eulerian framework to combine these two approaches in a single numerical technique. The differential equations that describe the dynamics of the fluid and the structure therefore must be written in this framework. Hence, there are three different fields that characterize a fluid structure interaction problem: fluid dynamics, structure dynamics and mesh dynamics which will be described below.

2.1. Fluid Dynamics

The incompressible Navier-Stokes Equations, in an ALE description, without thermal effects, in continuous form, can be written as:

\[
\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{c} \cdot \nabla) \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega \times (0,t)
\]

\[
\nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad \Omega \times (0,t)
\]

where \( \Omega \) is the spatial fluid domain, \( t \) is the time variable, \((0,t)\) is the time interval, \( \mathbf{u} \) is the velocity field, \( \nu \) is the kinematic viscosity, \( p \) is the pressure, \( \mathbf{f} \) is the external force vector, \( \nabla \) is the gradient operator and \( \Delta \) is the Laplacian operator, and \( \mathbf{c} \) is the relative velocity field between fluid and mesh. Considering that the Eqs. (1-2) must be manipulated in order to generate a Poisson equation for pressure field, the physical boundary was divided in two non-overlapping parts \( \Gamma_{in} \) and \( \Gamma_{out} \) in which the Dirichlet and Neumann boundary conditions are prescribed to each equation, respectively. The
Dirichlet and Neuman boundary conditions are:

\[ \mathbf{u} = \mathbf{u}_0 \text{ in } \Gamma_{du}, \quad p = \bar{p} \text{ and } \mathbf{n} \cdot \mathbf{\sigma} = \mathbf{t} \text{ in } \Gamma_{ne} \]  

(3)

where \( \mathbf{\sigma} \) is the viscous stress tensor, \( \mathbf{n} \) is the unit outward normal vector, and \( \mathbf{t} \) is the surface stress or traction. An upper bar refers to a prescribed value. Finally, initial conditions must be known in the whole domain at the initial time.

In this work, a Fractional Step method based in a LU factorization \([1,6-9]\) was applied. In this method the final system is analogous to the method proposed by Chorin \([10]\), and Temam \([11]\), that applies Helmholtz decomposition. The final discrete system is obtained by using a \( \theta \) method in time, resulting in a trapezoidal discretization, and a Finite Element Method for the spatial discretization \([12,13]\). The stabilization of the advective and the gradient pressure terms are obtained with an orthogonal projection of these terms in a finite element space \([7,14]\).

The system of equations from the discrete variational formulation is solved in a segregated way applying a Gauss-Seidel procedure. At the convergence of the block Gauss-Seidel, all system converges to the monolithic system. We are using an edge based data structure, which is advantageous in terms of CPU time, because, in the adopted procedure, most of the discrete terms do not need to be re-computed in each iteration by looping through the elements.

### 2.2. Structural Dynamics

In this work we only consider dynamics of rigid bodies. The movement of the body is obtained with a straightforward application of Newmark’s Method \([3,16]\). The formulation used in this work leads to an implicit, second order accurate and unconditionally stable time integration scheme. The unconditional stability of this scheme is important because the time step increment for the structural time evolution is taken as the same as the time increment chosen for the CFD solution. This time increment is determined by the stability requirements of the CFD algorithm, and therefore its time scale is completely unrelated to the dynamic behavior of the structure.

### 2.3. Mesh Dynamics

In the domains with a ALE formulation, the movement of the interface nodes causes distortions on the original shapes of the elements connected to these nodes.

In order to avoid excessive elements distortion, in this work we are solving a modified Laplace equation for the mesh problem that is solved by Finite Element Method using an edge based data structure. The diffusivity coefficient is based on the volume of the elements, and is designed to smooth the distortions caused by the structure displacements \([17,18]\).

The domain is divided as follow: \( \Omega \) is the domain of the equation, \( \Gamma_m \) is the moving boundary (structure surface), and \( \Gamma_f \) is the fixed boundary. So the classical formulation of the problem results:

given \( \mathbf{v} \) on the boundaries, find \( \mathbf{v} \) in the domain, such that:

\[ \nabla \cdot (\left[1 + \tau \right] \nabla) \mathbf{v} = 0 \]  

(4)

\[ \mathbf{v} = \mathbf{v}_0 \text{ in } \Gamma_m \]  

(5)

\[ \mathbf{v} = 0 \text{ in } \Gamma_f \]  

(6)

where \( \Gamma = \Gamma_m \cup \Gamma_f \), \( \mathbf{v} \) is the displacement vector, \( \mathbf{v}_0 \) the moving boundary displacement, and it’s required to solve the equation for each direction.

Note that the diffusivity coefficient must be defined in order that the smaller elements, generally...
defined close to the structure, where the small scale effects are present, suffer minimum deformations. This way, the displacements, are sent through the mesh until the bigger elements, close to the fixed boundaries, can smooth the displacements. Kanchi e Masud, [17], proposed the calculus of the coefficient to avoid excessive element deformations. The element coefficient used by Kanchi and Massud [17], Eq. (7), and adopted in this work is:

\[ \tau' = \frac{1 - \frac{V_{\text{min}}}{V}}{\frac{V'}{V_{\text{max}}}} \]  

where \( V_{\text{min}} \) = minimum element volume of the mesh, \( V_{\text{max}} \) = maximum element volume of the mesh, \( V' \) = element volume \( e \).

There are many practical aspects to a successful computational implementation of the procedures described above. Clearly, facilities for dealing with deformable domains, which involve automatic mesh generation, assessment of mesh quality, and automatic mesh movement are all important aspects. Also is important the choice of the coupling algorithm in order to warranty the kinematic and dynamic compatibility and the Geometric Conservation Law. All these were considered and implemented in this work [16].

3 PROBLEM DEFINITION

3.1 Acoustic Excitation

In the literature (Hiejima et al., [19] and references therein) several experimental and numerical results are reported in which acoustic excitation is applied to an external flow to increase the momentum transfer from the outside flow to the boundary layer and eliminating (or delaying) separation and suppressing (or reducing) vortex induced vibrations in different solid configurations. In this article, our computational system is used to perform an study on the behavior of the fluid-structure problem described by Hiejima et al. [19] in which an idealization of the acoustic excitation is obtained through the application of a periodic velocity excitation on two points at the cylinder surface (see Figure 1). The angle between the stagnation point and the excitation points at \( \phi_{a} \). The excitation velocity is given by:

\[ V_{a} = U_{a} \sin(2\pi f_{a} t) \]  

where, \( U_{a} \) and \( f_{a} \) refer to the periodic velocity excitation amplitude and frequency, respectively. The two excitation velocities are in phase.

![Figure 1: Computational model data and boundary conditions, and description of the periodic velocity excitation on the surface of the cylinder](image-url)
Figure 1 also shows the full description of the numerical model, including the: computational domain, boundary conditions, fluid properties and structural parameters of the mass-stiffness-damping system. The free stream velocity is $U = 0.0264 \text{ m/s}$ and the Reynolds number based on the cylinder diameter is 200. Initially, we performed some numerical simulations considering a fixed cylinder and the vortex shedding frequency obtained at $Re = 200$ was $f_s = 0.043478 \text{ Hz}$. This value is in good agreement with the experimental curve presented by Blevins [20]. The effect of the periodic velocity excitation was initially investigated considering different ratio between the values of the excitation frequency ($f_a$) and the vortex shedding frequency ($f_s$), i.e. $f_a / f_s = 1.00; 3.51$ and $4.45$, see Figure 2. Accordingly to Hiejima et al. (1997), the ratio value of 4.45 is close to the experimental value near the transition wave frequency, which is an effective value of frequency for an acoustic excitation to change the flow around a stationary circular cylinder. With such value of excitation they were able to get a considerable increase on the vortex shedding frequency that was quite effective in reducing the vortex induced vibration amplitude, as the experimental results suggests. We picked up two other values around 4.45 in order to study the influence of the excitation frequency on our results. Considering a fixed cylinder and the different ratio ($f_a / f_s$) mentioned previously, the frequency of the velocity transversal to the flow in a point located inside the vortex shedding region behind the cylinder was studied.

The same analyses were performed considering the cylinder free to vibrate in the direction transversal to the flow. The numerical simulation set up consists of starting with a fixed cylinder and after the vortex shedding becomes periodic we allow the transverse movement, and after the vibration amplitude stabilizes on a constant value we start applying the periodic excitation. In all three cases the cylinder is set free when the time is around 80 seconds and the excitation starts when the time is around 180 seconds. In Figure 2 the displacement histories are plotted for $f_a / f_s = 1.00, 3.51, 4.45$, respectively. For $f_a / f_s = 4.45$, there is reduction on the oscillatory amplitude, and the adopted excitation frequency has an effective effect. The results suggest that an even bigger variation on the vortex shedding frequency might reduce more or even suppress the vibration on the cylinder.

![Figure 2: Time history of transversal displacement for $f_a/f_s = 1.00, 3.51, 4.45$, respectively.](image)

It should be observed that the amplitude of the oscillations were small with the cylinder vibrating under the influence of the vortex formation and shedding behind the cylinder, and that the characteristic of the vortex induced vibrations were directly affected by the change on the frequency of such vortex formation and shedding. Also, further investigation considering different ration between the values of the excitation frequency ($f_a$) and the vortex shedding value ($f_s$), and also considering different application points and amplitude of excitation were pursued in order to gain a better insight on the behavior of this application to conduct further studies considering optimization techniques.

### 3.2 Positioning of a plate behind the cylinder

The model described by Zdravkovich [21] and studied by Correia [22] for the suppression (or reduction) of vortex shedding consists in positioning a flat plate behind the cylinder, where the vortex
street is developed. The length of the plate is assumed to be equal to the one studied by Correia [21], fixed as 1.14 times the cylinder diameter. The position of the plate relative to the cylinder is the design variable of the problem ($c$).

In order to compare the results with those obtained by the application of the acoustic excitation, the physics and geometric parameters were considered the same for both cases. Figure 3 shows the geometry scheme adopted for this case.

![Figure 3: Problem definition to solution by positioning a plate behind the cylinder.](image)

### 3.3 Formulation of the Optimization problem

Mathematically, a general optimization problem is formulated as:

Minimize $f(x)$

$s.t.$ $h_k(x) = 0, \quad k = 1, ..., ne$

$g_i(x) \leq 0, \quad i = 1, ..., ni$

$x_{lj} \leq x_j \leq x_{uj}, \quad j = 1, ..., ndv$

(9)

in which $x$ are the design variables. The function $f(x)$ is the objective of the problem. The functions $g_i(x)$ and $h_k(x)$ represent, respectively, the inequality and equality constraints. The side constraints have inferior limits $x_{lj}$ and superior limits $x_{uj}$. $ne$, $ni$ and $ndv$ are, respectively, the numbers of equality constraints, inequality constraints and design variables.

In preparation for the optimization process, for the particular problems here addressed, some procedures are required to be previously performed. Considering the cylinder fixed, a fluid problem at $Re=200$ is analyzed until is ensured the stabilization of the vortex shedding frequency. This frequency is calculated and registered through the transversal velocity of a point inside the vortex street behind cylinder. Then, the analysis proceeds and the cylinder's displacements are released, turning the simulation into a fluid-structure interaction problem. At this time, the structure's displacements start to increase until stabilization, when finally a vortex shedding suppression technique is ready to be applied.

When considering optimization tools, the objective function is the amplitude of the transversal displacements of the cylinder.

For the particular problem of the acoustic excitation, it should be observed from Figure 4 (a) that simulation's time after releasing the cylinder can be divided into three intervals. The first one corresponds to the time interval before the application of the acoustic excitation, followed by the second interval, when the acoustic excitation is applied and a modification on amplitude is observed. On the third and last interval, between 300 and 600 seconds, the amplitude of the cylinder’s transversal displacements is considered stabilized and the absolute maximum value of transversal displacement is calculated and sent to optimization algorithm as the objective function value. The parameters to be changed (design variables (dv) candidates) are: frequency, amplitude and the location of the acoustic excitation.
The objective function of the plate problem is evaluated by an analogous way to that of the acoustic excitation, but at the same time that cylinder is released, the plate slowly starts to move from its initial position to the position defined by the optimization algorithm. The movement of the plate to its final position lasts up to 200 seconds of simulation. For this case, the stabilization of the cylinder vibrations is considered only on the interval defined by the final 300 seconds of simulation. During the preliminary tests, it was observed that after the release of the cylinder, an initial and larger displacement is observed and the cylinder assumes a new position of vibration. This effect is shown on Figure 4 (b). Once we are here interested in the reduction of vibrations, this displacement was not considered for evaluation of the objective function. So, the amplitude of vibration considered is relative to its final position of vibration.

4 SURROGATE MODEL

The challenge in the process of surrogate model construction is to provide a substitute model with sufficient accuracy. Several strategies can be used to build the approximated model [23]. Here, kriging (ordinary) based data fitting approach is considered. The main idea of this model is to assume that errors are not independent but rather assume that the correlations between errors is related to the distance between corresponding points modeled by a Gaussian process around each sample point. The main advantages of this scheme are to easily accommodate irregularly distributed sample data, and the ability to model multimodal functions with many peaks and valleys. Kriging models provide exact interpolation at the sample points. Details of such procedure can be seen elsewhere [23].

5 OPTIMIZATION STRATEGY

In order to obtain the optimum design of the problem tackled in this work the SAO methodology will be employed. As surrogate models have a limited range of accuracy, the design space of the approximate optimization problem is restricted to the sub region called trust region whose dimensions are adaptively managed by the SAO strategy depending on surrogate accuracy [5]. Mathematically, each sub problem $k$ is defined as:
Minimize $\hat{f}(x)$

Subject to

$\hat{g}^k(x) \leq 0, i = 1, \ldots, m$  $\hat{g}^{\Delta}(x) \leq 0, i = 1, \ldots, m$

$x^k_i \leq x^k \leq x^u_i$, $k = 0, 1, 2, \ldots k_{max}$

Where

$x^k_i = x^k - \frac{\Delta^k}{2}$; $x^u_i = x^k + \frac{\Delta^k}{2}$

In the above equations, $\hat{f}(x)$ and $\hat{g}(x)$ are, respectively, the surrogate objective and constraints functions, $x^k_i$ is the center point of the trust region, $\Delta^k$ is the width of the trust region and $x^k_i$, $x^u_i$ are, respectively, the lower and upper bounds of the design variables at the $k^{th}$ SAO iteration [5].

5.1 Reuse of samplings

When considering the original algorithm of the SAO methodology, the high fidelity model simulations are conducted on a set number of points in the trust region for each iteration. The selected output computed at the samplings are used to build the surrogate model in the sub region of the design space. In the next iteration, other points (same quantity) are selected to build another surrogate model.

In this work, the algorithm used for application of SAO strategy is modified. The modification consists in save the samples of the high fidelity model in a data base during the SAO process. Then, in each SAO iteration, this data base is consulted in order to check the possibility to reuse the samples for the new surrogate model construction. With this approach, the information obtained by a high computational cost can be reused during the process. The details of such scheme can be found in [24].

6 EXAMPLES

6.1 Optimization of the acoustic excitation problem

An optimization problem considering as design variables the frequency and the amplitude is solved by two versions of SAO procedure. Initially, the SAO algorithm already used on the optimization of single design variables is considered. Next, some modifications were applied to the algorithm, resulting in a potential gain of performance and robustness. The modification encompasses the reuse of samplings from previous SAO iterations that are located in the current sub problem trust region. Those points together with the generated samples of current sub region make a richer model to build the surrogate to be used in the current SAO iteration. This higher level of information allows surrogate models to represent a better approximation to the high fidelity model without computational cost addition, once those samples were already evaluated which in the previous SAO version were being discarded. A preliminary study [24] considering an analytic function shows the potential gain of performance with the modifications of the algorithm.

For the problem here analyzed, using both SAO strategies, the angle was fixed as 80º and the initial design values for frequency and amplitude were considered 4.00 and 65 cm/s respectively, in design spaces limits between 0 to 10 and 50 to 90 cm/s.

In the initial SAO version, whose results are presented on Figure 5 (a) to (c), the parameters that lead to the maximum reduction of vibrations of the cylinder were found in nine SAO iterations. Relative to the initial design, the optimum value found represents a reduction of approximately 50%.

Figure 5: 2 dv problem (traditional SAO) – iteration histories: (a) Frequency, (b) Amplitude and (c) Objective function.

Figure 6 (a) to (c) present the results of SAO procedure with re-use of samples from previous iterations.

Figure 6: 2 dv problem (SAO with re-use sampling) – iteration histories (a) Frequency, (b) Amplitude and (c) Objective function.

It can be observed that with the later SAO alternative a reduction on the number of iterations when compared with the traditional procedure, which represents an economy of computational cost of almost a half. Moreover, the optimum values found with both procedures were similar.

Table 1 and 2 present in summary the numerical results obtained using the SAO strategy for the optimization of the acoustic excitation problem. The shaded cells of Table 1 refer to the fixed variables on the optimization process. In both tables the following cases are presented:

- **SAO 01**: Two design variables optimization: frequency and amplitude of acoustic excitation (traditional SAO scheme);
- **SAO 02**: Two design variables optimization: frequency and amplitude of acoustic excitation (SAO with re-use sampling scheme);

**Table 1: Summary of SAO procedures – Design Variables.**

<table>
<thead>
<tr>
<th>Optimization Procedure</th>
<th>Initial design</th>
<th>Optimum design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Amplitude</td>
</tr>
<tr>
<td>SAO 03</td>
<td>4.0000</td>
<td>65.0000</td>
</tr>
<tr>
<td>SAO 04</td>
<td>4.0000</td>
<td>65.0000</td>
</tr>
</tbody>
</table>

Table 2: Summary of SAO procedures - Objective Function and number of iterations.

<table>
<thead>
<tr>
<th>Optimization Procedure</th>
<th>SAO Iterations</th>
<th>Objective Function</th>
<th>Objective Function Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Initial Point</td>
<td>Optimum Point</td>
</tr>
<tr>
<td>SAO 04</td>
<td>9</td>
<td>0.4446</td>
<td>0.2170</td>
</tr>
<tr>
<td>SAO 05</td>
<td>5</td>
<td>0.4446</td>
<td>0.2210</td>
</tr>
</tbody>
</table>

The re-use scheme highlights its capability in terms of SAO iterations. Other optimizations processes considering one and three design variables were also conducted and can be found in [24].

6.2 Optimization of the flat plate problem

As described previously, only one design variable was considered for the problem of positioning a flat plate behind the cylinder. This design variable is the distance between the plate and cylinder, and is normalized by the cylinder diameter.

The initial design variable value considered is 3.50, based on the parametric study presented in previous section. The design space was defined between 1.45 and 4.34, referred by the studies of Correia [22]. The SAO algorithm that reuses samples is used for this problem. After five SAO iterations, the strategy converged to a minimum of 0.1362. This value represents a reduction on cylinder vibrations of almost 87% related to the original configuration (without plate). Taking the initial objective function value as reference, the reduction of the objective function on the optimum solution is equal to 3%. However, it is important to highlight that the initial point chosen was expected to be near the optimum, based on the parametric study previously conducted.

Figure 7 (a) and (b) present, respectively, the histories of the design variable and the objective function over the SAO iterations of this problem.

![Figure 7](image)

**Figure 7**: Flat plate problem – SAO histories: (a) design variable and (b) objective function.

7 CONCLUSIONS

In the present work an integrated tool to solve fluid-structure interaction problems. The VIV problem was investigated obtaining very good results. The overall conclusions of present study are:

- The use of surrogate models was fundamental to conduct the optimization of the fluid structure interaction problems.
- Reuse of samplings at SAO algorithm provided computational savings and increased its robustness.
- The optimization considering acoustic excitation represented 80% vibration reduction. So far the best literature reported improvement is 60%.
• For the flat plate behind the cylinder problem, the application of the developed optimization tool reduced the cylinder vibrations in 86%.
• The techniques employed were satisfactory for its aim and should be applied to other examples.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support for this research given by CNPq (National Research Council, Brazil) and Pernambuco Research Council (FACEPE).

REFERENCES

element scheme for incompressible flow problems”, In: 15th AIAA Computational Fluid
Dynamics Conference, Anaheim, EUA, June.
incompressible flow formulation”, Comp. Meth. Applied Mechanics and Engineering,
vol. 193, p. 2139-2154.
Computational System for the Simulation of Fluid-Structure Interaction Problem”. J. of
Finite Element Method for the solution of Mesh Movement problem in Fluid-Structure
applications”, VI National Congress of Mechanical Engineering, Campina Grande, PB,
Brazil, 11p, (in Portuguese).
suppression of the vortex-induced vibration of a circular cylinder by acoustic excitation.
Journal of Wind Engineering and industrial aerodynamics, pp. 325-335.
Malabar/Florida, USA, 363 p.
21- Zdravkovich, M., 1981. Review and classification of various aerodynamic and
145-189.
22- Correia, A. C., 2001. “Computational simulations of vortices formations on a circular
cylinder with emphasis on Vortex Induced Vibrations”. M. Sc. Thesis. Recife, Brazil (in
Portuguese)
23- Forrester, A., Sobester, A., Keane, A., Engineering Design via Surrogate Modeling: A
24- Farias Filho, U. P., 2012. “Minimization of the phenomenon of Vibrations Induced by
Vortices using Sequential Approximation Optimization”, M.Sc. Thesis, Recife, Brazil (in
Portuguese).
3d conservative coupling method between a compressible fluid flow and a deformable structure

Maria Adela Puscas*,†‡, Virginie Daru‡, Alexandre Ern*, Christian Mariotti†, Laurent Monasse*, Christian Tenaud‡

* Université Paris-Est, CERMICS, École des Ponts ParisTech
77455 Marne la Vallée cedex, France
email: {puscasa, ern, monassel}@cermics.enpc.fr
† CEA, DAM, DIF, F-91297 Arpajon, France
email: {adela.puscas, christian.mariotti}@cea.fr
‡ LIMSI-CNRS, 91403 Orsay, France
email: {adela.puscas, virginie.daru, christian.tenaud}@limsi.fr

ABSTRACT

In this work, we present a conservative method for three-dimensional inviscid fluid-structure interaction problems. On the fluid side, we consider an inviscid Euler fluid in conservative form. The Finite Volume method uses the OSMP high-order flux with a Strang operator directional splitting [1]. On the solid side, we consider an elastic deformable solid. In order to examine the issue of energy conservation, the behavior law is here assumed to be linear elasticity. In order to ultimately deal with rupture, we use a Discrete Element method for the discretization of the solid [2]. An immersed boundary technique is employed through the modification of the Finite Volume fluxes in the vicinity of the solid. Since both fluid and solid methods are explicit, the coupling scheme is designed to be globally explicit too. The computational cost of the fluid and solid methods lies mainly in the evaluation of fluxes on the fluid side and of forces and torques on the solid side. The coupling algorithm evaluates these only once every time step, ensuring the computational efficiency of the coupling. Our approach is an extension to the three-dimensional deformable case of the conservative method developed in [3]. We focus herein numerical results assessing the robustness of the method in the case of a undeformable solid with large displacements subjected to a compressible fluid flow.

1 Introduction

Many situations involve phenomena of fluid-structure interactions. The study of such phenomena is motivated by the fact that the consequences are sometimes catastrophic for the mechanical structure. For example, in the military or safety domains, the effects of an explosion on a building or on a submarine involve complex non-linear phenomena (shock waves, cracking, rupture, ...) [12]. In view of these applications, we consider on the fluid side an inviscid compressible fluid flow model so as to deal with air shock waves and on the solid side an elastic deformable solid.

The main challenges in fluid-structure interaction problems are the computation of the fluid forces that act on the rigid or deformable structure and the modification of the fluid domain due to the displacement of the solid. Another difficulty that arises in fluid-structure interaction problems is the coupling of models with different descriptions: the fluid is classically described in Eulerian formulation and the elastic structure in Lagrangian formulation. Several types of methods have been developed for this purpose. Fully Eulerian [10], fully Lagrangian [11], Arbitrary Lagrangian-Eulerian (ALE) methods [4] and fictitious domain methods ([5], [7], [8]) have been proposed to address the issue of fluid-structure interaction. In general, monolithic Eulerian or Lagrangian approaches are limited to the case where the fluid and solid behave according to similar equations with different parameters. The ALE method deforms the fluid domain in order to follow the movement of the structure. It is a method with adapted mesh on the solid boundaries which requires possibly costly re-meshing of the fluid domain when the solid goes through large displacement or rupture.

In fictitious domain methods the solid is superimposed on the fluid fixed grid, and additional terms are introduced in the fluid formulation to penalize or prevent the penetration of fluid inside the solid. Various
types of fictitious domain methods have been developed: Immersed Boundary methods [5], Ghost Fluid methods [6], Embedded Boundary methods [7], etc.

An important issue in compressible fluid-structure interaction is the conservation of mass, momentum and energy. The accurate capture of shocks is based on conservation properties and the preservation of physical properties is an important ingredient towards an effective numerical method. In addition, verifying conservation at a discrete level is a natural means to prove the numerical stability of the scheme. Embedded Boundary methods [8] are built in such a way that the spatial discretization satisfies mass, momentum and energy conservation.

We therefore use the Embedded Boundary method developed by Colella et al. [8] in combination with a Finite Volume method for the fluid and a Discrete Element method for the solid. The Finite Volume method is computed on a Cartesian grid, using high-order upwind fluxes [1] computed with a Lax-Wendroff approach. The Discrete Element method [2] is a particle method for elastodynamics, in which particles interact through forces and torques yielding the macroscopic behaviour of the assembly. Both methods being time-explicit and computationally expensive, we require that the coupling algorithm be explicit too. The method is tailored to yield the exact conservation of mass, momentum and energy of the system and exhibits consistency properties. A conservative explicit coupling algorithm between the Finite Volume method and 2d undeformable solid has already been developed in [3]. Herein we extend the results to the 3d deformable case.

This paper starts with a brief description of the fluid and solid methods. Then, we present the proposed conservative coupling method and the explicit time-integration coupling procedure between the inviscid fluid and the deformable moving structure. Next, we point out several properties of the coupling method. Finally, we present numerical results showing the energy and mass conservation by the coupling scheme and the ability of the method to compute the interaction of strong discontinuities with irregular moving boundaries.

2 Fluid and solid description

2.1 Inviscid compressible flow

The fluid is modeled by the Euler equations expressing conservation of mass, momentum and energy for an inviscid compressible flow, which are written in Cartesian coordinates:

\[
\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} + \frac{\partial G(U)}{\partial y} + \frac{\partial H(U)}{\partial z} = 0,
\]

where \( U = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho E \end{pmatrix} \), \( F(U) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ \rho u w \end{pmatrix} \), \( G(U) = \begin{pmatrix} \rho v \\ \rho u v \\ \rho v^2 + p \\ \rho v w \end{pmatrix} \), \( H(U) = \begin{pmatrix} \rho w \\ \rho u w \\ \rho v w \end{pmatrix} \), and \( \gamma = 1.4 \) the ratio of specific heats, assumed to be constant.

The numerical resolution of these equations is based on a Cartesian grid explicit Finite Volume method and directional operator splitting. For the flux calculation we use the OSMP numerical scheme which is a one-step high order scheme [1]. It is derived using a coupled space-time Lax-Wendroff approach, where the formal order of accuracy in the scalar case can be set at an arbitrary order. In this work we use order 11. The coupling method is actually independent from the numerical scheme used for the flux calculation.

2.2 Deformable solid discretization method

The deformable solid is modeled using a Discrete Element method. The solid is discretized with a finite number of rigid particles. Each particle is governed by the classical equations of mechanics and interparticle links ensure the macroscopic behaviour of the solid [2]. The solid is discretized into polyhedral
particles which have a rigid-body motion and interact through forces and torques. The expression of these forces and torques allows one to recover the macroscopic behaviour of materials. Furthermore, this approach facilitates the handling of rupture and fracture, because breaking the link between material grains ensures the loss of cohesion between solid particles.

The position and velocity of the solid particles are given respectively by the position of their center of mass \( \mathbf{X} \), the rotation matrix \( \mathbf{Q} \), the velocity of the center of mass \( \mathbf{V} \) and the angular momentum matrix \( \mathbf{P} \). The movement of solid particles is integrated in time by an explicit scheme using the Verlet scheme for translation and the RATTLE scheme for rotation [3].

3 Coupling method

Fluid-structure interaction needs to take into account the solid obstacles in the flow calculation as well as the fluid efforts on the solid. For the first point we use a method of immersed boundaries which allows for easy modification of existing fluid solvers. In this context, the solid and fluid grid overlap, leading to fluid-solid mixed cells, thereafter called “cut-cells”.

3.1 Cut-cells description

Since the solid is discretized into polyhedral particles, the fluid-solid interface is simply the set of faces of these polyhedra in contact with the fluid.

We denote with integer subscripts \( i, j, k \) quantities related to the center of cells and with half integer subscripts quantities related to the center of faces. Let \( C_{i,j,k} \) be a cut-cell of size \((\Delta x_{i,j,k}, \Delta y_{i,j,k}, \Delta z_{i,j,k})\) and \( \Omega_S(t) \) the solid domain. The relevant aspects of the intersection between the moving interface and cell \( C_{i,j,k} \) are:

- the **volume fraction** \( 0 \leq \Lambda_{i,j,k} \leq 1 \) occupied by the solid in the cell \( C_{i,j,k} \):
  \[
  \Lambda_{i,j,k}(t) = \frac{V_{i,j,k}(t)}{V_{i,j,k}},
  \]
  \( V_{i,j,k} = (\Delta x \Delta y \Delta z)_{i,j,k} \) being the volume of \( C_{i,j,k} \) where the solid occupies the volume \( V_{i,j,k}(t) \) at instant \( t \):
  \[
  V_{i,j,k}(t) = \int_{C_{i,j,k} \cap \Omega_S(t)} dx \ dy \ dz.
  \]
  The volume fraction is evaluated at the discrete time \( t^n \). The interface between cells \( C_{i,j,k} \) and \( C_{i+1,j,k} \) is denoted \( \partial C_{i+1/2,j,k} \).

- the **side area fraction** \( 0 \leq \lambda_{i,j,k} \leq 1 \) of each face; for example on face \( \partial C_{i+1/2,j,k} \) we have
  \[
  \lambda_{i,j,k}^{n+1/2} = \frac{A_{i+1/2,j,k}^{n+1/2}}{(\Delta y \Delta z)_{i,j,k}},
  \]
  where \( A_{i+1/2,j,k}^{n+1/2} \) is the area representing the intersection of the solid with face \( \partial C_{i+1/2,j,k} \) averaged over \([t^n, t^{n+1}]\):
  \[
  A_{i+1/2,j,k}^{n+1/2} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{\partial C_{i+1/2,j,k} \cap \Omega_S(t)} dy \ dz \ dt.
  \]

- the **boundary area** denoted \( A_{f_{i,j,k}}^{n+1/2} \), is the area of the surface formed by the intersection of the fluid with the solid in the cell during the time interval \([t^n, t^{n+1}]\):
  \[
  A_{f_{i,j,k}}^{n+1/2} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_{C_{i,j,k} \cap \partial \Omega_S(t)} ds \ dt.
  \]
  We also denote \( \vec{n}_{f_{i,j,k}}(t) \) the **outward normal** to the surface (see Fig. 1).
The term $\Delta U$ be very complex and 3d implementation would be highly time consuming [7]. Instead, we take $\lambda$.

The additional flux takes into account the exchange of energy and momentum between the solid and the fluid results from the presence of the solid boundaries $f$ in the cell $C_{i,j,k}$. This flux takes into account the exchange of energy and momentum between the solid and the fluid results from pressure forces.

The computation of average time of the side area fraction $\lambda^{n+1/2}$ and of the boundary area $A_f^{n+1/2}$ can be very complex and 3d implementation would be highly time consuming [7]. Instead, we take $\lambda^{n+1}$ and $A_f^{n+1}$, and we calculate directly the conservative state given by the discrete balance in the cell, as in [3]:

$$
(1 - A_{i,j,k}^{n+1}) U_{i,j,k}^{n+1} = (1 - A_{i,j,k}^{n}) U_{i,j,k}^{n} + \Delta t \left\{ \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta x_{i,j,k}} F_{i-1/2,j,k} - \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta x_{i,j,k}} F_{i+1/2,j,k} + \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta y_{i,j,k}} G_{i,j-1/2,k} - \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta y_{i,j,k}} G_{i,j+1/2,k} + \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta z_{i,j,k}} H_{i,j,k-1/2} - \frac{(1 - \lambda_{i,j,k}^{n+1/2})}{\Delta z_{i,j,k}} H_{i,j,k+1/2} \right\} + \sum_{f \in C_{i,j,k}} A_{f_{i,j,k}}^{n+1/2} \phi_f^{n+1/2},
$$

The additional flux $\phi_f$ results from the presence of the solid boundaries $f$ in the cell $C_{i,j,k}$. The additional flux takes into account the exchange of energy and momentum between the solid and the fluid results from pressure forces.

The term $\Delta U_{i,j,k}^{n+1}$ denotes the amount of $U^n$ swept by the movement of the interface $f$ during the time step; the amount swept by the movement of the interface $f$ is such that
\[
\sum_{C_{i,j,k}} \sum_{f \in C_{i,j,k}} \Delta U_{f_{i,j,k}}^{n,n+1} = \sum_{C_{i,j,k}} (A_{i,j,k}^{n+1} - A_{i,j,k}^n) U_{i,j,k}^n.
\]

In practice, we decompose the interface into triangles which are contained in one cell (not necessarily the same) at times \(n\Delta t\) and \((n+1)\Delta t\). We calculate \(\Delta U_{f_{i,j,k}}^{n,n+1}\) as the integral of \(U^n\) on the prism built on these triangular bases. The detailed procedure to obtain these terms will be given in [13].

The main limitation of immersed boundary methods is that they involve small control cells (“small” in the sense that the solid volume fraction is \(> 0.5\)). In order to ensure the CFL stability condition of explicit schemes on these cells, the time step should be decreased to an unacceptably small value.

To deal with these issues, we use a conservative mixing process following the ideas developed in [9] with minor changes. Let \(p\) be a small cell and \(g\) a completely fluid \((\Lambda_g = 0)\) neighboring cell. We define the following exchange terms:

\[
E_{pg} = \frac{(1 - \Lambda_g)}{(2 - \Lambda_p - \Lambda_g)}(U_g - U_p), \quad E_{gp} = \frac{(1 - \Lambda_p)}{(2 - \Lambda_p - \Lambda_g)}(U_p - U_g),
\]

and we set

\[
U_p = U_p + E_{pg}, \quad U_g = U_g + E_{gp}.
\]

The mixing procedure is fully conservative and ensures that the equivalent volume of a mixed cell is compatible with the usual CFL condition using the standard size cells.

### 3.3 Coupling algorithm with a deformable solid

At the beginning of the time step \(n^{+1} = (n + 1)\Delta t\), we known the state of the fluid \(U_{i,j,k}^n\), the position and rotation of the particle \((X^n, Q^n)\), as well as the velocity of the center of mass and the angular momentum \((V^n, P^n)\). The general procedure for the conservative coupling method can be described by the following five steps:

1. **The fluid fluxes are precomputed.** We denote by \(\overline{p}_x, \overline{p}_y\) and \(\overline{p}_z\) the mean pressures used to compute the fluxes in the \(x, y\) and \(z\) directions.

2. **The fluid forces \(\vec{F}_f\) acting on a planar solid boundary \(f\) of surface \(A_f\) and normal vector \(\vec{n}_f\) are equal to the force exerted by these pressures on the surface in contact with the fluid:**

   \[
   \vec{F}_f \cdot \vec{e}_x = -\overline{p}_x A_f \vec{n}_f^x,
   \]

   \[
   \vec{F}_f \cdot \vec{e}_y = -\overline{p}_y A_f \vec{n}_f^y,
   \]

   \[
   \vec{F}_f \cdot \vec{e}_z = -\overline{p}_z A_f \vec{n}_f^z.
   \]

3. **The solid is advanced in time:** internal forces are computed based on the position of the solid particles. The position of each particle (submitted to a constant external fluid force) is integrated using the Verlet scheme for translation and the RATTLE scheme for rotation [3].

4. **The volume fractions of solid in fluid cells \(A_{i,j,k}^{n+1}\) and surface fractions of solid on cell interfaces \(\lambda_{i,j,k}^{n+1} = \lambda_{i,j,k}^{n+1} \cup \lambda_{i,j,k}^{n+1} + 1/2\) can then be computed using the new position of the solid boundary. The fluid fluxes are modified using \(\Lambda_{i,j,k}, \Lambda_{i,j,k}^{n+1}\) and \(\lambda_{i,j,k}^{n+1} = \lambda_{i,j,k}^{n+1} + 1/2\), pressures \(\overline{p}_x, \overline{p}_y\) and \(\overline{p}_z\) and the velocity of the boundary in order to enforce the conservation of fluid mass and of the total momentum and energy of the system.

5. **The final value of the state \(U_{i,j,k}^{n+1}\) in the cell is calculated using (5).** Owing to the perfect slip conditions at the solid boundary, the flux \(\phi_{f_{i,j,k}}^{n+1}\) is given by

   \[
   \phi_{f_{i,j,k}}^{n+1} = \frac{1}{A_{f_{i,j,k}}^{n+1}} \left(0, \Pi_x, \Pi_y, \Pi_z, V_f^{n+1/2}, \Pi \right)^{T},
   \]
where
\[
\Pi = \left( \int f_x n_{f_i,j,k}^{n+1,x}, \int f_y n_{f_i,j,k}^{n+1,y}, \int f_z n_{f_i,j,k}^{n+1,z} \right)^{t},
\]
and \(V_f^{n+1/2}\) is the velocity in the center of the interface in the cell \(C_i,j,k\):
\[
V_f^{n+1/2} = V^{n+1/2} + \Omega^{n+1/2} \wedge (X_f^{n+1} - X_f^{n+1}),
\]
where \(X_f^{n+1}\) is the center of the part of interface \(f\) at time \((n+1)\Delta t\) in the cell, \(X_f^{n+1}\) the center of the particle containing \(f\), and \(V^{n+1/2}\) and \(\Omega^{n+1/2}\), respectively, the average velocity and rotation velocity of the particle in the time interval \([n\Delta t, (n+1)\Delta t]\).

The general structure of the explicit coupling scheme is presented in Fig. 2.

Figure 2: Structure of the explicit coupling scheme

### 3.4 Properties of the coupling scheme

The properties stated in this section can be proven and we have verified them numerically (see [3] for the proof in 2d; the proof in 3d is similar).

**Conservation of mass, momentum and energy**

These properties hold for periodic boundary conditions and in the cases where there is physical conservation (i.e. mass and energy with fixed boundaries, conservation when boundaries are far ...).

**Perfect slipping along a wall**

The coupling algorithm preserves exactly a uniform constant flow parallel to a rigid half-plane, even in the case when the solid is not aligned with the fluid grid. This last result shows that no artificial roughness appears on the solid walls.

**Galilean invariance**

Consider an arbitrary shaped rigid body moving at constant velocity and without rotation, immersed in a uniform fluid flowing at the same velocity. Then, the uniform movement of the fluid and the solid is retained by the coupling algorithm.
4 Numerical results

4.1 Conservation of mass and energy

In order to verify the conservation by the coupling scheme, we consider a test consisting in a simple shock tube in a straight rectangular channel and a rigid mobile body inside this channel. The computational domain is a simple rectangular box \((x, y, z) \in [0, 2] \times [0, 1] \times [0, 1]\) and the initial flow field is given by

\[
\begin{align*}
\rho &= 1.4, \quad u = v = w = 0, \quad p = 5 & \text{if } x < 0.16, \\
\rho &= 1.4, \quad u = v = w = 0, \quad p = 1 & \text{if } x \geq 0.16.
\end{align*}
\]

The solid is represented by the cuboid \((x, y, z) \in [0.4, 0.9] \times [0.4, 0.6] \times [0.4, 0.6]\). The computation is performed on a \((140 \times 70 \times 70)\) grid with periodic boundary conditions. The simulation time is 1 s. The pressure and density fields on the outer domain boundaries at time 1 s are shown in Fig. 3. We examine more precisely the pressure and density distribution along the line \(y = 0.5\) in the plane \(z = 0.75\) in Fig. 4. The shocks and rarefaction waves are well captured by our method, without spurious numerical oscillations. The resolution of the shocks is obviously moderate due to the coarseness of the fluid mesh.

![Figure 3: Distribution of pressure (a) and density (b) at time \(t = 1\).](image)

![Figure 4: Pressure distribution (a) and density distribution (b) along the line \(y = 0.5\) in the plane \(z = 0.75\) at time \(t = 1\).](image)

In Fig. 5 we present the relative conservation error of fluid mass, computed as the difference between the initial total mass value and the total mass value computed at different time steps. This mass difference
is normalized by the maximum amount of mass swept by the movement of the solid. In Fig. 6 we present the relative energy conservation error of the system, computed as the difference between the initial energy value and the energy value computed at different time steps. This energy difference is normalized by the maximum energy exchange between the fluid and solid, which is the relevant quantity to evaluate the relative effect of coupling on conservation issues.

We observe a small variation of both mass and energy, without any clear growth or decrease of either quantity. The variation of mass is as low as 0.01% of the mass swept and the variation of energy is as low as 0.4% of the energy flux in the system. We suppose that the main effect accounting for these variations are the rounding errors involved in the evaluation of geometric quantities in cut-cells, since both mass and energy are impacted at similar levels.

Figure 5: Relative conservation error on fluid mass.

Figure 6: Relative conservation error on energy.
4.2 Interaction of a shock wave and a sphere

In this problem, a planar shock interacts with a rigid mobile sphere in a rigid channel (Fig. 7). The side boundaries of the domain are rigid walls while the left and right are respectively inflow and outflow boundaries.

The computational domain is the rectangular box \((x, y, z) \in [0, 1] \times [0, 0.2] \times [0, 0.2]\). The shock is initially set up to a Mach number of 3, so that the initial values are

\[
\begin{aligned}
\rho &= 3.857, \quad p = 10.333, \quad u = 2.6929, \quad v = w = 0, \quad \text{if } x < 0.08, \\
\rho &= 1, \quad p = 1, \quad u = v = w = 0, \quad \text{if } x \geq 0.08.
\end{aligned}
\]  

The initial position of the center of mass of the sphere is \(C(0.535, 0.08145, 0.0984)\). The physical system is symmetric with regard to the plane \(z = 0.1\).

The impinging shock wave impacts the sphere and is partially reflected, while part of the shock wave moves past the sphere and part of it is transferred as kinetic energy to the sphere. The reflected shock then reflects itself on the lower wall \((y = 0)\), creating an overpressure under the sphere and lifting it up. Complex interactions between the sphere, the walls and the reflected shocks then occur. In Fig. 8 we display the trajectory of the sphere in the plane \((x, y)\). The final position of the center of mass of the sphere is \(C(0.535, 0.08145, 0.0984)\). The physical system is symmetric with regard to the plane \(z = 0.1\).
This feature is fairly well preserved visually by the numerical results, even though the polyhedron itself is not perfectly symmetric. As a result, the sphere mass center is no longer exactly at $z = 0.1$ at $t = 0.255$. 30 contours of density and pressure at the final time are plotted on Fig. 9a and Fig. 9b, respectively. This computation shows the ability of the coupling algorithm to compute interaction of strong discontinuities with irregular moving boundaries.

Figure 9: 30 contours of density (a) and pressure (b) at time $t = 0.255$.

REFERENCES


A COMPARATIVE STUDY OF IMMERSED-BOUNDARY INTERPOLATION METHODS FOR A FLOW AROUND A STATIONARY CYLINDER AT LOW REYNOLDS NUMBER

SEYED HOSSEIN MADANI*, JAN WISSINK, HAMID BAHAI

School of Engineering and Design - Brunel University West London
*Islamic Azad University – South Tehran Branch
email: Hossein.Madani@brunel.ac.uk

Key words: Immersed-Boundary, Reconstructive method, Interpolation methods, stationary cylinder, Vortex shedding

ABSTRACT

The accuracy and computational efficiency of various interpolation methods for the implementation of non grid-confirming boundaries is assessed. The aim of the research is to select an interpolation method that is both efficient and sufficiently accurate to be used in the simulation of vortex induced vibration of the flow around a deformable cylinder. Results are presented of an immersed boundary implementation in which the velocities near non-confirming boundaries were interpolated in the normal direction to the walls. The flow field is solved on a Cartesian grid using a finite volume method with a staggered variable arrangement. The Strouhal number and Drag coefficient for various cases are reported. The results show a good agreement with the literature. Also, the drag coefficient and Strouhal number results for five different interpolation methods were compared it was shown that for a stationary cylinder at low Reynolds number, the interpolation method could affect the drag coefficient by a maximum 2% and the Strouhal number by maximum of 3%. In addition, the bi-liner interpolation method took about 2% more computational time per vortex shedding cycle in companion to the other methods.

INTRODUCTION

Obtaining accurate solutions for Fluid-Structure Interaction (FSI) problems is of interest in many engineering and scientific applications. A broad classification of FSI methods is based on the type of mesh employed in the discretisation, where we can differentiate between boundary-conforming and non-boundary-conforming mesh methods [1]. A well-known conforming mesh method is the Arbitrarily Lagrangian-Eulerian method (ALE). For non-conforming mesh methods, usually an immersed boundary method is used and most recent developments in FSI methods are based on this approach. The latter is the subject of this review.

The immersed-boundary (IB) method is a technique for solving flow problems in regions with irregular boundaries using a simple structured grid solver. The term “immersed boundary method” was initially used for a method developed by Peskin [2] which was used to simulate blood flow in a cardiovascular system. It was specifically designed to handle deforming (elastic) boundaries interacting with low Reynolds number flow. The simulation was carried
out on a Cartesian grid and at those locations where the boundary did not align with a mesh line the solution algorithm was locally modified to enforce the desired boundary conditions on the flow. More recently, numerous modifications and refinements have been proposed to enhance the accuracy, stability, and application range of the IB method [3].

Depending on the way that the boundary conditions are imposed on the immersed boundary, the IB methods can be generally categorized into continuous and discrete forcing approaches. In the continuous forcing method, a forcing function is applied to the Navier-Stokes equation in order to maintain the boundary condition on the structure (e.g. enforcing a no-slip boundary condition on a stationary body). The most important issue in this method is the definition of the continuous forcing function needed to enforce the correct boundary condition. Several different functions have been developed by Peskin [2], Saiki and Biringen [4], Beyer and Leveque [5], and Lai and Peskin [6], among others. In all cases, a distributed function was used rather than a sharp function because firstly the solid boundaries do not coincide with the Cartesian mesh and, secondly, this way the Gibbs’ oscillations phenomenon [7] adjacent to the solid boundaries could be suppressed. Applying a continuous force to enforce the boundary conditions is attractive for elastic boundaries as it has a physical meaning and its implementation is relatively easy. However, the implementation of this method to enforce rigid boundaries is relatively cumbersome due to its definition. Another problem is that by using a smooth function the method cannot sharply represent the immersed boundaries which is not recommended for high Reynolds number flows [3].

Because the Navier-Stokes equations usually cannot be integrated analytically to find the forcing functions, it is usually not possible to derive an analytical forcing function to enforce specific boundary conditions. To tackle this problem, a method has been suggested by Mohd-Yusof [8] and Verzicco [9]. In this method, which is known as Indirect Discrete forcing approach, forcing functions are subtracted from the numerical solution after discretizing the Navier-Stokes equations. The important advantage of this method is that there is no need to define the forcing function parameters prior to solving the Navier-Stokes equations and there is no stability constraint due to using continuous forcing functions (Gibb’s oscillation). However, it is still needed to implement the distributed forcing functions which strongly depend on the discretization algorithm. Another division of the discrete forcing approach is Direct Discrete Forcing.

Due to the need for an accurate representation of the boundary layer in high Reynolds number flow, the use of distributed, smooth forcing functions near the immersed boundary is not desirable. In these cases it is recommended to use a sharp interface with a higher local accuracy near the boundary. This goal can be achieved by imposing the boundary conditions directly on the immersed boundary. There are two well-known methods that fit into this category: the Ghost-Cell Finite-Difference Approach and the Cut-Cell Finite-Volume Approach.

In the Ghost-Cell approach the immersed boundary is implemented by the use of ghost cells. Ghost cells are cells inside the solid boundary which have at least one neighbour on the fluid side. The parameters (imaginary velocity and pressure) in the ghost cell (inside the solid) are defined by an interpolation method which implicitly enforces the correct boundary condition for the immersed boundary. Iaccarino and Verzicco [10] showed that a linear interpolation method is acceptable for those cases in which the first points of the interpolation
in the fluid are inside the viscous sub layer. Other interpolation methods have been introduced by Ghias [11].

The entire immersed boundary methods discussed so far are not designed to consider the conservation laws near the solid boundary. However, the Cut-Cell method used in combination with a Finite-Volume approach is designed in order to preserve the conservation of momentum and mass near the boundary. In this method the cells, which have been cut by the immersed boundary, are reshaped or absorbed by neighbouring cells in order to form a new trapezoidal control volume cell shape. This method has been used by Mittal [12, 13] to simulate vortex-induced vibration around a stationary and a moving body and for free falling objects. Extending this method to 3D is not straightforward and needs complex polyhedral cells, which complicate the discretization of the Navier-Stokes equations.

As discussed earlier in the discrete forcing approach, the IB is imposed on the domain after the discretization of Navier-Stokes equations. This means that introducing the boundary conditions and forcing functions is not as straightforward as in the continuous forcing approach and depends on the discretization method and its implementation. Also, in discrete forcing approach the definition of the pressure on the boundary is not as straightforward as in the continuous forcing approach and requires special treatment. Advantages of the discrete forcing approach are that the boundary conditions can be introduced sharply without any extra stability constraint, while the fluid and solid domains are clearly separated and the equations that describe the flow are only solved in the fluid domain.

In this paper, we focus on the indirect discrete forcing approach, where the forcing function is not calculated directly and added to the Navier-Stokes equation. We are not intent to use any Cut-Cell or Ghost Cell methods as the applying the Cut-Cell method for fluid-structure interaction problems with moving boundaries takes lots of computational time [14, 15], while the Ghost-Cell approach will create non-physical results when solving the fluid equations in the solid domain.

In the next section, the formulation of the fluid dynamical problem which has been used in the simulation is introduced.

FORMULATION AND NUMERICAL METHODS

The governing equation for an unsteady, incompressible fluid flow in vector form is given by the Navier-Stokes equation which reads:

\[ \rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \mu \nabla^2 \mathbf{v} + \mathbf{f}, \]

\[ \nabla \cdot \mathbf{v} = 0 \]

where \( \mathbf{f} \) is the external force on the fluid domain which is used to implement the boundary condition on non-conforming solid boundaries. In this paper this force is not applied directly to the governing equations. Instead, the non-conforming boundary conditions are introduced by interpolating velocities close to the solid boundary.

The incompressible Navier-stokes equations in a 2D Cartesian domain are given by:

\[ \frac{\partial u_i}{\partial t} + \frac{\partial u_i u_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{1}{Re} \left( \frac{\partial^2 u_i}{\partial x_j \partial x_j} \right) \]

\[ \frac{\partial u_i}{\partial x_i} = 0 \]
where \( p \) is the generalised pressure which is defined by the static pressure divided by the density. Hence, to obtain the correct static pressure we need to multiply \( p \) by the density.

A staggered variable arrangement, as introduced by Harlow and Welsh [16], is used to discretize the governing equations on a Cartesian mesh. The continuity equation is enforced by taking the divergence of the momentum equation and using the continuity equation to simplify the results to form a Poisson equation for the pressure field. This equation is solved by strongly implicit procedure (SIP), Stone’s method, at every time step [17].

To maintain a consistent implementation, the pressure equation is discretized in a similar way as the momentum equation.

The extent of the computational domain was selected to be relatively large to ensure that the location of the boundaries does not affect the simulation. For this reason the size of the \( y \) direction was taken to be 20D and the size of the \( x \) direction was taken to be 15D which was deemed to be sufficient to capture the vortex shedding behind the cylinder. The mesh size was chosen was \( dx=dy=0.05D \) which was checked in a mesh refinement study.

![Figure 1: Fluid domain size and boundary conditions.](image)

As the entire domain was meshed using a Cartesian grid, the implementation of the grid-conforming inlet, outlet and side boundary conditions was straightforward, while the boundary conditions along the cylinder were implemented using immersed boundary methods.

In the next section a number of velocity-interpolation methods, both from the literature as well as a novel second-order interpolation method for the implementation of the non-conforming boundary conditions in the fluid domain are introduced and compared. It will be shown that the results show a good agreement with the literature.
INTERPOLATION METHODS

In this section, interpolation or reconstruction methods are compared. To enforce boundary conditions using the interpolation method, the forcing function $f$ is not calculated directly. Instead, the flow velocity is interpolated at the interface cells and the forcing term is imposed indirectly to the discrete equations. The interface points are defined as the points in the fluid domain near the solid boundary for which one of the neighboring points in the discretized equations is inside the solid domain. Therefore, the parameters related to these points cannot be updated through solving the governing equation (Figure 2). Any cells that contain one or more interface points are called interface cells. It is well known that most immersed boundary approaches need some sort of interpolation procedure. In the direct forcing approaches, interpolation is used in order to determine the forcing functions at the interface cells which enforce the correct boundary conditions to the governing equation. In the indirect forcing approach (interpolations approach), each time step the flow parameters in the interface cells are updated by direct interpolation and used as boundary condition for the flow solver. In this review, a number of interpolation procedures which could potentially be used in indirect discrete forcing approaches (interpolation or reconstruction approaches) will be compared.

![Figure 2: A 2D Cartesian mesh with a solid boundary (circle). Interface points, that require interpolation, are identified by arrows. Points A1 to A8 are all neighboring points of A. Note that A2 and A7 are inside the solid domain.](image)

Below it will be shown in detail how these interface cells have been treated and how possible problems that may occur near boundaries, like the decoupling of pressure and conservation of mass, may be overcome. To do so, first step is to define the interface cells for the specific geometry, which could be complicated for geometries with unknown functions [10]. Subsequently, it will be ensured that the flow governing equations are not solved inside any interface and solid cells. The most important step in the interpolation methods is to determine the flow parameters in the interface cells adjacent to the solid boundaries which will be used as boundary conditions for the rest of the flow domain that will subsequently be updated by the flow solver. Various interpolation methods have been developed to tackle this problem. In the following part, these methods are categorized and explained in more detail.

**Case A: No interpolation**

The simplest possible method is to select the interface cells at the solid boundaries and define the solid domain inside those cells. In fact, in this case there is no interpolation and the
solid boundary will have a stepwise shape (Figure 3). Also, the boundary itself is somewhat diffused, as in the staggered methods the boundary conditions for the different velocity components are applied at different sides of an element. Fadlun [18] proposed a similar method for imposing forcing functions for immersed boundaries. Here, however, the method is applied directly to define the solid boundaries, while the governing equation will be subsequently solved in the remainder of the computational domain assuming no-slip boundary conditions for the solid boundaries. As interpolation is not needed, this method will be relatively fast while still giving acceptable results. The disadvantage of this method - when used in the calculation of flow around a circular cylinder - is that on course meshes shape and size differences between the cylinder and the solid boundary could affect lift and drag forces.

Case B: Weighting method

This method is similar to the one discussed above. The major difference is that the boundary values for the velocity in those cells that are part fluid and part solid are weighted accordingly. Figure 3 (right) shows the location of these weighted boundary velocities in the cells that are part fluid and part solid. For each of the velocity components a coefficient is determined that corresponds to the ratio of the fluid part of the two adjacent cells to the whole area of the two cells.

Figure 3: left, the arrows shows velocity inside solid body (solid velocity), assumed zero for a stationary cylinder. Right, hatched cells used to define weighting coefficient.

Figure 4: linear interpolation method: interpolating $V_{i,j}$ for a stationary solid in vertical direction (left), interpolating $V_{i,j}$ for a moving solid ($V_{solid}$) in vertical direction (middle); interpolating $U_{ij}$ for a moving solid ($U_{solid}$) in horizontal direction.
Case C: linear interpolation method

The second method is a linear interpolation method where the velocities in the interface cells are calculated by interpolation between the velocity at the solid boundary (applying the no slip condition) and one point in fluid domain. Fadlun [18] suggested using this interpolation method to enforce the boundary condition to the fluid domain in the indirect discrete forcing approaches. In this paper, however, we are only interested in his interpolation method.

Case D: Bilinear interpolation method

Kang [19] presented various methods to interpolate the velocity near the boundary in two directions considering the effect of the pressure near the boundary. As before, only the interpolation is of interest here. At first, in the Standard Reconstruction method (SRM) the two velocities in the horizontal and vertical directions that are closest to the immersed boundary are used to obtain the interpolated velocity at each of the interface points. The resulting interpolation formula for the velocity in horizontal direction has the following form.

$$U_{i,j} = \omega_{i+1,j}U_{i+1,j} + \omega_{i,j+1}U_{i,j+1} + \omega_{\text{solid}}U_{\text{solid}}$$

where the various $\omega$ represent the interpolation weights.

Kang [19] has revised the above interpolation and also uses the velocity field in the previous time step to obtain a better interpolation at the next time step and explicitly used the difference between the velocities at two consecutive time steps in the interpolation of the interface velocities. In this case the interpolation formula becomes:

$$\Delta U_{i,j} = \omega_{i+1,j} \Delta U_{i+1,j} + \omega_{i,j+1} \Delta U_{i,j+1} + \omega_{\text{solid}} \Delta U_{\text{solid}}$$

In addition Kang [19] introduced a quadratic interpolation formula to incorporate the local pressure gradient in the velocity interpolation to compensate for a decoupling between the
pressure and the velocity near the solid boundary. Figure 6 illustrates their interpolation method that uses four adjacent velocities and enforces the momentum equation by a quadratic interpolation in the two-dimensional case. As we only focus on comparing velocity interpolation methods Kang’s method is excluded from the comparison.

**Case E: Proposed interpolation method**

The bilinear interpolation method proposed in this paper is based on interpolating the boundary velocity values in the direction perpendicular to the solid boundary. In this method, a line from the boundary velocity position is drawn perpendicular to the boundary surface and extended to cut the line between the first two known velocities in the fluid domain (Point A, Figure 7 right). The velocity will be interpolated at the intersection point A. Then, the boundary cell velocity values will be interpolated using the solid boundary velocity (for a stationary cylinder with no-slip conditions this velocity is zero) and the velocity at point A. Figure 7 (left) shows this interpolation for velocities in y direction and Figure 7 (right) shows the interpolation for the velocity in the x direction.

**RESULTS AND DISCUSSION**

The flow around a stationary cylinder at Re=100 has been simulated with different interpolation treatments to represent the immersed boundary. The Strouhal number, drag and lift coefficients for various cases are compared.

The Strouhal number is the non-dimensional frequency of the vortex shedding around the body and is defined by:

\[ St = \frac{f_t D}{U_{in}} \]

where \( f_t \) is the frequency of the vortex shedding, \( D \) is the cylinder diameter and \( U_{in} \) is the far-field velocity.

The drag coefficient on a body in a fluid flow includes both the shear stress and the pressure drag on the solid surface. The dimensionless drag coefficient is defined by:

\[ C_D = \frac{F_D}{\frac{1}{2} \rho u_c^2 D} \]

The lift force on the cylinder is generated when the vortex shedding starts around the structure. The dimensionless lift coefficient is defined by:

\[ C_L = \frac{F_L}{\frac{1}{2} \rho u_c^2 D} \]
For any solid body both the pressure distribution and the friction along the solid surface may contribute to the lift and drag forces. In the present study, the pressure at the surface is obtained by taking the wall-nearest pressure values in the flow domain on the outside of the solid body, thereby assuming that the wall normal gradient of the pressure near the surface is negligibly small. The component of the drag and lift forces due to pressure distribution is calculated by integrating the pressure along the solid boundary. On the other hand, the shear-force component of the lift and drag forces is calculated from near the surface of the solid. The tangential velocity near the solid surface is obtained at the wall-nearest point outside of the body and is subsequently used to calculate the wall-shear stress at the cylinder surface.

Table 1: Real computational time, 20 vortex shedding

<table>
<thead>
<tr>
<th>Case</th>
<th>Real time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case A</td>
<td>3231</td>
</tr>
<tr>
<td>Case B</td>
<td>3225</td>
</tr>
<tr>
<td>Case C</td>
<td>3379</td>
</tr>
<tr>
<td>Case D</td>
<td>4441</td>
</tr>
<tr>
<td>Case E</td>
<td>3383</td>
</tr>
</tbody>
</table>

Figure 8 shows a comparison of the drag coefficients obtained in calculations of flow over a stationary cylinder at Re=100 using various interpolation methods. It can be seen that in the cases C, D and E, (linear and Bilinear interpolation methods) the results were converging to a value of C_D = 1.43. However, Case A (without interpolation) leads to a higher drag coefficient, C_D=1.46 and Case B (weighting method) leads to a lower drag coefficient C_D=1.42. Once vortex shedding commenced all simulations were found to run at virtually the same speed (Table 1) showing that the computational effort needed for the interpolation was negligible. However, for a non-stationary cylinder, it is expected that the required repeated calculation of interpolation coefficients may lead to a reduction in execution speed.

Figure 8 (left) shows that Case C (linear interpolation) is the quickest method to develop vortex shedding, which indicates that it the implementation of boundary conditions with linear interpolation causes significant numerical noise. In Case E (proposed bilinear method),
on the other hand, the vortex-shedding instability kicks in much later evidencing that the level of numerical noise introduced by this type of interpolation is very small.

Figure 9: Lift coefficient for the flow around a stationary cylinder at Re=100, Case A, without interpolation; Case B: area weighting method; Case C, Linear interpolation method; Case D, Bilinear interpolation1; Case E, Bilinear interpolation2

Figure 9 shows the comparison of lift coefficients for the various interpolation cases. It can be seen from the figure that like the drag coefficient, the lift coefficient for the linear and bilinear cases are nearly the same (Case C, D and E) CL=0.33. However, the Case B (related to weighting area method) shows lower values for the lift (CL=0.27). In Case A (without interpolation), the lift due to shear stress is out of range, but the lift due to pressure is acceptable. The reason for the unacceptable results for the lift due to shear is that the velocity for case A was selected out of the boundary layer (the aim was to choose similar conditions for all cases).

Table 2: Strouhal number, lift and Drag coefficient for the flow around a stationary cylinder and Re=100.

<table>
<thead>
<tr>
<th>simulation methods</th>
<th>Strouhal Number</th>
<th>Drag Coefficient</th>
<th>Lift coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case A</td>
<td>0.174</td>
<td>1.46</td>
<td>0.26</td>
</tr>
<tr>
<td>Case B</td>
<td>0.175</td>
<td>1.42</td>
<td>0.27</td>
</tr>
<tr>
<td>Case C</td>
<td>0.169</td>
<td>1.432</td>
<td>0.325</td>
</tr>
<tr>
<td>Case D</td>
<td>0.169</td>
<td>1.434</td>
<td>0.305</td>
</tr>
<tr>
<td>Case E</td>
<td>0.168</td>
<td>1.432</td>
<td>0.31</td>
</tr>
<tr>
<td>Park [22] fitted method</td>
<td>0.165</td>
<td>1.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Williamson(exp.)[20]</td>
<td>0.166</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>Kim [22]</td>
<td>0.165</td>
<td>1.33</td>
<td>0.32</td>
</tr>
<tr>
<td>Roshko (exp.)[20]</td>
<td>0.164</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>Lai and Peskin [6]</td>
<td>0.165</td>
<td>1.4473</td>
<td>0.3299</td>
</tr>
<tr>
<td>Choi [7]</td>
<td>....</td>
<td>1.351</td>
<td>0.315</td>
</tr>
<tr>
<td>Corbalan &amp; Souza [24]</td>
<td>....</td>
<td>1.44</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Table 2 shows the comparison of the Strouhal number, lift and drag coefficient for various methods; from the experimental methods (Roshko and Williams reported by [20]) to the body fitted mesh [21] and immersed boundary methods [22, 6, 23, 24], for the flow around a stationary cylinder at Re=100. It can be seen that the Strouhal number varies between 0.16 and 0.18; the Drag coefficient between 1.33 and 1.4473 and the lift coefficient between 0.31 and 0.33.

CONCLUSION

The objective of the present study is to compare the accuracy and expenses of different IB interpolation methods and select the most accurate and least expensive method for future use in simulations of flow around a deformable cylinder. A finite-volume method on a Cartesian grid with a staggered variable arrangement has been used. In this IB implementation the velocities near non-confirming boundaries were interpolated in the normal direction to walls, thereby considering the curvature of the geometry. The Strouhal number and Drag coefficient for different cases are reported. The results show a good agreement with the literature for most of the interpolation methods for the stationary cylinder. The drag coefficient and Strouhal number results for five different interpolation methods were compared it was shown that for a stationary cylinder at low Reynolds number, the interpolation method could affect the drag coefficient by a maximum 2% and the Strouhal number by maximum of 3%. In addition, the bi-liner interpolation method took about 2% more computational time per vortex shedding cycle in companion to the other methods.

REFERENCES

A COUPLED MODE – hp FEM FOR HYDROELASTIC ANALYSIS OF SHEAR-DEFORMABLE FLOATING BODIES OF GENERAL THICKNESS IN VARIABLE BATHYMETRY

K. A. BELIBASSAKIS†, G. A. ATHANASSOULIS†,

T. K. PAPATHANASIOU* AND S. FILOPOULOS*

† School of Naval Architecture and Marine Engineering
National Technical University of Athens
Herou Polytechniou 9, Zografos 15773, Athens, Greece
e-mail: kbel@fluid.mech.ntua.gr

* School of Applied Mathematical and Physical Science
National Technical University of Athens
e-mail: paphth@gmail.com

Key words: Hydroelastic analysis, Coupled Modes, FEM, Shear Deformable Plates, VLFS.

Abstract. An efficient computational procedure is presented for the solution of coupled hydroelastic problems involving bodies of general thickness, floating over variable bathymetry regions. The problem is treated by the coupled mode system of horizontal equations derived by Athanassoulis and Belibassakis [1], for the analysis of floating, shear deformable plates or beams. The proposed beam (or plate) model is based on the addition of extra vertical elastic deformation modes, at each horizontal position along the floating body, permitting shear strain and stress to vanish on both the upper and lower boundaries and extending third-order plate theories [2]. The final coupled mode system is derived from a variational principle combining the one – field functional of the elastodynamics in the plate region with the pressure functional in the water region. The wave potential in the water column is represented by means of a local – mode series containing an extra mode, accounting for not mildly sloped bottom variations [3]. The addition of the additional modes results to increased convergence rate, enabling high accuracy with the use of a relatively small number of vertical modes. In the present work the hp-version of the Finite Element Method [4] is applied to the solution of a simplified version of the resulting system of coupled horizontal differential equations with respect to the modal amplitudes, providing good convergence rates and adaptivity capabilities, and increasing the overall efficiency of the solution strategy. Numerical results are presented demonstrating the applicability of present method.

1 INTRODUCTION

The effect of water waves on floating deformable bodies is related to both environmental and technical issues, finding important applications. A specific example concerns the interaction of waves with thin sheets of sea ice, which is particularly important in the Marginal Ice Zone.
(MIZ) in the Antarctic, a region consisting of loose or packed ice floes situated between the ocean and the shore sea ice [5]. As the ice sheets support flexural–gravity waves, the energy carried by the ocean waves is capable of propagating far into the MIZ, contributing to break and melting of ice glaciers [6,7] thus accelerating global warming effects and rise in sea water level. In addition, the interaction of free-surface gravity waves with floating deformable bodies is a very interesting problem finding applications in hydrodynamic analysis and design of very large floating structures (VLFS) operating offshore (as power stations/mining and storage/transfer), but also in coastal areas (as floating airports, floating docks, residence/entertainment facilities), as well as floating bridges, floating marinas and breakwaters etc. For all the above problems hydroelastic effects are significant and should be properly taken into account. Extended surveys, including a literature review, have been presented by Kashiwagi [8], Watanabe et al [9]. A recent review on both topics and the synergies between VLFS hydroelasticity and sea ice research can be found in Squire [10].

Taking into account that the horizontal dimensions of the large floating body are much greater than the vertical one, thin-plate (Kirchhoff) theory is commonly used to model the above hydroelastic problems. Although non-linear effects are of specific importance, still the solution of the linearised problem provides valuable information, serving also as the basis for the development of weakly non-linear models. The linearised hydroelastic problem is effectively treated in the frequency domain, and many methods have been developed for its solution. These include hydroelastic eigenfunction expansion techniques [11,12,13], Boundary Element Methods [14,15], B-spline Galerkin method [16], integro-differential equations [17], Wiener-Hopf techniques [18], Green-Naghdi models [19], and others. Moreover, Meylan [20] derived a variational equation for the plate-water system by expressing the water motion as an operator equation. In the case of hydroelastic behaviour of large floating bodies in general bathymetry, a new coupled-mode system has been derived and examined by Belibassakis & Athanassoulis [3] based on local vertical expansion of the wave potential in terms of hydroelastic eigenmodes, and extending a previous similar approach for the propagation of water waves in variable bathymetry regions [21]. Similar approaches with application to wave scattering by ice sheets of varying thickness have been presented by Porter & Porter [5] based on mild-slope approximation and by Bennets et al [22] based on multi-mode expansion.

In the above models the floating body has been considered to be very thin and first-order plate theory has been applied, neglecting shear effects. In Athanassoulis & Belibassakis [1] an extension of the previous coupled-mode model is presented with application to the case of hydroelastic analysis of a thick, non-uniform, shear deformable floating elastic body, lying over variable bathymetry regions. The problem addressed therein is the scattering of linear, coupled, hydroelastic waves propagating through an inhomogeneous sea-ice environment, containing ice sheets of variable, finite thickness characterized by mildly sloped interface. The enhanced coupled-mode system of horizontal equations is obtained on the basis of higher-order theory of shear deformable plates and beams, and is derived by the local vertical expansion of the wave potential in the water region, in conjunction with an enhanced representation of the elastic displacement field in the plate, containing additional elastic vertical modes. The latter permit the shear strain and stress to vanish on both the upper and lower boundaries of the finite floating plate, and extends third-order plate theories by Reddy.
[23] and Bickford [24] (see also [2]) to plates and beams of general shape. The above representations are used in a variational principle composed by the one-field functional of elastodynamics in the plate region, and Luke [25] pressure functional in the water region, reducing to the coupled-mode system (CMS). In the case of plates of general but slowly varying and relatively small thickness, the above shear-enhanced CMS simplifies in a form suitable for long-range, large-domain calculations, and extends the first-order model by Belibassakis & Athanassoulis [3] to shear deformable floating plates of variable finite thickness, lying over general bottom topography. In the present work, an efficient computational procedure is presented, to treat the extended coupled-mode system, using the hp-version of the Finite Element Method [4] for the solution of the resulting system of horizontal differential equations, providing very good convergence rates and adaptivity capabilities, and increasing the overall efficiency of the solution strategy. Numerical results are presented demonstrating the applicability of the proposed method.

2. FORMULATION OF THE PROBLEM

The studied environment consists of a water layer bounded above by a floating plate of general shape, as e.g., an ice sheet of variable thickness \( b(x) \), and below by a rigid bottom; see Fig. 1. For simplicity we restrict ourselves to a 2D problem, however the method can be straightforward extended to 3D. Also, the upper face of the floating elastic plate is flat-horizontal (e.g. by considering that the static plate deflection to be very small), however, our analysis can be easily extended to the case of a more general upper boundary. It is also assumed that the bottom and the plate surfaces exhibit general one-dimensional variation in a subdomain of finite length. The bathymetry is characterised by straight and parallel bottom contours lying between two regions of constant but possibly different depth: \( h = h_i \) (region of incidence) and \( h = h_j \) (region of transmission), where \( h(x) \) is the depth function. A Cartesian coordinate system is introduced, with its origin at some point on the upper elastic-plate surface (in the variable bathymetry region), the \( z \)-axis pointing upwards and the \( y \)-axis being parallel to the bottom contours. The functions \( h(x) \) and \( b(x) \) are smooth functions such that \( h(x) = h(x_i) = h_i \) and \( b(x) = b(x_i) = b_i \), for all \( x \leq x_i \), and \( h(x) = h(x_j) = h_j \) and \( b(x) = b(x_j) = b_j \), for all \( x \geq x_j \). Also, the slope of the elastic-plate deflection \( w(x,y;t) \) assumes small, so that linear theory can be applied. We consider the scattering problem of harmonic incident plane waves of angular frequency \( \omega \), under the combined effects of variable bathymetry and the infinite, floating elastic plate of general and finite thickness \( b \).

We shall concentrate here in the case of normally incident linear waves (as shown in Fig.1), leaving the treatment of obliquely-incident waves propagating with directions \( \theta_i \) and \( \theta_t \) with respect to the \( x \)-axis in the regions of incidence \(( x \leq x_i )\) and transmission \(( x \geq x_j )\), respectively, and more complex 3D systems to be examined in future works. For the above problem an extended coupled-mode model (eCMS) has been derived by Athanassoulis & Belibassakis [1], for the hydroelastic analysis of floating, shear deformable plate or ice sheet in general bathymetry, characterised by sloped boundaries and interfaces, taking into account finite, general thickness effects.
The extended model is based on enhanced representations of both the elastic displacement field in the plate region and the wave potential field in the water region, permitting to model effects of shear stresses and to consistently treat the end-conditions on the non-horizontal boundaries (wetted plate surface, bottom surface).

2.1 Modal expansion of the wave potential in the water subregion

In a series of works presented by the authors, starting with the linearised water wave problem in general bathymetry (Athanassoulis & Belibassakis [22]), a vertical local mode series expansion is used to consistently represent the wave field in the water region:

$$\phi(x,z) = \phi_{-1}(x)Z_{-1}(z;x) + \sum_{n=0}^{\infty} \phi_{n}(x)Z_{n}(z;x), \quad -h(x) \leq z \leq -b(x).$$

(1)

The major part of the set of vertical modes \(\{Z_{n}(z;x), n = 0,1,2,...\}\) is obtained through the solution of a vertical eigenvalue problem, formulated at each horizontal position and \(\phi_{-1}(x)Z_{-1}(z;x)\) is an appropriate term, called the sloping-bottom mode, accounting for the satisfaction of the bottom boundary condition on the non-horizontal parts of the bottom. The idea of the sloping bottom mode, in conjunction with the above type of modal expansion, has been first introduced by the authors (Athanassoulis & Belibassakis [22]) in the case of water waves propagating in variable bathymetry. Since then, it has been used for many problems exhibiting similar features, such as nonlinear water waves (Belibassakis & Athanassoulis [26]), hydroacoustics (Athanassoulis et al [27]), and hydroelastic applications in variable bathymetry regions, formulated in the context of classical thin plate theory (Belibassakis & Athanassoulis [3]) and high-order shear deformable plate theory (Athanassoulis & Belibassakis [1]). In accordance with the latter work, the infinite set \(Z_{n}(z;x), n = 0,1,2,3,...,\)
of functions describing the vertical structure of each mode, at each horizontal position \( x \), are generated by

\[
\frac{\partial^2}{\partial z^2} Z_n(z) - \kappa_n^2 Z_n(z) = 0, \quad \text{in the vertical interval } -h < z < -b, \quad (2a)
\]
\[
\dot{Z}_n(z = -h) = 0, \quad \text{at the bottom } z = -h(x), \quad (2b)
\]
\[
\alpha \dot{Z}_n(z = -b) - \mu Z_n(z = -b) = 0, \quad \text{at the water-elastic body interface } z = -b(x), \quad (2c)
\]

where \( \alpha \) is a function depended on \( \kappa \). The solution the above problem is given by

\[
Z_n(z) = \cosh^{-1}(\kappa_n H) \cosh[\kappa_n(z + h)], \quad n = 0, 1, 2, 3, \ldots , \quad (3)
\]

where the eigenvalues \( \{\kappa_n, n = 0, 1, 2, \ldots \} \) are obtained as the roots of (local) dispersion relation:

\[
\mu H = \alpha(\kappa) \kappa H \tanh(\kappa H), \quad \text{where } \alpha(\kappa) = D e_\nu \kappa^4(1-\delta) + 1 - e^{\left(1 + \frac{\kappa^2 h^2}{12}(1-\delta)\right)}. \quad (4)
\]

In Eq. (4), \( \mu = \omega^2 / g \) is the frequency parameter, \( H = h - b \) is the thickness of the water layer, \( D = Eb^3 / (12\rho g\left(1 - \nu^2\right)) \) denotes the plate flexural rigidity (with \( E \) Young’s modulus and \( \nu \) Poisson’s ratio). Moreover, \( \rho \) is the water density and \( g \) acceleration of gravity, \( \varepsilon = m\omega^2 / \rho g \) the plate mass coefficient (with \( m \) the plate mass distribution per horizontal area). Parameter \( e_\nu = \left(1 + \nu^2(1-2\nu)^{-1}\right) \) is a material constant, involved in the expression of \( \sigma_\alpha = E e_\nu (1-\nu^2)^{-1} e_\alpha \).

Additionally, \( Xb^{-2} \) and \( \delta \) are newly introduced non-dimensional parameters given as follows:

\[
\delta = 0.0135\left(0.0135 + \left(X / b^2\right)^{-1}\right), \quad \frac{X}{b^2} = \left(2\kappa^2 b^2 e_\nu\right)^{-3} \cdot 0.132(1-\nu) \end{equation}, \quad (5)
\]

\[
\text{and } d = D e_\nu H^{-4}. \quad \text{More details can be found in Athanassoulis & Belibassakis[1]. We note here that the above system (Eqs.3,4) with } \alpha(\kappa) \to (D\kappa^4 + 1-\varepsilon), \text{ which is asymptotically obtained for small plate thickness } (\kappa b \to 0, \delta \to 0), \text{ reduces exactly to the standard hydroelastic relations based on thin-plate theory (Athanassoulis & Belibassakis [3])}. \]

On the basis of the above complete expansions of the wave potentials in the two semi-infinite strips \( (x \leq x_c) \) and \( (x \geq x_c) \) are obtained (see also Belibassakis & Athanassoulis [1]), describing plane hydroelastic waves in these regions. Demonstrative results concerning the phase speed of propagating hydroelastic waves in homogeneous floating shear deformable plate of constant thickness, non-dimensionalised with respect to the phase speed of linearised water waves in shallow conditions \( C / \sqrt{gH} \), are shown in Fig.2. We have used a density ratio \( \rho e_\nu / \rho = 923/1025 \) corresponding to ice/water. In this case the plate has uniform finite thickness \( b/h=0.5, \) and Poisson ratio \( \nu=0.3. \) Results are presented for three representative values of flexural rigidity \( d=1, 10, 100 \) and three values of \( \varepsilon=0, 0.5, 1, \) for shoaling ratio \( \kappa H \) ranging from very shallow to very deep water conditions \( 0 < \kappa H < 16 \).
Figure 2. Phase speed of harmonic flexural waves for shear deformable plate of uniform finite thickness \( b/h = 0.5 \), and Poisson ratio \( \nu = 0.3 \), for various values of \( \varepsilon \), \( d \) and shoaling ratio \( (kH) \).

We clearly observe that the effect of flexural rigidity that leads to significant increase of phase speed. Also, in the limit of very shallow conditions \( kH \ll 1 \), we observe in this figure that the effect of mass parameter \( \varepsilon \) becomes important.

2.2 The extended coupled-mode system (eCMS)

The eCMS of horizontal differential equations has been obtained by means of a variational principle composed by the one-field functional of elastodynamics in the plate region (see, e.g. [28]), and the Luke's [25] pressure functional in the water region. The wave potential in the water column is represented by means of the local mode series expansion Eq. (1), and an enhanced fourth-order vertical expansion of the elastic displacement field in the floating plate is used containing additional elastic vertical modes, permitting the shear strain and stress to vanish on both the upper and lower boundaries of the thick floating plate. More details can be found in Athanassoulis & Belibassakis [1]. In the case of plates of general, finite thickness, but slowly varying characteristics, elimination relations are approximately derived between the vertical plate deflection \( (w) \) and the rest of the elastic displacement modes. In this case the eCMS takes the following form

\[
\sum_{n=-1}^{\infty} a_{mn}(x) \frac{\partial^2 \varphi_n(x)}{\partial x^2} + b_{mn}(x) \frac{\partial \varphi_n(x)}{\partial x} + c_{mn}(x) \varphi_n(x) = i\omega w(x), \quad m = -1, 0, 1, \ldots ,
\]  

(6)

in conjunction with the following equation providing the coupling between the water-wave modes \( (\varphi_n) \) and the elastic plate deflection \( (w) \):

\[
\frac{\partial^2}{\partial x^2} \left( dH^4 \left( 1 - \delta \right) \frac{\partial^2 w}{\partial x^2} \right) - \varepsilon b^2 \frac{\partial^2 w}{12 \partial x^2} + (1 - \varepsilon) \omega = \int \frac{\varphi_n(x)}{\omega} \sum_{n=0}^{\infty} \varphi_n(x).
\]  

(7)
In the above equations, the $x$-dependent coefficients $a_{mn}(x)$, $b_{mn}(x)$ and $c_{mn}(x)$ are given by the following expressions

$$a_{mn}(x) = \langle Z_n, Z_m \rangle_{-h}^{-b},$$

$$b_{mn}(x) = 2 \left( \frac{\partial Z_n}{\partial x} \right)_{-h}^{b} + \frac{dh}{dx} Z_n(z = -h; x) - \frac{db}{dx} Z_m(z = -b; x),$$

$$c_{mn}(x) = \left( \frac{\partial^2 Z_n}{\partial x^2} + \frac{\partial^2 Z_n}{\partial z^2} - q^2 Z_n, Z_m \right)_{-h}^{b} - \left( \frac{\partial Z_n(z = -b; x)}{\partial z} + \frac{db}{dx} \frac{\partial Z_n}{\partial x}(z = -b; x) \right) Z_m(z = -b; x) + \frac{dh}{dx} \frac{\partial Z_n}{\partial x}(z = -h; x),$$

where $\langle f, g \rangle_{-h}^{b} = \int_{z=-h(x)}^{z=-b(x)} f(z) g(z) dz$. After solving the system of Eqs. (6), (7) the wave characteristics can be obtained all over the domain by means of the calculated wave modes $\varphi_n(x), n = -1, 0, 1, 2, 3, \ldots$, using the expansion (1). Also, the elastic strain and stress distributions in the thick plate cross section are obtained from the solution, using expressions connecting the vertical deflection ($w$) and the rest of the shear deformable plate modes (see Athanassoulis & Belibassakis [1]).

The eCMS is supplemented by specific boundary conditions ensuring complete matching between the wave and the elastic fields at the two vertical interfaces (at $x = x_1$ and $x = x_3$) separating the variable bathymetry inhomogeneous subdomain from the regions of incidence ($x \leq x_1$) and transmission ($x \geq x_3$), respectively. More specifically at the left interface

$$\varphi_0'(x_1) + i \kappa_0^{(1)} \varphi_0(x_1) = 2i \kappa_0^{(1)} \exp(ik_0^{(1)} x_1), \quad \varphi_n'(x_1) + i \kappa_n^{(1)} \varphi_n(x_1) = 0, \quad n=1,2,3\ldots, (9a)$$

$$\varphi_n'(x_1) + \left( \kappa_n^{(1)} \right)^2 \varphi_n(x_1) = 0, \quad \text{at} \ x = x_1, \quad \text{and}$$

$$w(x_1) = \frac{i}{\omega} \sum_{n=0}^{\infty} \kappa_n^{(1)} \tanh(\kappa_n^{(1)} H_1) \varphi_n(x_1), (9c)$$

where a prime denotes $x$-differentiation. Moreover on the right vertical interface

$$\varphi_n'(x_3) - i \kappa_n^{(3)} \varphi_n(x_3) = 0, \quad n=0,1,2,3\ldots, \quad \text{and} \quad \varphi_n'(x_3) + \left( \kappa_n^{(3)} \right)^2 \varphi_n(x_3) = 0, \quad \text{at} \ x = x_3, (10a,b)$$

$$w(x_3) = \frac{i}{\omega} \sum_{n=0}^{\infty} \kappa_n^{(3)} \tanh(\kappa_n^{(3)} H_3) \varphi_n(x_3). (10c)$$

In the above equations $\kappa_n^{(m)}, m = 1, 3$ denote eigenvalues obtained from the extended hydroelastic dispersion relation (4) formulated at the constant depth and plate thickness.
subregions, respectively. The forcing of the eCMS appears only Eq. (9a) and is associated with the mode representing the incident wave exciting the hydroelastic waveguide.

2.3 Reformulation as a second-order system

The discrete version of the present CMS is obtained by truncating the local-mode series (1) to \( n = N \) keeping the first \( N + 2 \) modes \( \{ \phi_n(x), n = -1, 0, 1, ..., N \} \). Subsequently, by setting \( \phi_{N+1}(x) = w(x) \) and \( \phi_{N+2}(x) = dH^4(1-\delta)w(x)^* \), Eq. (7) is equivalently written as follows

\[
dH^4(1-\delta)\phi_{N+1}^*(x) - \phi_{N+2}(x) = 0, \quad (11a)
\]

\[
\phi_{N+2}(x) + \frac{\epsilon b^2}{12dH^4} \phi_{N+2}(x) + (1-\epsilon)\phi_{N+1}(x) - \frac{i\mu}{\omega} \sum_{n=1}^{N} \phi_n(x) = 0. \quad (11b)
\]

Thus, the present eCMS is put in the following, second-order form

\[
\sum_{n=1}^{N+2} \left( a_{mn}(x)\phi_n^* + b_{mn}(x)\phi'_n + c_{mn}(x)\phi_n \right) = 0, \quad m = -1, 0, 1, ..., N, N+1, N+2. \quad (12)
\]

where the definition of general \( a_{mn}, b_{mn}, c_{mn} \) coefficients is obtained from Eqs. (8) and (11).

3. CMS SOLUTION WITH hp-ADAPTIVE FINITE ELEMENTS

Assuming that the matrix \( \tilde{a}_{nn} \) is invertible, the coupled mode system is written in the form:

\[
-\mathbf{u}'' + \mathbf{B}\mathbf{u}' + \mathbf{C}\mathbf{u} = 0, \quad x \in (x_1, x_3) \quad \text{where} \quad \mathbf{B} = -\tilde{a}^{-2}\tilde{b} \quad \text{and} \quad \mathbf{C} = -\tilde{a}^{-2}\tilde{c}, \quad (13)
\]

where \( \mathbf{u}(x) \) is the vector of unknown modal amplitudes of dimension \( M = N + 4 \). The system is accompanied by the boundary conditions

\[
\mathbf{u}' + \mathbf{S}_1\mathbf{u} = \mathbf{T}_1, \quad \text{at} \quad x = x_1, \quad \text{and} \quad \mathbf{u}' + \mathbf{S}_3\mathbf{u} = \mathbf{T}_3, \quad \text{at} \quad x = x_3, \quad (14)
\]

where the coefficients \( \mathbf{S}_n, n = 1, 3, \) and the forcing \( \mathbf{T}_i \) are obtained from Eqs.(9) and (10). The weak formulation of boundary value problem (13),(14) consists of finding \( \mathbf{u} \in V \), where \( V \)

is the Cartesian product of Sobolev spaces \( V = [H^1(x_1, x_3)]^M \) defined over \( \mathbb{C} \), such that

\[
r(\mathbf{w}, \mathbf{u}) = \mathbf{w}^*(x_3)\mathbf{T}_3 - \mathbf{w}^*(x_1)\mathbf{T}_1 = F(\mathbf{w}), \quad \forall \mathbf{w} \in V, \quad (15a)
\]

where \( \mathbf{w}^* \) denotes the complex conjugate of \( \mathbf{w} \in V \) and \( r(\cdot, \cdot): V \times V \to \mathbb{C} \) is the following continuous sesquilinear form

\[
r(\mathbf{w}, \mathbf{u}) = \int_{x_1}^{x_3} \mathbf{w}' \mathbf{u}' dx + \int_{x_1}^{x_3} \mathbf{w}^* \mathbf{B} \mathbf{u}' dx + \int_{x_1}^{x_3} \mathbf{w}^* \mathbf{C} \mathbf{u} dx + \mathbf{w}^*(x_3)\mathbf{S}_3\mathbf{u}(x_3) - \mathbf{w}^*(x_1)\mathbf{S}_1\mathbf{u}(x_1). \quad (15b)
\]

Assuming that \( r(\cdot, \cdot) \) is \( V \)-elliptic (i.e. \( \exists \mu \in \mathbb{R}^+ \) such that \( r(\mathbf{u}, \mathbf{u}) \geq \mu \|\mathbf{u}\|^2 \), with ellipticity constant \( \mu \) ) direct application of the Lax – Milgram lemma guarantees the existence of a
unique \( u \in V \), solution to variational problem (15), \( \forall F \in V' \) (the dual of \( V \)). Further, we have the a priori estimate

\[
\|u\|_{V'} \leq \frac{1}{\mu} \|F\|_{V'},
\]

Finite element approximation of solutions of the variational problem are constructed by considering the subspaces \( V^h \subset V \) and finding \( u^h \in V^h \) such that \( r(u^h, u) = 0 \) for all \( u \in V^h \). For the implementation of the Finite Element Method, we assume a partition of the interval \([x_1, x_N]\) of the form \( x = s_1 < s_2 < \ldots < s_{N+1} = x_N \), \( N \in \mathbb{N} \).

Let \( P_p(s) \) be a polynomial of degree \( p \). We now define the subspaces \( V^h \subset V \) as

\[
V^h = \left\{ u^h \in \left[ H^1(x_1, x_N) \right]^M : u^h|_{[s_i, s_{i+1}]} \equiv P_p(s) \right\}, i = 1, 2, \ldots, N, j = 1, 2, \ldots, M, p \in \mathbb{N}\}
\]

From standard theory (e.g., Babuška & Ihlenburg [29]) we expect the following estimate of the error associated with the present approximate solution

\[
\|u - u^h\|_{V'} \leq ch^p \|u\|_{H^1(x_1, x_N)},
\]

holding for some positive constant \( c \) that does not depend on \( h \) or \( p \).

## 4 NUMERICAL EXAMPLES AND DISCUSSION

In order to illustrate the applicability of our method, a specific example is presented in Figs. 3-5 concerning the propagation of harmonic hydroelastic waves of period \( T=10 \text{sec} \) \((\omega=0.628 \text{rad/sec})\) on floating ice sheet, characterised by modulus of elasticity \( E=5 \text{GPa} \), Poisson’s ratio \( \nu=0.3 \) and values of ice/water densities \( 923/1025 \text{kg/m}^3 \) (see also Squire et al. [6]). In this case, except of uniform ice sheet of finite thickness \( b=1 \text{m} \), in semi-infinite strips of constant depths \( h=13 \text{m} \) and \( h=7 \text{m} \), respectively, we also consider the effect of inhomogeneous ice thickness in \( 100 \text{m}<x<400 \text{m} \), with specific form as shown in Fig.3.

In this case the beam thickness varies from 1m at the ends of the domain to 3m in the middle part. We also consider the effect of an underwater shoal, extending from \( x=0 \text{m} \) to 500m, connecting two regions of constant but different depths: the left region of wave-incidence, where \( h_1 =13 \text{m} \), and the region of transmission (right half strip), where \( h_3=7 \text{m} \). In the latter case, the average and maximum values of the slope of the bottom profile are 1.2\% and 6\%, respectively. Present method results have been obtained by retaining 5 modes in the local series Eq. (1), which was found to be enough for numerical convergence. Furthermore numerical results are based on a discretization using \( N=251 \) elements for subdivision of the segment from \( x_1 = 0 \) to \( x_5 = 500 \text{m} \), and \( p=1 \), which is shown to provide reasonable accuracy. Future work will focus on the detailed investigation of the rates of convergence and evaluation of the efficiency of present FEM scheme for higher \( p \), demonstrating the overall robustness of the solution procedure.
Figure 3. Real part (solid line) and imaginary part (dashed line) of the wave field $\varphi$ at the top of the water layer. The middle subregion containing floating body and bottom inhomogeneity extends from $x_i = 0$ to $x_i = 500m$.

Figure 4. Plot of the wave field $\varphi$ (real part) in the water layer, as calculated by the present method, using equipotential lines. The ice layer is indicated by using cyan lines. Extension of equipotential lines below the bottom profile is maintained in order to visualize the fulfilment of corresponding boundary condition both on the flat and sloping parts.

Figure 5. Real part (solid line) and imaginary part (dashed line) of the elastic deflection $w$, as calculated by the present method, in the middle subregion containing floating body and bottom inhomogeneity extending from $x_i = 0$ to $x_i = 500m$. 
CONCLUSIONS
A novel coupled-mode system of horizontal differential equations has been applied to the hydroelastic analysis of large floating bodies or ice sheets of general, finite thickness, lying over variable bathymetry regions. The present method is based on the theory of shear deformable plates (or beams), and is derived by an enhanced representation of the elastic displacement field, containing additional elastic vertical modes and permitting the shear strain and stress to vanish on both the upper and lower boundaries of the thick floating plate derived by Athanassoulis and Belibassakis[1], for the analysis of floating, shear deformable floating plates or beams of general shape. The proposed plate (or beam) model is based on the addition of extra vertical elastic deformation modes, at each horizontal position along the floating body, permitting shear strain and stress to vanish on both the upper and lower boundaries and extending third-order plate theories. First numerical results are obtained by applying the hp-version of the Finite Element Method to the solution of the resulting system, indicating good convergence rates and adaptivity capabilities.

ACKNOWLEDGEMENTS
This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: ARCHIMEDES III. Investing in knowledge society through the European Social Fund.

REFERENCES


A NUMERICAL APPROACH FOR STATIC AND DYNAMIC ANALYSIS OF DEFORMABLE JOURNAL BEARINGS

DENIS BENASCIUTTI*, MIRCEA GH. MUNTEANU* AND FABIO FLUMIAN†

* Università degli Studi di Udine - Dipartimento di Ing. Elettrica, Gestionale e Meccanica (DIEGM)
via delle Scienze 206, 33100 Udine, Italy
e-mail: denis.benasciutti@uniud.it, munteanu@uniud.it, www.uniud.it/

† Centro Ricerche Danieli - Danieli Officine Meccaniche S.p.A.
via Nazionale 41, 33042, Buttrio (UD), Italy
f.flumian@danieli.it, www.danieli.com/

Key words: Journal bearing, finite elements, elastic deformation, dynamic analysis

Abstract. This paper presents a numerical approach for the static and dynamic analysis of hydrodynamic radial journal bearings. In the first part, the effect of shaft and housing deformability on pressure distribution within oil film is investigated. An iterative algorithm that couples Reynolds equation with a finite elements (FE) structural model is solved. Viscosity-to-pressure dependency (Vogel-Barus equation) is also included. The deformed lubrication gap and the overall stress state are obtained. Numerical results are presented with reference to a typical journal bearing configuration at two different inlet oil temperatures. Obtained results show the great influence of bearing components structural deformation on oil pressure distribution, compared with results for ideally rigid components. In the second part, a numerical approach based on perturbation method is used to compute stiffness and damping matrices, which characterize the journal bearing dynamic behavior.

1 INTRODUCTION

Journal bearings are machine elements in which the applied force is entirely supported by an oil film pressure. They are used in many different engineering applications, for example as supports of rotating shafts. They are considered superior to roll-bearings because of their higher load-bearing capacity, higher operating angular speed, lower cost and easier manufacturing. Furthermore, a proper design can assure very large service lives. The early studies on the fluid-dynamic behavior of journal bearings based on the numerical solution of Reynolds equation date back to the fifties, thanks to the work of Raimondi and Boyd (R&B) [1]-[2]. They summarized results in useful dimensionless charts ready for design, which are nowadays accepted also in code standards [3].

Raimondi and Boyd analysis is based on some simplifying assumptions, as the hypothesis of constant viscosity of oil film, independency of viscosity on pressure and finally the postulation of perfectly rigid components (shaft and bushing). Such assumptions, however, can be somewhat oversimplified, considering for example that deformation of journal bearing components under imposed oil film pressure is expected to produce a change in lubrication
gap and thus a modification in the resultant pressure distribution. Moreover, also the assumption of constant viscosity and its independence from pressure should be critically reviewed, as it is experimentally known how viscosity depends, other than temperature, also on pressure, as summarized by many constitutive models [4].

It would be then of interest to investigate in more detail the correlation existing between all the above-mentioned aspects and journal bearing performance and design.

In light of the above considerations, the present paper aims to present a general numerical approach to study the static and dynamic behavior of hydrodynamic radial journal bearings, by including in the analysis the effect of the aforementioned aspects.

In the first part, attention will focus on computation of pressure distribution as a function of temperature variation within lubrication gap, viscosity-to-pressure sensitivity (according to the Vogel-Barus constitutive model [4]) and components flexibility [5]. An iterative algorithm using a finite difference scheme will be developed to solve the Reynolds equation, based on the deformed lubrication gap calculated by a coupled structural finite elements (FE) analysis. The numerical approach will compute the pressure distribution and the local stress field including shaft and bushing structural deformation. Results will clearly emphasize the strong influence of component flexibility on journal bearing performance, with a significant reduction of peak pressure caused by components deformation.

In the second part of the paper, the dynamic behavior of journal bearing will be also investigated. A numerical procedure implementing the so-called "perturbation approach" will be developed to compute the stiffness and damping matrices characterizing the dynamic behavior of hydrodynamic journal bearings. Numerical examples considering a typical journal bearing configuration will be presented.

2 JOURNAL BEARING: BASIC CONCEPTS

A typical configuration of radial journal bearing under a vertical load (see Fig. 1) consists on a shaft rotating inside a fixed support (choke), where it is usually fitted a bush. The nominal radial clearance between shaft (diameter \( d=2r \)) and choke (diameter \( D=2R \)) is \( c=R-r \).

The steady-state response of a journal bearing is governed by the fundamental equation of lubrication theory (Reynolds equation) [6]:

\[
\frac{1}{r^2} \frac{\partial}{\partial \theta} \left( h^3 \frac{\partial p}{\partial \theta} \right) + \frac{\partial}{\partial z} \left( \frac{h^3}{\mu} \frac{\partial p}{\partial z} \right) = \frac{6U}{r} \frac{dh}{d\theta}
\]

where \( h(\theta) = c - e \cos \theta \) is the oil film thickness as a function of angular coordinate \( \theta \), symbol \( e \) is the eccentricity, \( U=\omega r \) is the tangential velocity of shaft, \( \omega \) is its angular velocity, \( p(\theta, z) \) is the resultant oil pressure distribution, \( \mu \) is the oil dynamic viscosity. The numerical solution of Reynolds equation gives the pressure distribution \( p(\theta, z) \) within the lubrication gap and the system operating parameters (eccentricity, minimum lubrication gap, force resultant components, etc.).

Due to the relative velocity between shaft and support, the oil generates a pressure \( p(\theta, z) \) over the attitude angle \( \beta \), where \( p_{\text{max}} \) is the peak pressure that occurs at angle \( \theta_{p_{\text{max}}} \). The system moves in a new equilibrium configuration, where the eccentricity \( e \) characterizes the
position of shaft axis with respect to the fixed support axis, along direction defined by angle $\theta_{h0}$ (which also identifies the direction of minimum oil thickness $h_0$).

Figure 1: Sketch of a hydrodynamic journal bearing

Several design charts are available in literature [1]-[2], which provide journal bearing operation parameters as a function of Sommerfeld number $S=(r/c)^2 (\mu N/p_m)$, defined in terms of shaft radius $r$ and rotational speed $N$, while $p_m=F/(LD)$ is the average (specific) pressure defined as the ratio of the applied radial force $F$ and the nominal projected area ($L$ is the length of journal bearing). Such charts were determined by R&B through numerical solution of Reynolds equation under the hypothesis of constant temperature (and thus viscosity) of lubrication film and also under the assumption of perfectly rigid components (shaft and support).

An improvement of the analysis can be obtained by including in Reynolds equation a more sophisticated constitutive model for the viscosity. For example, a coupled temperature-pressure dependency can be summarized by the experimentally determined Vogel-Barus equation $\mu=\mu_0 \exp(\alpha p)$, in which $\mu_0$ is a pressure-independent viscosity term (only function of temperature) and $\alpha$ is a sensitive parameter related to oil film pressure (typical values are $\alpha=0.01\div0.02$ MPa$^{-1}$). In accordance to this constitutive model, an increase in dynamic viscosity occurs for high pressures, with a solid-like behavior for very high pressures. This effect, well-known in elasto-hydrodynamic studies (e.g. lubricated contacts), has not been actually investigated in the field of journal bearings.

A further improvement in journal bearing analysis can be obtained by including in the solution of Reynolds equation the deformed shape of lubrication gap caused by deformation of shaft and support under imposed oil pressure $p(\theta, z)$.

This paper will present a general numerical approach to compute the pressure distribution by also including the above mentioned effects. A typical journal bearing configuration (see Table 1), operating at two different inlet oil temperatures ($T_{in}=40$ and 70 °C), will be investigated. A viscosity-temperature curve typical of an oil ISO VG 680 will be used in all simulations [4].
Table 1: Geometrical dimensions used in numerical simulations

<table>
<thead>
<tr>
<th>(d) (mm)</th>
<th>(D) (mm)</th>
<th>(L) (mm)</th>
<th>(F) (kN)</th>
<th>(N) (rpm)</th>
<th>(p_m) (MPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>500.5</td>
<td>300</td>
<td>3600</td>
<td>65</td>
<td>24</td>
</tr>
</tbody>
</table>

3 STEADY-STATE ANALYSIS

In numerical simulations two models for the journal bearing were adopted: a 2D model and a 3D one. In the first part of this paper the hypotheses used are rigid components and viscosity function of both temperature and pressure according to the Vogel-Barus model. In the second part, while shaft and support deformation will be included into the analysis, pressure effect will be considered too although it could be neglected in this case due to the drastic reduction of the maximum pressure.

3.1 Temperature and pressure effect (with rigid components)

Reynolds equation (1) is solved by using the finite difference method based on central difference scheme [7]. The unknown function in (1) is the pressure \(p(\theta, z)\) that, upon integration, gives the resultant applied load \(F\), which in fact is a given input.

It is worth noting that the problem is actually not linear for several reasons. Although the pressure \(p(\theta, z)\) is the unknown function, equation (1) does not explicitly depends on load \(F\) (i.e. the resultant of pressure), but on eccentricity through the lubrication gap \(h(\theta) = c - e \cos(\theta)\).

Several iterations (Newton–Raphson method was used) are then required to first impose the input force \(F\) (as resultant of pressure) and to find the appropriate pressure distribution \(p(\theta, z)\) that solves (1).

Table 2: Overall comparison of results from numerical simulations for rigid components, 2D model \((L/D \sim \infty)\)

<table>
<thead>
<tr>
<th>Configurations</th>
<th>(T_m) (^\circ)C</th>
<th>(T_m) (^\circ)C</th>
<th>(T_{out}) (\circ)C</th>
<th>(\mu) P(\mu)-S (\mu)-S</th>
<th>(S) (\mu)-S</th>
<th>(e) mm</th>
<th>(p_{max}) MPa</th>
<th>(\theta_{p_{max}}) deg</th>
<th>(h_0) mm</th>
<th>(\theta_{h_0}) deg</th>
</tr>
</thead>
<tbody>
<tr>
<td>R&amp;B</td>
<td>(T_m) cost. (\alpha = 0)</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>0.1678</td>
<td>0.00786</td>
<td>0.2352</td>
<td>87.30</td>
<td>15.50</td>
<td>0.0148</td>
</tr>
<tr>
<td>&amp; (\alpha = 0.01)</td>
<td>40</td>
<td>60</td>
<td>80</td>
<td>↓</td>
<td>↓</td>
<td>0.2286</td>
<td>83.74</td>
<td>15.03</td>
<td>0.0214</td>
<td>27.03</td>
</tr>
<tr>
<td>&amp; (\alpha = 0)</td>
<td>(T_{in}) ~(T_{out}) lin.</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>Not</td>
<td>Not</td>
<td>0.2392</td>
<td>83.17</td>
<td>22.82</td>
<td>0.0108</td>
</tr>
<tr>
<td>&amp; (\alpha = 0.01)</td>
<td>(T_{in}) ~(T_{out}) lin.</td>
<td>defined</td>
<td>defined</td>
<td>0.2350</td>
<td>80.18</td>
<td>22.43</td>
<td>0.0150</td>
<td>33.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R&amp;B</td>
<td>(\alpha = 0)</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>0.0655</td>
<td>0.00298</td>
<td>0.2447</td>
<td>205.50</td>
<td>6.60</td>
<td>0.0053</td>
</tr>
<tr>
<td>&amp; (\alpha = 0.01)</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>↓</td>
<td>↓</td>
<td>0.2440</td>
<td>136.92</td>
<td>10.27</td>
<td>0.0060</td>
<td>16.27</td>
</tr>
<tr>
<td>&amp; (\alpha = 0)</td>
<td>(T_{in}) ~(T_{out}) lin.</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>Not</td>
<td>Not</td>
<td>0.2412</td>
<td>149.54</td>
<td>9.34</td>
<td>0.0088</td>
</tr>
<tr>
<td>&amp; (\alpha = 0.01)</td>
<td>(T_{in}) ~(T_{out}) lin.</td>
<td>defined</td>
<td>defined</td>
<td>0.2431</td>
<td>173.29</td>
<td>9.80</td>
<td>0.0069</td>
<td>16.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Secondly, the force-eccentricity relationship \(F-e\) is highly non-linear, especially for eccentricity values \(e\) approaching the nominal radial clearance \(c\). Another source of non-
linearity is that negative pressure values must be set to zero during the iterative process.

Table 3: Overall comparison of results from numerical simulations for rigid components, 3D model ($L/D=0.6$)

<table>
<thead>
<tr>
<th>Configurations</th>
<th>$T_{in}$ °C</th>
<th>$T_{out}$ °C</th>
<th>$\mu$ Pa⋅s</th>
<th>$S$ -</th>
<th>$e$ mm</th>
<th>$p_{max}$ MPa</th>
<th>$\theta_{pmax}$ deg</th>
<th>$h_0$ mm</th>
<th>$\theta_{h0}$ deg</th>
</tr>
</thead>
<tbody>
<tr>
<td>R&amp;B</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>0.2405</td>
<td>158.00</td>
<td>6.00</td>
<td>0.0095</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$T_{m}$ cost.</td>
<td>$\alpha = 0$</td>
<td>↑</td>
<td>↑</td>
<td>0.1678</td>
<td>0.00786</td>
<td>0.2397</td>
<td>152.43</td>
<td>6.15</td>
<td>0.0103</td>
</tr>
<tr>
<td>JB1 $T_{m}$ cost.</td>
<td>$\alpha = 0.01$</td>
<td>40</td>
<td>60</td>
<td>80</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>0.2357</td>
<td>219.08</td>
<td>7.50</td>
</tr>
<tr>
<td>$T_{in}$-$T_{out}$ lin.</td>
<td>$\alpha = 0$</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>Not</td>
<td>Not</td>
<td>0.2442</td>
<td>179.06</td>
<td>8.28</td>
<td>0.0058</td>
</tr>
<tr>
<td>$T_{in}$-$T_{out}$ lin.</td>
<td>$\alpha = 0.01$</td>
<td>defined</td>
<td>defined</td>
<td>0.2411</td>
<td>285.85</td>
<td>6.97</td>
<td>0.0089</td>
<td>14.67</td>
<td></td>
</tr>
<tr>
<td>R&amp;B</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>0.2456</td>
<td>269.39</td>
<td>3.10</td>
<td>0.0044</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$T_{m}$ cost.</td>
<td>$\alpha = 0$</td>
<td>↑</td>
<td>↑</td>
<td>0.0655</td>
<td>0.00298</td>
<td>0.2452</td>
<td>203.07</td>
<td>5.95</td>
<td>0.0048</td>
</tr>
<tr>
<td>JB2 $T_{m}$ cost.</td>
<td>$\alpha = 0.01$</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>0.2424</td>
<td>407.09</td>
<td>5.25</td>
</tr>
<tr>
<td>$T_{in}$-$T_{out}$ lin.</td>
<td>$\alpha = 0$</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>Not</td>
<td>Not</td>
<td>0.2464</td>
<td>221.00</td>
<td>6.11</td>
<td>0.0036</td>
</tr>
<tr>
<td>$T_{in}$-$T_{out}$ lin.</td>
<td>$\alpha = 0.01$</td>
<td>defined</td>
<td>defined</td>
<td>0.2438</td>
<td>674.38</td>
<td>4.56</td>
<td>0.0062</td>
<td>10.97</td>
<td></td>
</tr>
</tbody>
</table>

To evaluate the effect of temperature on viscosity, and consequently on pressure distribution, the journal bearing configuration in Table 2 (2D model) and Table 3 (3D model) was studied at two operating conditions (JB1, JB2) characterized by two different inlet temperatures ($T_{in}=40, 70°C$). Two hypotheses are then adopted to compute the pressure-independent viscosity term $\mu_0$ as a function of oil temperature: in the first, using an average constant oil temperature $T_m$ resulting by a thermal balance inside the oil film (as in R&B approach), in the second using, as a first approximation, a linear temperature variation from inlet value $T_{in}$ to the outlet value $T_{out}$ (that has been calculated by previous thermal balance). Note that in both cases the same average oil film temperature $T_m$ is obtained.

For both temperature distributions within lubrication gap (constant $T_{m}$, linear $T_{in}$-$T_{out}$), the Vogel-Barus equation has been implemented with two different cases ($\alpha=0$ and $\alpha=0.01$).

Table 2 shows an overall comparison of obtained results for the 2D model and Table 3 the comparison for the 3D model. Figures 2 and 3 compare the pressure distribution under an imposed vertical load, with a linear temperature variation within oil film and assuming different pressure sensitivity values for viscosity. As expected with 3D model it results larger values for maximum pressure because the 2D model assumes a constant distribution along shaft axis.

The effect of temperature variation of oil film is first analyzed. Referring to JB1 configuration in Table 2, a negligible difference is observed between the case of constant and linearly varying temperature, for both $\alpha=0$ and $\alpha=0.01$ values. Instead, larger differences (with a 10-12% increase of $p_{max}$ value) are observed for JB2 configuration, considering both $\alpha=0$ and $\alpha=0.01$ values. On the contrary, for 3D model the value of maximum pressure, much larger with respect of 2D case, strongly depends on $\alpha$ sensitivity factor. Also $p_{max}$ depends on temperature variation law.

This emphasizes how the variation of oil film temperature could have some effect on
pressure distribution, especially for high temperature values. Considering the viscosity-
temperature strong correlation, this seems to confirm that pressure distribution is more
sensitive to a change of viscosity values within lubrication gap. Constant viscosity assumption
used in R&B calculations (no temperature and no pressure influence) seems too simplified.

\[ \alpha = 0, \ p_{\text{max}} = 83.17 \text{ MPa} \]
\[ \alpha = 0.01, \ p_{\text{max}} = 80.18 \text{ MPa} \]

\[ \alpha = 0, \ p_{\text{max}} = 151.08 \text{ MPa} \]
\[ \alpha = 0.01, \ p_{\text{max}} = 173.29 \text{ MPa} \]

**Figure 2:** Results for JB1 and JB2 configurations, 2D model

\[ \alpha = 0, \ p_{\text{max}} = 179.06 \text{ MPa}, \ e = 0.2442 \text{ mm} \]
\[ \alpha = 0, \ p_{\text{max}} = 221.00 \text{ MPa}, \ e = 0.2464 \text{ mm} \]

**Figure 3:** Results for JB1 and JB2 configurations, 3D model
On the other hand, in the case of rigid components (shaft and bushing), the 2D model to solve Reynolds equation (1) is not suitable as pressure distribution is very different from 3D model, see figures 2 and 3.

Numerical solutions for constant $T_m$ and $\alpha=0$ were also compared with results given by R&B charts, showing a good agreement only for JB1 configuration, while some difference characterizes JB2 configuration. The observed discrepancy can be attributed to the very low Sommerfeld number ($S=0.00298$) characterizing JB2 configuration, which makes difficult using R&B design charts and thus can be source of interpolation errors.

A non-zero viscosity-to-pressure sensitivity ($\alpha=0.01$) determines an important variation in the overall pressure distribution (change of attitude angle $\beta$) and in its maximum value $p_{max}$, depending on the general pressure levels attained. In the case of 2D model with $T_{in}-T_{out}$ linear temperature variation, for peak pressures $p_{max}<100$ MPa (2D model, case JB1), the pressure effect is actually negligible, as shown in Fig. 2a, with only a small decrease of the maximum pressure of about 3.5%. For larger pressure levels (case JB2), an increment of $p_{max}$ of about 12% is observed, see Fig. 2b. The minimum oil thickness increment ($h_0=c-e$) produced by the pressure effect is relevant in both cases, with a variation respectively of 28% and 32%.

For the 3D model the maximum pressure are much higher and the effect of $\alpha$ is extremely important. The minimum film oil thickness has the same tendency as in the previous case but the values are larger: 37% and 42% respectively.

The obtained results can be summarized by saying that, if the influence of pressure on viscosity is taken into consideration, when $\alpha$ increases the peak pressure $p_{max}$ increases, while the eccentricity $e$ decreases. The conclusion of detailed study shows that the pressure-to-viscosity effect is smaller compared to temperature influence if the maximum pressure is smaller than 90...100 MPa and in this case could be neglected.

3.2 Effect of component deformation (T linear)

In the second part of this work, the pressure distribution will be calculated by considering the real geometry of lubrication gap resulting from component deformation. Pressure values calculated by solving the Reynolds equation (1) are used, as input in a FE model, to compute the geometry of lubrication gap after deformation, which is next used to solve again equation (1) with an iterative analysis scheme. Details on the numerical algorithm can be found in [7].

A fluid-structural coupled numerical procedure was developed in Matlab environment. The first analysis step is the calculation of pressure distribution $p(\theta, z)$ and eccentricity $e$ for the case of not deformable components, by solving Reynolds equation (1). The obtained pressure distribution is next applied as input mechanical load in a plane structural FE model, which gives the relative radial displacements between shaft and support after deformation, and the resulting gap deformation $g(\theta, z)$. A new oil film geometry $h'(\theta, z)=c-e\cos(\theta)+g(\theta, z)$ that incorporates mechanical deformation (thus it differs from the case of perfectly rigid components) can be thus calculated. At second iteration step this updated gap geometry $h'(\theta)$ is entered in (1) to get a new pressure distribution $p'(\theta, z)$ that balances the input force $F$. This iterative procedure is repeated until convergence is achieved with respect to an imposed threshold tolerance on the maximum pressure [7].
The 3D FE models of both shaft and support used in the analysis are shown in figures 4a and 4b. The shaft and the support are modeled by a mesh with 8-node brick isoparametric finite elements. The 2D model uses 4-node plane isoparametric finite element. The support is fixed on the lower surface, while the shaft is clamped on one end (z=300 mm, Fig. 4a).

Shaft and support are loaded by the same oil pressure distribution \( p(\theta, z) \) applied on the outer and inner surfaces, respectively. Analysis assumes small displacements. In the case of 2D model a plane strain condition was considered. Material has linear elastic behavior, with properties typical of a structural steel.

It is worth noting that the use of a plane FE model for the structural analysis of a journal bearing requires a special attention in modeling mechanical constraints. In fact, in a real journal bearing the applied load \( F \) and the resulting pressure distribution are actually applied along different longitudinal locations along the shaft axis. Instead, in the plane FE model here adopted the external load \( F \) that balances the oil pressure is replaced by an appropriate constrain on shaft geometry. For this purpose, the shaft has been modeled with a central hole and all nodes on the inner circumference have imposed zero radial displacements, Fig 4c; the support, instead, has all the external edges constrained. This modeling strategy, however, affects the shaft structural stiffness: a large inner radius determines an anomalous increment of shaft stiffness, while a very small inner hole gives rise to very large deformations and abnormally high reaction forces at constrained nodes. A proper sensitivity analysis has been preliminary carried out, in order to find the optimal radius of inner hole.

The coupled numerical approach was applied to study the JB1 configuration (with \( \alpha=0.01 \) and linear temperature variation in \( T_{in}=40^\circ C-T_{out}=80^\circ C \)). Fig. 5 shows the result for the case of deformable components, 3D model. The comparison with the case of rigid components in Fig. 3a clearly emphasizes how component deformation determines a reduction to about 20% (285.85 MPa to 56.04 MPa) of the maximum peak pressure in the median plane of the shaft (Fig. 5b) and, accordingly, an increase in the attitude angle \( \beta \) (under the same applied resultant

Figure 4: Finite element model of shaft and support
force $F$). The pressure profile, more uniform than the case of rigid components (R&B solution), seems to support the idea of using the average pressure $p_{m}$ as a structural design parameter, as suggested in some design codes [3].

The absolute maximum of the pressure is around 65…85 MPa and it is reached at the two ends of the shaft. This value depends on the local mesh fineness and this aspect will be discussed in a future paper. The 2D model gives a value of maximum pressure (considered valid in the median shaft section) of 49.98 MPa, closed enough to 3D model.

![Image](image1.png)

**Figure 5**: Pressure distribution, JB1 configuration ($\alpha=0, T=40^\circ - 80^\circ$ C), deformable components, 3D model

![Image](image2.png)

**Figure 6**: Stress distribution (MPa units)

Fig. 5c also compares the geometry of lubrication gap for the case of deformable and rigid components (angles are referred to the position of minimum oil gap $\theta_{h0}$). It is observed that
for deformable components the gap is not symmetric and that eccentricity can assume values greater than the nominal clearance, as deformation can increase the gap between shaft and support.

For what concerns the calculated mechanical stresses, Fig. 6 shows stress distribution in the shaft (radial stress) and in the support (von Mises stress). Compared to the case of perfectly rigid components, this explains the relatively small values of von Mises stress calculated in the support, which is actually comparable with static strength of materials usually employed in the bush (for instance, white metal generally used as internal coating has a yield stress of about 50 MPa [5]).

4 DYNAMIC ANALYSIS

The dynamic behavior of rotating shaft supported by journal bearings is strongly influenced by the hydrodynamic forces produced in lubricant film that oppose to shaft movement. Determination of rotor dynamics then requires full characterization of the dynamic response of bearing lubricant film, which is a non-linear function of position and velocity of journal center.

In the dynamic analysis of a rotor-bearing configuration, a simple spring-dashpot model is usually adopted to account for journal bearing contribution [6]. With small displacements increments (δx, δy) and small velocities increments (δ̇x, ̇δy) in the vicinity of the journal bearing static equilibrium position, a linearized relationship, between the incremental oil-film forces F̃i and journal displacements and velocities increments that cause them, can be written as [6]:

\[
\begin{bmatrix}
F̃_x \\
F̃_y
\end{bmatrix} =
\begin{bmatrix}
k_{xx} & k_{xy} \\
k_{yx} & k_{yy}
\end{bmatrix}
\begin{bmatrix}
δx \\
δy
\end{bmatrix} +
\begin{bmatrix}
c_{xx} & c_{xy} \\
c_{yx} & c_{yy}
\end{bmatrix}
\begin{bmatrix}
δ̇x \\
δ̇y
\end{bmatrix}
\]  

where \( k_{ij} = (\partial F_i / \partial x_j) \) and \( c_{ij} = (\partial F_i / \partial \dot{x}_j) \) are the linear stiffness and damping coefficients, respectively.

A classical "perturbation method" is followed to compute the increase in oil film forces resulting from a departure (perturbation) from the static equilibrium position. A journal bearing configuration, characterized by given displacement \((u, v)\) and velocities \((\dot{u}, \dot{v})\) of journal centre, is first assigned. The reference pressure distribution \(p(\theta, z)\) and oil film forces \(F_x\) and \(F_y\) are then calculated by Eq. (3), the Reynolds equation in dynamic regime:

\[
\frac{h^3}{\mu} \frac{\partial}{\partial \theta} \left( \frac{1}{\mu} \frac{\partial p}{\partial \theta} \right) + 3h \frac{\partial h}{\partial \theta} \frac{\partial p}{\partial \theta} = \frac{6U\mu}{r} \frac{\partial h}{\partial \theta} + 12 \frac{\partial h}{\partial \theta}
\]  

which explicitly depends also on the time derivative of lubrication gap \(\dot{h}(\theta) = -\dot{u} \cos \theta - \dot{v} \sin \theta\). Independent displacement and velocity perturbations are next applied and the corresponding force increments calculated. Solution of (3) gives the increased pressure distribution (say \(p + \Delta p\)), and therefore the increased resultant of oil film forces \(F_x + \Delta F_x\) and \(F_y + \Delta F_y\), with respect to the reference equilibrium position, for shaft displacement and velocity increments \((\delta x, \delta y, \delta \dot{x}, \delta \dot{y})\). Thus stiffness and damping coefficients can be thus
determined, as for example:

\[
[k] = \begin{bmatrix}
\frac{\partial F_u}{\partial u} & \frac{\partial F_u}{\partial v} \\
\frac{\partial F_v}{\partial u} & \frac{\partial F_v}{\partial v}
\end{bmatrix}
\]  \tag{4}

To include also the contribution of the structure deformation into stiffness and damping matrices, in the above expression the displacement increments are substituted by \(\delta u_o\) and \(dv_o\), total displacement increments of the bearing center. Absolute displacements \(u_o\) and \(v_o\) of the bearing center and their increments \(\delta u_o\) and \(dv_o\) are found as explained in section 3.2. A similar approach is used to determine damping matrix.

Stiffness \([k]\) and damping \([c]\) matrices characterize the dynamic behavior of journal bearing and they enter into the dynamic equilibrium equations of the shaft. It is worth noting that, due to the non-linear nature of the Reynolds equation (3), both matrices explicitly depend on the assigned journal bearing displacement \((u, v)\) and velocities \((\dot{u}, \dot{v})\), that is they have to be interpreted as tangent matrices. Therefore, a transient dynamic analysis of a rotor supported by journal bearings is non-linear and \([k]\), \([c]\) matrices must be calculated at every time integration step. In addition, \([k]\) and \([c]\) are in general not symmetric.

A numerical algorithm has been specifically developed to compute stiffness and damping matrices by the perturbation method previously described. The procedure has been applied to characterize different journal bearing configurations.

An example of calculated pressure distribution for JB1 configuration for two different velocities, \(\dot{v} = 0\) and \(\dot{v} = 0.5\) mm/s, is shown in Fig. 7 (2D model). The calculated stiffness and damping coefficients are reported in Fig. 8 and Fig. 9: a high non-linear dependence on the eccentricity \(e\) is observed.

5 CONCLUSIONS

The present papers developed a numerical procedure for the steady state and dynamic analysis of hydrodynamic radial journal bearing. Influence of temperature and pressure on viscosity and thus on resultant pressure distribution were studied. A mechanical plane finite element model, coupled with solution of Reynolds equation, was also developed to study journal bearing structural behavior and its influence on pressure distribution. Finally, a perturbation approach was implemented to evaluate stiffness and damping coefficients.

The main findings of the work can be summarized as follows:

- temperature increase was shown to give a decrease of attitude angle \(\beta\) and an increase in pressure peak;
- an increase of viscosity-to-pressure sensitivity (\(\alpha\) value) gives a general increase of peak pressure for pressure peaks greater than about 90...100 MPa;
- component deformation gives a more uniform pressure distribution, with a considerable reducing of the peak pressure compared to the case of ideally rigid components;
- stiffness and damping coefficient were calculated and a high non-linear trend with journal bearing eccentricity \(e\) was observed.
Figure 7: Pressure distribution calculated for $u=0$, $\dot{u} = 0$, $\nu=0.23$ mm and $\dot{v} = 0$ (left) and $\dot{v} = 0.5$ mm/s (right)

Figure 8: Stiffness coefficients [N/mm/mm], JB1, versus eccentricity (for $u=0$, $\dot{u} = 0$, $\nu = 0$)

Figure 9: Damping coefficients [Ns/mm/mm], JB1, versus eccentricity (for $u=0$, $\dot{u} = 0$, $\nu = 0$)

REFERENCES

ADDED RESISTANCE OF SHIPS IN QUARTERING SEAS

SHUKUI LIU* AND APOSTOLOS D. PAPANIKOLAOU †

School of Naval Architecture and Marine Engineering
National Technical University of Athens
9, Heroon Polytechniou, 15773,
Zografou-Athens, Greece
http://www.naval.ntua.gr/sdl
*email: liushukui@deslab.ntua.gr, †email: papa@deslab.ntua.gr

Key words: added resistance in waves, quartering seas, short sea formula, far field method.

Abstract. In this paper, the 3D panel method of NTUA-SDL for calculating the added resistance of a ship in bow seas has been extended and validated for the quartering seas case. At first we studied the drift force on a semi-submersible in quartering seas. Comparison has been made between numerical results of a near field and far field approach, and good agreement has been observed. In the following, we applied our method to the study of the added resistance of the S175 ITTC standard container ship and a bulk carrier for different forward speeds and different heading angles. The obtained numerical results agree also well with available experimental data, except for the short waves range. In the short waves range, the actual physical phenomenon is very complicated, involving strong viscous and other nonlinear effects, consequently the present method based on potential theory, cannot be satisfactory. In order to improve the prediction, a new semi-empirical formula is proposed to include the aforementioned effects arising in the short, oblique waves scenario. Initial tests with this formula have shown that the proposed formula provides a good approximation for the added resistance in short oblique seas. We expect that this practical formula can be further improved if additional experiments will be available in the future.

1 INTRODUCTION

The accurate prediction of ship’s added resistance in seaways is nowadays of high scientific and practical interest, for both researchers, ship designers, yards, and shipowners/operators; this because it greatly affects the selection of ship’s engine/propulsion system and influences ship’s performance in terms of sustainable speed and safety, fuel consumption and engine air emissions in realistic sea conditions. This calls for optimizations of ship’s hull in terms of her total resistance in waves. Even more, accurate and efficient predictions of the added resistance in natural seaways are necessary for the implementation of modern and reliable on-board ship routeing systems[1].

In a previous study[2], the authors presented a numerical method to predict the added resistance of ships sailing in waves, but the validation has been limited to head seas only. The current paper presents a follow-up validation work. In addition, emphasis has been put on the prediction of added resistance in short waves, thus in waves of length of about less than half ship’s length. Noting the continuous increase of ship sizes in recent years (in view of the
economy of scale) it is obvious that the range of relative wave length $\lambda/L$ of practical interest is being shifted to lower values, which makes the prediction of added resistance of ships in such ranges more and more important. For the short wave range added resistance prediction, the asymptotic formula of Faltinsen\cite{3} and the improved expression of Takahashi\cite{4} are the most widely used formulas. However, Faltinsen’s formula, which is based on theoretical considerations for very short waves, is practically not very accurate, because it does not account for viscous effects. On the other hand, Kuroda & Tsujimoto’s\cite{5,6} proposed empirical formula, which is based on conducted tank tests, yields improved prediction results, but their recommendation appears not to cover all practical cases. In this paper, starting from the above pioneering achievements, we propose a new practical formula to predict the added resistance of ships in short oblique seas.

2 THEORETICAL BACKGROUND

2.1 Far field method for the added resistance in waves

The theoretical background for the calculation of the added resistance in waves has been described in detailed in our previous paper\cite{2}. Here only the final formulation will be given without further explanation.

The added resistance based on Maruo’s theory\cite{7} may be expressed by using the Kochin function as:

$$ R_{AW} = \frac{\rho}{8\pi} \left\{ \int_{-\alpha_0}^{\alpha_0} \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{\frac{3\pi}{2}}^{\frac{5\pi}{2}} \left| H(k_1, \theta) \right|^2 \frac{k_1 \left[ k_1 \cos \theta - K \cos \chi \right]}{\sqrt{1 - 4\Omega \cos \theta}} d\theta \right\} d\phi ight\} + \frac{\rho}{8\pi} \int_{2\pi - \alpha_0}^{2\pi} \left| H(k_2, \theta) \right|^2 \frac{k_2 \left[ k_2 \cos \theta - K \cos \chi \right]}{\sqrt{1 - 4\Omega \cos \theta}} d\theta $$

The complex Kochin function describes the elementary waves radiated from the ship and is expressed as:

$$ H(k_j, \theta) = \int \left( \frac{\partial}{\partial n} \cdot \frac{\partial \phi}{\partial n} \right) G_j(\theta) ds $$

where $G_j(\theta) = \exp \left[ k_j(\theta) z + ik_j(\theta)(x \cos \theta + y \sin \theta) \right]$. It is an integrated effect of the body’s geometry and motions.

In current study, the 3D frequency domain panel code NEWDRIFT\cite{8,9} is used to solve the diffraction and radiation velocity potentials on each panel of ship’s wetted surface and then the above formula is implemented to obtain the added resistance. Results from application of this formula to various case studies are denoted as $ND \text{ far}$ in the following graphs.

2.2 Semi-empirical correction for short waves range

In our previous study, the semi-empirical formulae proposed by Faltinsen\cite{3} and Kuroda & Tsujimoto\cite{5,6} have been investigated. However, Faltinsen’s formula is practically not very accurate, while Kuroda & Tsujimoto’s proposed formula, which is based on the tank tests and yields much better results, appears not to cover some practical cases, at least theoretically. We focus in the following on deriving a more versatile, simple but rational formula.
In the short waves range we may assume that radiation (ship motion) effects are negligible, thus we deal mainly with phenomena related to diffracted/transmitted/reflected waves, viscous and other nonlinear, hull form phenomena, encountered by a ship sailing in oblique seas at certain speed.

The added resistance in short waves based on Takahashi’s formula\(^\text{[4]}\) is written as:

\[
R_{AW}^{D,v} = \frac{1}{2} \rho g \zeta \alpha \beta B_\alpha \alpha \left[ 1 + \alpha V^* f(Fn) \right]
\]

where \( B_\alpha = \left[ \int \sin^2(\theta - \alpha) \sin \delta d\delta + \int \sin^2(\theta + \alpha) \sin \delta d\delta \right] / B \)

\[
\alpha = \frac{\pi I^T_{\alpha} (kd)}{\pi I^T_{\alpha} + K^T_{\alpha} (kd)}
\]

The first part, i.e., \( \frac{1}{2} \rho g \zeta \alpha \beta B_\alpha \alpha \), accounts for ship’s waterplane contour, the main particulars of the hull and the wave characteristics; it is obviously an approximation of the added resistance due to the zero speed diffraction phenomenon. Figure 1 shows the integration interval in \( B_\alpha \)'s expression, i.e., the non-shaded part (A-F-B) of the waterplane, facing the incoming wave. The term in the bracket is a correction term for the speed of advance of the ship and the wave heading. A critical review of results of the experimental study\(^{[5,6]}\) with respect to the validity of the above formula reveals that the coefficient for the speed correction is related to ship’s hull form as well. Nevertheless, the above formula exhibits the beauty of simplicity as it expresses the effect of each fundamental physical phenomenon separately, and not coupled.

![Figure 1: Coordinate system for short wave approximate method](image)

It is well established that the added resistance due to zero speed diffraction can be easily calculated based on theoretical methods, and at least theoretically, these calculations can exactly include the hull form and wave excitation effects. So it is rational to replace the above approximation term in (3) with it and then we have a formula as follows:

\[
R_{AW}^{D,v} = R_{AW}^{D,0} \left[ 1 + \alpha V^* f(Fn) \right]
\]

Then the problem is how to improve the forward speed effect. As shown in the original work of Takahashi, this term is actually equal to \( 3.5 \sqrt{F_n} (-\cos \chi) \). This expresses the effect of forward speed in the way similar to the concept of calculating the frequency of encounter and clearly states the necessity to include the effect of forward speed on the diffraction problem.
But it also says that in beam seas the correction vanishes. However, we know that the added resistance in actual (short) seaways is certainly not only due to the diffraction effect but also due to other effects, such as wave and hull form nonlinearities and viscosity effects, which should be linked to ship’s projected area to the wave direction. This has been also revealed by the known experimental work \[^{[5,6]}\], based on the recommended line of $C_U$ for practical use, the speed parameter $C_U$ has a minimum value of 10, which indicates the necessity of correction for forward speed in whatever heading. Reinvestigating the results of the experimental work and in consideration of deriving a more straightforward expression, we propose the following formula:

$$R_{AW}^{\text{Refl,v}} = R_{AW}^{D,0} \left\{ 1 + C_1 F_n (-\cos \chi) + C_2 \left[ (L-B) \sin \chi + B^* C_B (1+ \cos \chi C_B) \right] F_n \right\} \quad (5)$$

This expression expresses the influence of forward speed on diffraction problem and other contributions more clearly. It says that the added resistance in short waves has two parts, one part is due to diffraction, which can be calculated based on zero speed diffraction problem and corrected for speed, corresponding to the first two term in the last bracket; the other part is due to nonlinearity, viscosity etc and it has to do with ship’s projected area in the wave heading direction, which is denoted approximately by $((L-B) \sin \chi + B) / B$. The hull form influence is further corrected by using the block coefficient $(1-C_B) / (1+ \cos \chi C_B)$. Finally, the added resistance in short waves is calculated as follows:

$$R_{AW} = R_{AW}^{Ra} + R_{AW}^{D,0} + R_{AW}^{D,0} \left\{ C_1 F_n (-\cos \chi) + C_2 \left[ (L-B) \sin \chi + B^* C_B (1+ \cos \chi C_B) \right] F_n \right\} \quad (6)$$

where $R_{AW}$ is the total added resistance, $R_{AW}^{Ra} + R_{AW}^{D,0}$ is calculated based on the far field method and the third term will be the correction; but we have to first determine the related parameters for practical use of this formula.

3 VALIDATIONS AND DISCUSSIONS

3.1 Semi-submersible case

The first study has been conducted for a semi-submersible, which has been investigated extensively in previous ITTC studies\[^{[10,11,12]}\]. The scope of this study is to validate the newly extended computer code for the added resistance. The principal dimensions of the model used in the calculations are shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Main dimensions and other data of the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Draft</td>
</tr>
<tr>
<td>$X_G$</td>
</tr>
<tr>
<td>$Y_G$</td>
</tr>
<tr>
<td>$Z_G$</td>
</tr>
</tbody>
</table>

Figure 2 shows the comparison of the results from far field method (ND far), near field method (direct integration method)\[^{[13]}\] with free motion mode (ND FM) and due to diffraction effect only (ND DIFRAC), all of which are based on NEWDRIFT potential. As shown, a good agreement between the results from far field method and near field method is observed.
and they all fall into the experimental data range (though not presented here), which verifies the correctness of the extended computer code.

Figure 2: Drift forces of a semi-submersible in quartering seas

3.2 S-175 containership case

The S175 ITTC standard hull\textsuperscript{[14]} is chosen for the second case study. This ship has a low block coefficient $C_B=0.5716$ and large length to beam ratio $L/B = 6.89$; her main particulars are shown in Table 2. In order to identify the influence of forward speed and heading angle on the added resistance of S175 containership in short waves, the following numerical experiment is carried out. Table 3 shows the available experimental data collected from different sources\textsuperscript{[4,14]}. The drift force from numerical calculation by using NEWDRIFT is also included. As can be seen, there are sufficient experimental data only for head seas, hence we will try to determine the parameters based on these data.

Table 2: Main dimensions of the S175 containership

<table>
<thead>
<tr>
<th>Lpp</th>
<th>175.0 m</th>
<th>$C_B$</th>
<th>0.5716</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>25.4 m</td>
<td>KM</td>
<td>10.52 m</td>
</tr>
<tr>
<td>T</td>
<td>9.5 m</td>
<td>GM</td>
<td>1 m</td>
</tr>
<tr>
<td>D</td>
<td>15.4 m</td>
<td>$K_{XX}/B$</td>
<td>0.328</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>24 742 t</td>
<td>$K_{YY}/L_{pp}$</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Based on the available 9 experimental points, the sum $C_1+C_2$ is approximated by using a weighted least square method with the results of $C_1+C_2\approx12$, as plotted in Figure 3. In order to determine their value respectively, they are assumed with different ratios, as listed in Table 4.

In the following we present the obtained results from the above elaborated methods, in comparison with the experimental data, as shown in Figure 4. From the comparison at both $F_n=0.15$ and $F_n=0.25$ for heading angle of $\gamma=150^\circ$, $120^\circ$ and $90^\circ$, it is observed that the far field method can predict the added resistance of S175 ship with satisfactory accuracy in the range other than short waves, i.e. $\lambda/L<0.5$. For short waves, the results based on three
different assumptions are plotted. It is observed that, the predicted added resistance decreases from case-1 to case-3, which demonstrates the sensitivity/dependency of added resistance on these chosen parameters. However, due to the limitation/shortage of experiment data, we cannot make further comments on this point. On the other hand, the predicted value does agree reasonably with the available experimental data, in term of both quantity and tendency, thus proves that the proposed formula is capable of capturing the added resistance in short waves.

Table 3: Added resistance of S175 hull, $\lambda/L=0.5$

<table>
<thead>
<tr>
<th>$R_{AW}$</th>
<th>$F_n=0.0$ (NEWDRFIT)</th>
<th>$F_n=0.15$ (EXP)</th>
<th>$F_n=0.20$ (EXP)</th>
<th>$F_n=0.25$ (EXP)</th>
<th>$F_n=0.275$ (EXP)</th>
<th>$F_n=0.30$ (EXP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi=180^\circ$</td>
<td>0.584</td>
<td>1.06</td>
<td>1.09</td>
<td>3.23</td>
<td>2.942</td>
<td>2.6</td>
</tr>
<tr>
<td>$\chi=150^\circ$</td>
<td>0.290</td>
<td>1.8</td>
<td>-</td>
<td>5.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\chi=120^\circ$</td>
<td>0.475</td>
<td>3.7</td>
<td>-</td>
<td>4.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\chi=90^\circ$</td>
<td>0.850</td>
<td>2.9</td>
<td>-</td>
<td>4.9</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Assumption of $C_1 : C_2$ ratio for S175 ship

<table>
<thead>
<tr>
<th>Case-1</th>
<th>$C_1:C_2=1:2$</th>
<th>$C_1=4, C_2=8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-2</td>
<td>$C_1:C_2=1:1$</td>
<td>$C_1=6, C_2=6$</td>
</tr>
<tr>
<td>Case-3</td>
<td>$C_1:C_2=2:1$</td>
<td>$C_1=8, C_2=4$</td>
</tr>
</tbody>
</table>

Figure 3: Added resistance VS Froude number for S175 ship, $\lambda/L=0.5$, heading angle $\beta=180^\circ$
Figure 4: Added resistance of S175 ship in quartering seas
3.3 Bulk carrier ship case

The bulk-carrier that has been studied by Kadomatsu\cite{15} is chosen as a 2nd investigation object, because it covers the range of applications of full type ships. This ship has a $L_{pp}=285m$, $B=50.0m$ $D=18.5m$ with $C_{B}=0.829$ and $L/B = 5.7$. Based on the available 4 experimental points, the sum $C_1+C_2$ is approximated by using a weighted least square method with the result of $C_1+C_2\approx16.5$, as plotted in Figure 5. In order to determine their value respectively, they are assumed with different ratios, as listed in Table 5.

The added resistance results are shown in Figure 6 against experimental data. Generally speaking, the calculated results based on the far field method agree well with experimental data except in short wave range, where the added resistance is under predicted. For this range, the results based on the newly proposed formula with different assumptions are presented and they are very close to the available experimental data. Considering the fact that the $C_1$ and $C_2$ parameters are regressed from data corresponding to $\lambda/L=0.4$, we believe that more added resistance data in truly short waves will be necessary to find/verify the correctness of the parameters. For beam seas, the deviation is obvious; however, as the amplitude is much lower, it is of secondary importance from the design point of view.

![Figure 5: Added resistance VS Froude number for a bulkcarrier ship, $\lambda/L=0.4$, heading angle $\beta=180^\circ$](image)

**Table 5**: Assumption of $C_1 : C_2$ ratio for a bulkcarrier

<table>
<thead>
<tr>
<th>Case</th>
<th>$C_1:C_2$</th>
<th>$C_1$</th>
<th>$C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-1</td>
<td>1:2</td>
<td>5.5</td>
<td>11</td>
</tr>
<tr>
<td>Case-2</td>
<td>1:1</td>
<td>8.25</td>
<td>8.25</td>
</tr>
<tr>
<td>Case-3</td>
<td>2:1</td>
<td>11</td>
<td>5.5</td>
</tr>
</tbody>
</table>
4 SUMMARY AND CONCLUSIONS

1) The developed analytical/numerical method and computer code for the calculation of the added resistance of ships of different types sailing in head quartering seas based on the far field method, in combination with velocity potential solver NEWDRIFT of NTUA-SDL, yields satisfactory predictions, except in the short waves range.

2) For the short waves range, it appears that the newly developed semi-empirical formula, with parameter setting case-2, i.e. $C_1=C_2$, can give good prediction, i.e.

$$R_{AW}^{\text{Ref},v} = R_{AW}^{D,0} \left[ 1 + C Fn(-\cos \gamma) + C \frac{[(L-B) \sin \gamma + B]}{B} \frac{(1-C_B)/(1+\cos \gamma C_B) Fn} \right]$$

The constant $C$ should be determined by careful studying of available experimental data. In case of lack of experimental data, we recommend for fine hull form, $C=6$; for full hull form, $C=8$ is recommended. The introduction of the forward speed correction factor $[(L-B)\sin \gamma + B]/B$ indicates that viscous effects are highly correlated to the projected area.

3) Experimental data on added resistance scatter a lot and this is in the nature of this complicated phenomenon; repeating the experiments for the same conditions may lead to significant data variation, even if measurement techniques are highly accurate; hence we cannot expect an excellent agreement with theoretical predictions which rely on well-defined deterministic models.

4) The present research needs to be continued, as its outcome should be continuously improved, considering that: a) in the employed experiments there may be still motion effects, not considered in our recommended formula, b) there are not enough experimental samples. Obviously, further experiments need to be carried out to verify the validity of the proposed practical formula.
5 ACKNOWLEDGEMENTS

The presented work is a PostDoc study of the first author; the financial support of NTUA-SDL and the continuous guidance by Professor A. Papanikolaou is acknowledged. Professor G. Zaraphonitis’ contribution to the current study through many discussions is highly appreciated.

REFERENCES

An immersed boundary level-set based approach for fluid-shell interaction with impact

RODOLFO A. K. SANCHES*, ROGÉRIO CARRAZEDO* AND HUMBERTO B. CODA†

*Academic Department of Civil Engineering
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus de Pato Branco, via do conhecimento Km 1, Pato Branco PR, Brazil
e-mail: rodolfosanches@utfpr.edu.br, web page: http://www.pb.utfpr.edu.br/ecv/

†Structural Engineering Department of São Carlos Engineering School
University of São Paulo (USP)
Campus de São Carlos, Av. do Trabalhador Sãocarlense, 400, São Carlos, SP, Brazil
e-mail: hbcoda@sc.usp.br - Web page: http://www.set.eesc.usp.br/

Key words: Fluid shell interaction, Embedded coupling, large displacement, geometric nonlinear shell dynamics, Finite elements

Abstract. Fluid-shell interaction modeling is a challenging problem with application to several engineering fields. In this research we develop a partitioned algorithm for large displacements fluid-shell coupling with impact. The structure is modeled in a total Lagrangian description, using a novel shell finite element formulation to deal with geometric nonlinear dynamics of thin or thick shells. This formulation is based on the principle of minimum potential energy considering positions and generalized unconstrained vectors as nodal parameters, instead of displacements and rotations. As a consequence, the formulation eliminates the need for large rotation approximations and presents constant mass matrix, allowing the use of Newmark time integrator for the nonlinear problem. The Newton-Raphson method is employed to solve the resulting nonlinear system and contact between structures is modeled by enforcing non-penetration conditions based on a signed distance function. The flow is assumed to be compressible and the fluid dynamics solver is explicit with time integration based on characteristics. The fluid governing equations are written in the Eulerian description generating a fixed mesh method. The coupled problem is solved by using an embedded boundary technique where the fluid-shell interface is tracked inside the unstructured fluid mesh by level sets of a signed distance to boundary function. The versatility and efficiency of the proposed approach is demonstrated by selected three-dimensional examples.
1 INTRODUCTION

Fluid structure interaction problems are found in various engineering activities, such as civil buildings, mechanical devices, aeronautics, ocean structures and biomechanics. Many of these problems can be modeled as shell structures interacting with compressible flows and are a challenging field, specially if contact/impact may occur.

Mathematical modeling of mechanical problems is traditionally done in a Lagrangian or in a Eulerian description. Lagrangian description expresses the continuum medium movement in terms of the initial configuration and time, being very efficient for problems where finite displacements are the main variables, such as in solid mechanics. On the other hand, the Eulerian description is defined in terms of final configuration and time, being well used for problems where the variables are velocities instead of displacements, such as for fluid mechanics.

Both fluid and solid mechanics are involved in the study of fluid-structure interaction problems, implying the need to couple Eulerian description to Lagrangian description. One widely used way to deal with such situations is to solve the solid based on a Lagrangian description and the fluid based on an Arbitrary Lagrangian-Eulerian (ALE) description, in which an arbitrary velocity may be applied to the reference domain.

Using ALE description for Navier-Stokes equations together with some mesh moving technique is a methodology able to deal with many fluid-structure interaction problems [10, 3, 9]. However, some problems of large scale of displacements, such as air-bag or parachute deployment, will require also a remesh technique if the ALE description is employed.

Some authors have proposed immersed methods for Eulerian-Lagrangian coupling, most of them in the finite difference context, considering immersed boundary in a structured mesh [1, 4, 6].

The technique proposed here for coupling the Lagrangian shell finite element solver to the Eulerian fluid finite element solver considers the shell boundary moving inside the fluid unstructured mesh in which it is immersed. The shell position is tracked with level sets of a boundary signed distance function, and the fluid Dirichlet boundary conditions are applied by enforcing a ghost flow over the nodes immediately outside the shell boundary and, at same time, limiting the velocity slope based on the signed distance function. Shell-Shell multi-body contact is modeled by imposing non-penetration conditions based on a body to body signed distance function.

The outline of the paper consists in first briefly describe shell and fluid formulations, then describe the coupling algorithm and the impact algorithm, and finally present examples of inflatable structure problems giving a qualitative demonstration of feasibility and quality of the proposed technique.
Shell structures are solids with one dimension much larger than the others. Therefore, the mid surface serves as a reference to the solid mapping. The mappings \( f^0 \) and \( f^1 \), from an auxiliary non-dimensional space respectively to the initial and current configurations may be written as follows:

For any point out of the middle surface, its position at initial and final configuration may be written as:

\[
f^0_i = X_i = N_j(\xi_1, \xi_2)X^m_{ji} + \frac{h_0}{2} \xi_3 N_j(\xi_1, \xi_2)e^0_{ij},
\]

and

\[
f^1_i = x_i = N_j(\xi_1, \xi_2)x^m_{ji} + \frac{h_0}{2} [\xi_3 + a_j N_j(\xi_1, \xi_2)\xi^3_{s_i}] \ N_j(\xi_1, \xi_2) \bar{G}_{ij},
\]

where \( \bar{G}_{ij} \) are the nodal values (unknowns) for the generalized vector at node \( j \) at final configuration, \( h^0 \) is the initial thickness, \( e^0_{ij} \) is the \( i-th \) component of the unitary vector \( \vec{e}^0 \), normal to the middle surface at initial and \( a \) is the strain rate along thickness.

Finally, change of configuration from initial to current is represented by:

\[
f = f (X) = (f^1) \circ (f^0)^{-1}.
\]

The gradient \( A \) of the configuration change function may be expressed by:

\[
A = \nabla f = (A^1) (A^0)^{-1}.
\]

After evaluating the gradient \( A \), the Green strain tensor and the specific strain energy may be obtained, following [8]:

\[
E_{ij} = \frac{1}{2} [A_{kl}A_{kj} - \delta_{ij}] = \frac{1}{2} [C_{ij} - \delta_{ij}].
\]

The variables \( C_{ij} \) and \( \delta_{ij} \) are the right Cauchy-Green stretch tensor and the Kronecker delta, respectively. The following quadratic strain energy per unit of initial volume is adopted,

\[
u_e = \frac{1}{2} E_{ij} C_{ijkl} E_{kl}
\]

resulting into a linear elastic constitutive law relating second Piola-Kirchhoff stress and Green strain, usually called Saint-Venant–Kirchhoff elastic law, i.e.:

\[
S_{ij} = \frac{\partial u_e}{\partial E_{ij}} = C_{ijkl} E_{kl}
\]

where The \( C_{ijkl} \) are the components of the elastic constants tensor.
From preceding developments, one may write the equilibrium equation as the minimization of the energy functional as:

$$\frac{\partial U_e}{\partial x} - F + M\ddot{x} + C\dot{x} = 0,$$

(8)

where \( F \) is the external forces vector, \( C \) is the dissipative matrix and \( M \) is the mass matrix.

[2] proved that for a positional total Lagrangian description, the Newmark \( \beta \) with \( \gamma = 1/2 \) presents momentum conservative properties for most of shell dynamics problems and conserves energy for small strains if the time step is sufficiently large that the asymptotic energy convergence dominates or small enough that a uniform bound on the energy is achieved (see [5] for more details with respect to energy conservation for constant mass matrix nonlinear dynamics with the average acceleration time integration).

From Newmark \( \beta \) method, the equilibrium equation for a given instant \( s + 1 \) becomes:

$$\left. \frac{\partial U_e}{\partial x} \right|_{s+1} - F_{s+1} + \frac{M}{\beta \Delta t^2} x_{s+1} - MQ_S + CR_S + \frac{\gamma C}{\beta \Delta t} x_{s+1} - \gamma \Delta t CQ_S = 0,$$

(9)

where \( Q_S = \frac{x_s}{\beta \Delta t^2} + \frac{\dot{x}_s}{\beta \dot{\Delta} t} + \left( \frac{1}{2\beta} - 1 \right) \ddot{x}_S \) and \( R_S = \dot{x}_S + \Delta t (1 - \gamma) \ddot{x}_S \).

Equation (9) represents a nonlinear system, which we solve employing Newton-Raphson method. Each node will have 7 nodal parameters: 3 position vector components \( x_i \) with \( i = 1, 2 \) or 3, 3 components of the generalized position vector \( \bar{G}_i \) with \( i = 1, 2 \) or 3 and the strain ratio along thickness \( a \).

3 FEM FOR FLUID DYNAMICS

If there is no diffusion, the time variation of \( \phi \) over a characteristic coordinates \( x' \) is by definition null. For the Navier-Stokes equations we can write:

$$\frac{\partial \phi(x', t)}{\partial t} - Q(x') = 0,$$

(10)

where \( Q(x') \) contains all the non convective terms.

We assume the following approximation for Eq. (10) [11]:

$$\frac{\phi(y)_{n+1} - \phi(x)_n}{\Delta t} \approx \theta(Q(y)_{n+1}) + (1 - \theta)(Q(x)_n),$$

(11)

where \( x \) and \( y \) means respectively the characteristic positions at \( t = n \) and \( t = n + 1 \), \( \theta \) is a constant with value 0 for explicit solution and may be chosen larger than zero 0 and smaller than 1 for semi-implicit or implicit solution.

The product \( w\phi \) and the term \( Q(x) \) may be approximated by Taylor resulting the following expressions:
\[ u\phi(x)_n = u\phi(y)_n - (y - x) \frac{\partial (u\phi(y))_n}{\partial x} + \frac{(y - x)^2 \partial^2 (u\phi(y)_n)}{2 \partial x^2} + O(\Delta t^3), \]  
\tag{12}

\[ Q(x)_n = Q(y)_n - (y - x) \frac{\partial Q(y)_n}{\partial x} + O(\Delta t^2). \]  
\tag{13}

Assuming \( \Delta t = (y - x)/u \), from Eqs. (12), (11) and (13), and assuming \( \theta = 0 \) (explicit form), we have:

\[ \phi(y)_{n+1} = \phi(y)_n - \Delta t \left( \frac{\partial (u\phi(y))_n}{\partial x} - Q(y)_n \right) + \frac{(\Delta t)^2}{2} u \frac{\partial}{\partial x} \left( \frac{\partial (u\phi(y)_n)}{\partial x} - Q(y)_n \right) + O(\Delta t^2). \]  
\tag{14}

One important point about this procedure is that the high order terms of Eq. (14), obtained due to time integration along characteristics, introduce dissipation on stream lines direction, which as shown by [11] are equivalent to the Petrov-Galerking schemes when the time interval tends to the critical time interval, and gets smaller effects as the time interval gets larger.

Applying the procedure of Eq. (14) to the Navier-Stokes equations, one may write for momentum and energy equations one may write:

\[ \Delta (\rho u_i)_{n+1} = \Delta t \left( - \frac{\partial (u_i \rho u_i)}{\partial x_j} + \frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} + \rho g_i \right) + \frac{\Delta t^2}{2} \left( u_k \frac{\partial}{\partial x_k} \left( \frac{\partial (u_i \rho u_i)}{\partial x_j} - \frac{\partial \tau_{ij}}{\partial x_j} + \frac{\partial p}{\partial x_i} - \rho g_i \right) \right)_n \]  
\tag{15}

and

\[ \Delta (\rho E)_{n+1} = \Delta t \left( - \frac{\partial (u_i \rho E)}{\partial x_i} + \frac{\partial}{\partial x_i} \left( k \frac{\partial T}{\partial x_i} \right) - \frac{\partial (u_i p)}{\partial x_i} + \frac{\partial (\tau_{ij} u_j)}{\partial x_i} - \rho g_i u_i \right) + \frac{\Delta t^2}{2} u_k \frac{\partial}{\partial x_k} \left( \frac{\partial (u_i \rho E)}{\partial x_i} \right) + \frac{\Delta t^2}{2} u_k \frac{\partial}{\partial x_k} \left( - \frac{\partial}{\partial x_i} \left( k \frac{\partial T}{\partial x_i} \right) + \frac{\partial (u_i p)}{\partial x_i} - \frac{\partial (\tau_{ij} u_j)}{\partial x_i} + \rho g_i u_i \right)_n, \]  
\tag{16}

where all the right hand side terms are known at the instant \( t = n \).

Based on the Eulerian mass conservation equation, [11] suggest the following expression for explicit solution:

\[ \Delta \rho_{n+1} = -\Delta t \frac{\partial (\rho u_i)_{n+1}}{\partial x_i} = -\Delta t \left( \frac{\partial}{\partial x_i} (\rho u_i)_n + \theta \frac{\partial (\Delta (\rho u_i))_{n+1}}{\partial x_i} \right), \]  
\tag{17}
where $\theta$ is an arbitrary constant with value between 0.5 and 1.

Applying the Galerkin method to Eq. (15), (17) and (16), we obtain the spatial discretization and solve the resulting system getting the weak solution.

We still need to deal with the discontinuities due to the presence of shock waves, as the standard Galerkin method is unable to deal with strong discontinuities. Therefore, the artificial diffusion term based on pressure second derivative is added:

$$f_{\mu a} = \Delta t \mu_a \frac{\partial}{\partial x_i} \left( \frac{\partial \phi}{\partial x_i} \right),$$

where $\phi$ is the variable to be smoothed and $\mu_a$ is the artificial viscosity given by:

$$\mu_a = q_{dif} h^3 \left( |u| + c \right) \left| \frac{\partial}{\partial x_i} \left( \frac{\partial p}{\partial x_i} \right) \right|,$$

where $|u|$ is the velocity absolute value, $p_{av}$ is the pressure average over the element, $q_{dif}$ is an user specified coefficient taken between 0 and 2, $c$ is the sound speed and $h$ is the element size [11, 7].

4 IMMERSED FLUID-STRUCTURE COUPLING PROCEDURE

The proposed method for enforcing boundary conditions on the vicinity of a shell immersed in an unstructured fluid mesh requires all the fluid elements close to the boundary $\Gamma_s$ to be identified, and to know if they are inside or outside the physical domain $\Omega_f$. To this end, a computationally efficient and scalable approach is to use a signed distance function (or, level set function):

$$\Phi(x, \Gamma) = \begin{cases} 
\text{distance}(x, \Gamma) & \text{if } x \in \Omega \\
0 & \text{if } x \in \Gamma \\
-\text{distance}(x, \Gamma) & \text{otherwise}
\end{cases}$$

whose zero-th level set determines the resulting body shape.

In contrast to the usual parametric mesh based boundary representations (using segments or facets), level set based representations are more suitable for problems with large deformations and topology changes. There are efficient and scalable algorithms for converting a mesh based representation into an implicit representation.

Next, all the fluid elements are tagged as physical, fictitious or boundary depending on their position with respect to the physical domain. This classification is performed by computing for each fluid element $\Omega_{ef}$ the minimum and maximum signed distance $\min \Phi(\Omega_{ef})$ and $\max \Phi(\Omega_{ef})$, respectively, then the classification is applied as (see Fig. 1):

- physical element: $\min \Phi(\Omega_e) > 0$;
- fictitious element: $\max \Phi(\Omega_e) \leq 0$;
• boundary element: neither a physical nor a fictitious element.

The purpose of the element tag is to identify which nodes and elements should be deactivated from the analysis as well as the ghost nodes. To this purpose, all the fluid mesh nodes are tagged as active or inactive. A node $k$ is active if $\Phi(k) > 0$ or if $k$ bellows to some boundary fluid element, and inactive otherwise. This tags are computed for each time step and the inactive nodes as well as the fictitious elements are deactivated from the analysis.

For each fluid node $k$, we find the closest point $l$ on shell mesh, and store the shell element $\Omega_{se}$ for which $l \in \Omega_{se}$ and the non-dimensional shell coordinates $(\xi_1, \xi_2)$ for point $l$.

The active nodes $k$ outside the physical domain ($\Phi(k) < 0$) need to be populated. For this purpose we project the point $k$ to the closest physical element determining a new point $m$ from where the values of density, specific energy and momentum are linearly extrapolated.

A way to prescribe the velocity at the boundary position would be to change the velocity nodal values of the active nodes $k$ with $\Phi(k) < 0$ (ghost nodes) in order to modify the values over the boundary. However this procedure may imply on very large velocity values as $\Phi(k)$ becomes close to the element size.

To avoid this problem, we modify the velocity nodal values for the active nodes outside the boundary and also the velocity nodal inside a strip of width $\delta$ according to the following equation if the flow is inviscid:

$$u_f = u_f + (1 - \Phi) \left[ (u_s - u_f) \cdot n \right] n.$$  \hfill (21)

or, for a viscous flow:

$$u_f = u_f + (1 - \Phi) \left[ u_s - u_f \right]$$ \hfill (22)

where $u_f$ is the fluid nodal fluid velocity vector, $u_s$ is the shell velocity vector evaluated at the shell closest point to the fluid node. The term $(1 - \frac{\Phi}{\delta})$ limits the slope of velocity on
direction normal to the boundary but also introduces an artificial stiffness to the problem. However if we adopt a \( \delta \) equal to the element size, this artificial stiffness is equal to the one naturally produced by an mesh of same elements size adapted to the boundary.

Taking advantage of the fluid shape functions, the stress tensor may be evaluated directly over the position of the embedded shell nodes \( k \) or directly over the shell quadrature points. the shell loads with respect to the Cartesian axes are given by:

\[
q_{kj} = [-\tau_{jl} n_l - p n_j]_{P_k},
\]

where the indexes \( j \) and \( l \) represent Cartesian direction and \( n_l \) is the \( l \) component form the normal vector to \( \Gamma_s \).

5 SHELL-SHELL CONTACT

Contact between structures is modeled by enforcing non-penetration conditions based on a body-body signed distance function.

Each Newton-Raphson iteration, over each body \( i \), we calculate the signed distance to the other bodies \( k \) nodal values \( \Phi_{si}(k) \). This value is positive if the node did not cross other bodies, regarding its initial position, or negative if it crossed.

If the value is negative, the node position is projected back a distance of \( \Phi_{si}^2 \) along its normal direction (slip wall contact). this procedure is repeated after each Newton-Raphson iteration until reach the prescribed error.

6 NUMERICAL EXAMPLE

In this example a simulation of an airbag deployment and crash with an half sphere moving on the airbag direction. We consider it as a qualitative example, once due to computational the airbag mesh is not fine enough to represent the wrinkles that appears
in high frequency and also because the formulation is not yet ready to simulate self contact, what is common in a problem like this. The airbag on its flat initial condition is filled with an ideal gas at rest with density $\rho_f = 1.3\, \text{kg/m}^3$ pressure $p = 107,34\, \text{kPa}$ and specific heat ratio $\gamma = 1.4$. A gas with $\rho_{fi} = 10\, \text{kg/m}^3$, sound speed $c = 370\, \text{m/s}$ and $\gamma = 1.4$ enters the airbag producing a shock wave.

The input condition is kept constant until $t = 0.06\, \text{s}$, when the input is closed and the applied boundary condition is that of slip wall. The fluid mesh where the airbag is immersed has 263667 elements and 47491 nodes.

We discretize 1/4 of the problem assuming that the problem is symmetric according to the planes xz and yz and. The airbag is discretized by 258 elements and 1237 nodes and its material has Young’s modulus $E = 3\, \text{GPa}$ and specific mass $\rho_s = 1000\, \text{kg/m}^3$ and thickness $h = 0.5\, \text{mm}$. The abag is clamped over all the input area and simmetry boundary conditions are applied to the planes xz and yz.

A half sphere with specific mass $\rho_{s2} = 3000\, \text{kg/m}$, thickness $h = 15\, \text{mm}$ and Young’s modulus $E = 20\, \text{GPa}$ is initially positioned at $z =$ and moves on the airbag direction with a speed $w = -25\, \text{m/s}$. The 1/4 of the half sphere is discretized by 6 curved elements and 37 nodes.

The simulation produced results according to the expectations. Figure 4 plots the top displacement versus time and figure 5 presents a snapshots of pressure distribution and the airbag deformation for some instants.

From these results we conclude that the present procedure is a robust method for analysis of inflatable structures and should further be improved to enlarge computing capabilities and shell self contact.

7 CONCLUSION

We have proposed one numerical model for analysis of shell high-speed flows coupling and for impact between shell structures. We presented the shell solver, which is able to deal with geometrical nonlinear dynamics of shells and uses a methodology based on the minimum potential energy theorem written regarding nodal positions and generalized unconstrained vectors, not displacements and rotations, avoiding the use of large rotation
approximations. The resulting time integration is stable for problems like the example presented here and based on the Newmark method due to the presence of constant mass matrix. Finally we developed the coupling procedure. The immersed approach furnishes a general algorithm for explicit coupling of Lagrangian shell solvers with unstructured-mesh-based Eulerian fluid solvers considering the shell immersed in a block of unstructured fluid mesh. The coupled algorithms are tested by one selected example. The employed fluid and shell solver showed to be robust and completely adequate for simulating fluid-shell interaction with impact. Further improvements on computer capability and shell self contact are recommended.

Acknowledgments

The authors would like to thank the Brazilian council CNPq (Conselho Nacional de Desenvolvimento Cientifico e Tecnologico) and the State of Paran Foundation Fundao Aracaria for the financial support.

REFERENCES


Figure 5: Pressure snapshots


Institute of Technology, Pasadena, California, 2009.


EFFECTS OF SLIP CONDITION ON FLOW NEAR THE SURFACE OF HYDRO-GEL

M. YOKOYAMA*, Y. KUBOTA†, K. KIKUCHI†, O. MOCHIZUKI† AND G. YAGAWA†

* School of Information Science, Meisei University
2-1-1 Hodokubo, Hino, Tokyo, 191-8506, Japan
e-mail: masao.yokoyama@meisei-u.ac.jp

†Faculty of Science and Engineering, Toyo University
5-28-26 Hakusan, Bunkyou-ku, Tokyo, 112-8606, Japan
Email: yagawag@gmail.com

Key words: Slip condition, Hydro-gel, Particle method.

Abstract. The influence of surface properties of a wall of the body on the flow behavior is investigated numerically by the Moving Particle Semi-implicit (MPS) method. Considering the difference between the no-slip condition and the with-slip condition of a wall, a new method to calculate the flow near the hydro-gel wall and that inside the hydro-gel is proposed in this paper.

1 INTRODUCTION

Most of fluid simulation has been calculated under the boundary condition which is no-slip condition between fluid and structure. The wall is usually assumed in numerical simulation as that water does not permeate and it satisfies non-slip condition. But, though the non-slip condition in numerical simulations expresses the existence of surface of a body, it cannot describe the surface properties. Namely, it means that we cannot distinguish the difference between the qualities of the iron and the skin of an animal in water.

The problem of interface between solid and liquid is a significant topic from the viewpoint of water-repellent and drag reduction for material design of ship, car, micro channel and so on. Manservisi et al.[1] dealt with the interaction between a droplet and the surface of a wall and numerically simulated the spreading of a single droplet impacting over horizontal dry surfaces. Many studies focused on the contact angle of a droplet on a solid surface [2-4], and the surface tension is also well studied with numerical system as the problem of multiphase flow[5,6]. Rebouillat et al. [7] and Idelsohn et al. [8] calculated the sloshing of a liquid filled partially in a container as problems of solid and fluid interaction. They used the Finite Element Method (FEM) to calculate the free surface with large deformation and many droplets. Alam et al. [9] indicated that the surface tension of a splash becomes prominent in small-scale phenomena.

From the viewpoint of the field of bionics, the interaction between flow and object such as
a cell or living things was studied [10,11]. Kikuchi et al. [11] investigated the flow on a hydrogel surface as strong hydrophilic material to clarify the effect of the surface flow. They found the slip flow on the hydrogel surface, and determined the relation between the strength of the hydrophilicity and the slip velocity. And, creatures living in water such as frog or fish have a slimy mucus skin, and the principal ingredient of mucus is a hydrogel known as mucin[12].

In order to calculate free surface of water, the Moving Particle Semi-implicit (MPS) method is useful. Some studies were reported for splash and droplet dealing with free surface flow [13-15]. However, they ignored surface properties related to different materials, so the effects of a surface on splash are not clear yet. Therefore, since there is no distinction of a surface condition, the form of the splash by the hydrogel object and the splash by the acrylic resin object becomes the same. The purpose of the present paper is to discuss how to express the difference of the surface conditions when using the MPS method. Introducing the interaction models between flow and wall in order to describe the wall conditions, we consider the effect of the slip ratio as hydrophilicity in our numerical simulation.

2 SWELLING DEGREE AND SLIP RATIO

The governing equation of a flow is the incompressible Navier-Stokes equations as the following,

\[
\frac{Du}{Dt} = \frac{1}{\rho} \nabla P + \nu \nabla^2 u
\]

Here, \( u \) is the velocity vector of fluid, \( \rho \) is the density of fluid, \( P \) is the pressure, \( \nu \) is the kinematic viscosity of fluid. The \( F \) is added to the Navier-Stokes equation as the external force such as gravity.

Employed in the present paper is the MPS method, which is one of the particle methods. Assuming two particles \( i \) and \( j \), which possess scalar quantities of pressure \( p_i \) and \( p_j \), the gradient model between these two particles is written as

\[
\nabla p_i = \frac{d}{n^0} \sum \frac{p_i - p_j}{|\vec{r}_i - \vec{r}_j|} (\vec{r}_i - \vec{r}_j)|(|\vec{r}_i - \vec{r}_j|)
\]

Here, \( d \) is the number of the space dimension. Parameter \( n^0 \) is called particle number density in MPS method. Since each water particle has the same mass, every particle number density should be constant and equal to \( n^0 \).

The Laplacian model in MPS method is written as

\[
\nabla^2 u_i = \frac{2d}{\lambda n^0} \sum (u_j - u_i) \kappa (|\vec{r}_j - \vec{r}_i|)
\]

where \( \lambda \) is a parameter which is introduced in order to coincide the statistical distribution with an analytic solution [13]. In the MPS method, the interaction between particles is evaluated by
weight function $\kappa$ as follows,

$$\kappa(r) = \begin{cases} \frac{r_e}{r} - 1 & (0 \leq r \leq r_e) \\ 0 & (r_e < r) \end{cases}$$

(4)

Here, $r$ is the distance between two particles and $r_e$ is the cut-off radius.

The swelling degree is employed here as the slip ratio for treating the hydrophilic surface of an object, which is related to the weight of water included in a hydrogel [16], defined as the ratio of the weight of water against that of agar as follows,

$$S = \frac{m_{\text{water}} + m_{\text{gel}}}{m_{\text{gel}}}$$

(5)

where $m_{\text{water}}$ is the weight of water and the $m_{\text{gel}}$ is the weight of hydrogel.

The increase of $S$ means the water contained in the agar increases. Agar is a hydrogel, and is convenient for making a suitable shape and it's easy to control its degree of swelling. Agar has the effect to increase the slip on the surface of an object. The difference of the velocity of water flow versus $S$ was investigated by Kikuchi and Mochizuki [11], who measured the flow velocity $u$ at the distance $y$ from the surface of the acrylic slope and the agar slope as shown in Fig. 2. The $\tau$ is the wall shear stress on the no-slip condition, and $\tau'$ is that on the agar slope. The coefficient $\alpha$ is defined as the parameter on the slip condition wall as follows,
\[ \alpha = \tau' / \tau \quad (6) \]

Figure 3 shows the experimental relation between the $S$ and $\alpha$. The solid line is the observed data, and the dotted line is an estimated data. The $\tau'$ was smaller than $\tau$, since $\alpha$ decreased with the increasing of $S$. This relationship can be expressed as

\[ \alpha = 1 - \beta S \quad (7) \]

where $\beta$ is estimated to be $3.7 \times 10^{-4}$. It is noted that larger $S$ gives more slip on the surface. When the boundary condition is complete-slip, $\alpha$ is 0. This means that shearing stress decreases by about 10%, when $S$ is 250.

As shown in Fig. 4, the interaction as the shear force between each particle is evaluated by weight function $\kappa$. The range, which is the effect of weight function acts, is $r_e$. The shear force acting between wall and liquid is calculated by the Laplasian term of Eq.(3). Namely, $\alpha$ is multiplied to the $\kappa$ only for the water particles near the boundary of hydrogel wall, because the effect of slip is caused as the shear force near the hydrogel wall. Thus, Eq. (3) is rewritten using $\kappa_H(r) = \alpha \kappa(r)$ as follows,

\[ \nabla^2 u_i = \frac{2d}{\Delta n} \sum_{j} (u_j - u_i) \kappa_H(|\vec{r}_j - \vec{r}_i|), \]

\[ \kappa_H(r) = \alpha \kappa(r) \quad (0 \leq \alpha \leq 1) \quad (8) \]

\[ i: \text{water particle near hydrogel wall}, \quad j: \text{surface particle of hydrogel wall} \]

Namely, the way for adjusting the slip effect to the calculation in the present study is as follows: i) select the $S$ according to the hydrophilicity of a target object, ii) estimate the $\alpha$ using Eq.(7), iii) apply the $\alpha$ to the weight function of Laplacian term for the calculation of shear force near the hydrogel wall.

### 3 SIMULATION RESULT

The MPS simulation system was developed in three dimensional domain. The liquid was assumed to be water, namely the density of water $\rho_{\text{water}}$ was 1000 kg/m$^3$. The radius $R$ of sphere was 10 mm assumed as small living things. The buoyancy of sphere is considered in the calculation. The range of $S$ was from 1 to 350 in this simulation. The cut-off radius $r_e$ was 2.1.
The initial distance \( l_0 \) between particles was \( 4 \times 10^{-3} \)m. The benchmark test by the collapse of liquid columns was performed in order to confirm the influence of \( l_0 \) on simulation result. The collapse of liquid columns is commonly used for the evaluation of the program of MPS method. With the difference of \( l_0 \), there was little difference in the velocity of collapsing speed and in the form of collapse. The formation of collapse was also compared with the experimental result[31]. We confirmed the formation of water collapse was similar to that of experimental result.

3.1 Flow on hydro-gel wall

Figure 4 shows the profile of flow velocity on the hydro-gel wall of a conduit calculated by our simulation. The distributions of velocity of water in the \( x \) direction on the no-slip wall and with-slip wall (\( S = 100 \) and 250) are shown in Fig.4. Calculation domain is the length = 80 \( l_0 \), the height 20 \( l_0 \), and the width = 10 \( l_0 \). The upper surface of water is a free surface and is given velocity \( u_0 = 4 \) cm/s to the \( x \) direction as initial velocity. In addition, the distribution of \( u \) is normalized by \( u_0 \), and height \( y \) is normalized by \( h \), respectively. The distribution was obtained on the center of bottom wall.

About the simulation result, as it is known the flow under no-slip conditions becomes the Poiseuille flow as shown in the dashed line by the theoretical value, the velocity distribution of no-slip condition flow by our calculation is almost correct to that theoretical value. From the enlargement figure shown in the right bottom of Fig.4, the flow velocity near the surface of hydro-gel wall increased by the effect of slip on the hydro-gel surface with the increase of a swelling degree. For example, the flow velocity near the wall in the with-slip (\( S = 250 \)) condition was increasing about 10\% from that of no-slip conditions, namely, it was same as the above-mentioned experimental result. The gradient of velocity near the hydro-gel wall is decreasing by the increase of the swelling degree. Since the wall stress was proportional to the velocity gradient near the surface as shown in the formula (6), this result shows that the wall
stress of surface was reduced. By introducing the slip ratio near the surface of a wall, the calculation result which was same as the experimental result was obtained.

3.1 Flow inside of hydro-gel

The simulation schema of the flow inside of hydro-gel is shown in Fig.5. The bottom wall is the fixed wall without slip, and the layer of hydro-gel with the height $h$ is placed on it and the upper plate is placed on the top of hydro-gel. The upper plate was moved at the fixed speed $U$. The hydro-gel particles were arranged at equal intervals $l_H$ in the distance which was decided in inverse proportion to the degree of swelling by the primitive cubic lattice. Between the lattice the water particle was arranged uniformly, and calculated the motion of the particle of inside hydro-gel. Slip ratio $\alpha$ of the shear force between a hydro-gel particle and a water particle was applied. The rate of the number of particles and that of water particle was same as the swelling degree. The simulation result about the flow velocity distributed in the distance $y$ from a floor in hydro-gel is shown in Fig.6. The flow in hydro-gel was induced by movement of the top plate of a wall, and the profile of velocity was parabola. This result showed same tendency with the experiment result.

12 CONCLUSIONS

The influence of slip condition on the flow near the wall was simulated by the Moving Particle Semi-implicit (MPS) method. Considering the difference between the no-slip condition and the with-slip condition of object’s wall, a new method to calculate the flow near the hydro-gel wall and inside of hydro-gel is proposed in this paper. From the viewpoint of the field of bionics, the wall condition is an important topic of interaction between the fluid dynamics and the movement dynamics of living things such as frog or fish which have a slimy mucus skin.
Introducing the slip ratio according to the swelling of hydro-gel to the MPS method, the difference of flow near the wall between the no-slip condition and the slip condition such as hydro-gel was investigated numerically. As the result, the water flow in a conduit with the hydro-gel wall obtained by our simulation showed good agreement with the experimental result, which the velocity near the hydrogel wall was higher than that of no-slip wall. Furthermore, we applied this method to the simulation of the flow inside of the hydro-gel, which the flow became a parabola to the depth and the flow was similar to the experimental result.

REFERENCES


[16] Alupeia IC, Marcel Popa M, Hamcerencu M, Abadiec MJM. Super absorbant hydrogels
FSI OF HIGH PERFORMANCE HIGH-LIFT DEVICES WITH CIRCULATION CONTROL VIA CONDITIONED COANDĂ-JETS

KAY SOMMERWERK, MATTHIAS C. HAUPT AND PETER HORST

Institute of Aircraft Design and Lightweight Structures (IFL)
TU Braunschweig
Herrmann-Blenk-Str. 35, 38106 Braunschweig, Germany
e-mail: k.sommerwerk@tu-braunschweig.de, web page: http://www.tu-braunschweig.de/ifl

Key words: High-Lift, Coandă effect, Circulation Control, Aerodynamic performance, Fluid-structure interaction

Abstract. Current transport aircraft are limited to airports with comparatively long runways for take-off and landing. An aircraft with short take-off and landing capabilities is under investigation at the Collaborative Research Center 880. The aircraft employs circulation controlled high lift devices where high velocity air is blown through a slot in front of the flap. These high performance high-lift devices allow take-off and landing of the aircraft at runways of 800 m length. The curved flap leading edge induces a Coandă effect with the jet resulting in an attached flow up to the tip of the flap even at high deflection angles of up to 85°. Examination of the aeroelasticity of the wing is of high importance because of the sensitivity of the Coandă effect to perturbation through deformation and consequential change in flow.

The large pressure gradients can play a significant role in the effective use of this jet system. Preliminary studies have shown an influence on aerodynamic performance due to slot deformation. Small changes in the aerodynamic characteristics can have adverse effects on the stall behavior. A flap section model of the wing is used to analyze the performance for several flight states. This high detail model allows capturing fine effects over the whole wing chord and on the slot region while still including wing deformations.

The aerodynamic performance of the aeroelastic flap section model is compared to the characteristics of the rigid airfoil. The analyzed flight states give an insight into the influence of the deformation on the flow. Additionally the effects of jet momentum variation on aerodynamics is shown and the dominant stall phenomena presented. The local relative change in pressure can reach values between 10 and 20 % and has an influence on the stall behavior of the section. The change in aerodynamic performance illustrates the influence of small deformations on the sensitive circulation control.
1 INTRODUCTION

Airfoils employing high lift systems with circulation control have shown very high lift coefficient values of up to 5.2 for 2D analyses [1]. The collaborative research center SFB 880 investigates an aircraft which is distinguished through short takeoff and landing capabilities by making use of the Coandà effect. High velocity air is blown through a slot directly upstream of a deflected plain flap. This jet interacts with the boundary layer and induces a Coandà effect which keeps the flow attached to the high-lift device allowing high deflection angles of up to 85°. Figure 1 displays the 100 passenger high wing monoplane with an engine bleed air system to generate the pressurized air for the flap jets. Other configurations that integrate compressor systems for each flap are also under investigation. The aeroelastic phenomena of wings with blown air configurations are of high importance to the overall aerodynamic performance of the aircraft, as they can deviate from the phenomena of conventional wings. Different stall phenomena have been identified which depend on the jet performance. The jet performance is dependent on the deformation of the duct system due to the pressurized air in the duct as well as due to the wing deformation because of aerodynamic loads. The slot height in comparison to the overall model dimensions lies in the order of magnitude of $10^{-3}$ therefore requiring high resolution structural and aerodynamic models for the analyses.

The presented study gives insight into the overall aerodynamic performance of a flap sectional model under aeroelastic behavior. The stall phenomena of a Coandà flap are presented and the different jet performances for static airfoils are shown. One configuration from previous aeroelastic studies has been chosen for analyses of the stall behavior [3]. The flap sectional model is deformed via CFD-CSM coupling and the resulting performances are computed for several configurations with varying angle of attack and momentum coefficient.

2 MODELS

2.1 Structure

A flap sectional model as displayed in Fig. 2(a) is used for high detail analyses. Because of the low taper ratio of 0.38 of the aircraft wing the flap sectional model is untapered, which has a negligible effect on the accuracy on the analyses. The mean aerodynamic chord of the wing of 3.428 m is used as the chord length of the model. The flap width is 2.142 m and the slot height is 0.061 % of the chord length. The model is fixed with
connecting boundary conditions on the inside rib. The cutting loads of the unrepresented outer wing are applied to the outboard rib. The model uses the external duct integration configuration with a slot stiffener distance of 102 mm [3]. A layered composite is employed for the material model and sizing is performed based on a fully stressed approach adapted for composites [3]. The aerodynamic surface loads are computed via CFD analyses and are transferred to the structure via the ifls-coupling environment [4]. The ANSYS solver is used for the structural analyses.

2.2 Aerodynamics

The 2D section of the fluid grid including a slot detail is shown in Fig. 2(b). It is a hybrid grid designed to capture the aerodynamic effects on the flap and includes a concave section for the air duct system. The grid is extruded in the spanwise direction 160 times for the required resolution. The extruded grid consists of 28 million points with 85% of cells in the structured boundary layer. The aerodynamic computations are conducted with the TAU code 2011.2 of the German Aerospace Center (DLR), to solve the Reynolds-averaged Navier-Stokes equations (RANS) [5].

2.3 Coupling

The fluid structure interaction (FSI) process uses a partitioned approach to solve the fluid and structural problem. After solving the fluid problem for a given displacement the resulting boundary conditions are applied to the structure grid, which in turn alters the fluid grid through its computed deflection. In detail the pressure distribution on the wetted surface after reaching steady state in the CFD analysis is integrated to obtain the forces on each node of the fluid grid. These forces are transferred conservatively to the structure resulting in force vectors on the structural nodes. A FEM simulation is performed to obtain the nodal displacements, which can be used for stress-strain analyses.
Figure 3: Flap section coupling process, (a) pressure distribution of the wetted aerodynamic surface, (b) integrated force vectors, (c) transferred force vectors on structural grid, (d) structural nodal displacements or fluid grid deformation.

Figure 3 illustrates the coupling process over the span of the flap section. Figure 3(a) illustrates the pressure distribution of the wetted aerodynamic grid which is then integrated to the resulting force vectors shown in Fig. 3(b). The forces are transferred to the structural grid via the coupling environment resulting in the load vectors on the structural grid Fig. 3(c). Solving the finite element model with the imposed loads results in the nodal displacements in Fig. 3(d) used for sizing and mesh deformation.

3 Coandă effect and momentum coefficient

The circulation controlled high lift system makes use of the Coandă effect to mitigate flow separation at high flap deflection angles. The desired effect is an interaction of a blown jet directly upstream of the flap with the oncoming flow. The boundary layers intermix and the oncoming flow is accelerated. The high velocity jet flows by the curved flap leading edge, where a pressure gradient is induced because of the Coandă effect. Due to the low pressure on the flap surface the jet stays attached and is bend around the flap leading edge. The jet characteristic can be expressed dimensionless with the dynamic pressure of the flow to attain the momentum coefficient $c_u$. 
It is defined as

\[ c_\mu = \frac{\dot{m}_{\text{jet}} v_{\text{jet}}}{S_{\text{ref}} q_{\text{inf}}} \]  

(1)

where \( \dot{m}_{\text{jet}} \) is the mass flow through the jet exit section and \( v_{\text{jet}} \) is the velocity of the jet at the slot. \( S_{\text{ref}} \) is the reference surface, which is the chord length multiplied by the span-wise flap length. The dynamic pressure \( q_{\text{inf}} \) is defined as

\[ q_{\text{inf}} = \frac{1}{2} \rho_{\text{inf}} v_{\text{inf}}^2 \]  

(2)

with the density \( \rho_{\text{inf}} \) and the velocity \( \rho_{\text{inf}} \) of the flow.

The momentum coefficient is used as an indicator of jet performance. With an increase in \( c_\mu \) the flow separates later from the flap resulting in increased lift. Figure 4 illustrates the jet performance to lift correlation. The increase in lift is linear to the increase in \( c_\mu \) up to the point where the flow is attached on the entire flap. This point is optimal with respect to power consumption to generate the jet and increase in lift. Further increases in \( c_\mu \) cause considerably lower lift increments. The flow around the airfoil with optimal \( c_\mu \) is illustrated in Fig. 5 for a flap deflection of 65°, an angle of attack of -2° and a \( c_\mu \) of 0.06. The jet intermixes with the oncoming flow and has enough momentum to reach the trailing edge. Leading edge stall for this configuration occurs at angles of attack of 3°. Note the stagnation point at about 20 % chord for the airfoil. Droop nose configurations are under investigation which move the stagnation point closer to the leading edge, thus raising the stall point 15°.

4 RESULTS

4.1 Stall phenomena

Because of the additional parameter \( c_\mu \) the stall phenomena of an circulation controlled airfoil are more complex. An increase in \( c_\mu \) does not always have a positive effect on lift and the result depends also on the angle of attack \( \alpha \) and the flap deflection angle \( \delta \). Considering a fixed flap deflection angle of 65° the occurring stall behavior is shown in
Kay Sommerwerk, Matthias C. Haupt and Peter Horst

Figure 6: Stall phenomena of Coandă flap illustrated with stream lines and Spalart-Allmaras eddy viscosity

Fig. 6(a)–(c). Figure 6(a) shows the case of low momentum coefficient and is designated as type I stall. Here the jet and the outer flow mix effectively. However the flow separates before reaching the flap trailing edge as it has not enough momentum to overcome the adverse pressure gradient. With an increase of $c_\mu$ into the region of supercirculation the flow detaches at the leading edge of the airfoil thus leading to stall at lower angles as shown in Fig. 6(b). This is due to the increase in boundary layer thickness, which can not be caught entirely by the jet. Additionally, the large pressure gradient on the leading edge induced by the Coandă effect results in separation. Only small parts of the detached flow can be reattached to the flap. This stall is denoted as type II. A complete flow separation is denoted Type III stall. Here only a very thin part of the boundary layer stays attached. The jet momentum is to large to adequately mix with the boundary layer which changed with the angle of attack. Most of the flow detaches at the Coandă surface. This stall type does not necessarily include leading edge stall. Note in Fig. 6(c) the streamlines from the flap trailing edge up to the Coandă surface due to the adverse pressure gradients.

These phenomena show the sensitivity of the flow to the jet characteristics. Variation of the airfoil and slot geometry due to aeroelasticity not only influence the outer flow but can also affect the momentum coefficient as the jet velocity decreases with constant mass flow at higher slot cross sectional area.

4.2 Rigid airfoil characteristics

For reference the aerodynamic performance of the undeformed airfoil was evaluated for different angles of attack and different moment coefficients. The results are presented in Figure 7(a)–(c), with variable scale for the eddy viscosity computed with the on equation Spalart-Allmaras turbulence model [6]. The maximum lift coefficient $c_{L,max}$ is attained at the maximum $c_\mu$ at an angle of attack of 1°. It is evident, that for decreasing $c_\mu$ the value of $c_{L,max}$ decreases. Additionally the angle at which $c_{L,max}$ is attained increases.
The stall behavior for high $c_\mu$ values are first type II and then type III. It can clearly be seen that at higher angles of attack, the flow might briefly reattach itself. This is the range where the stall changes from type II to III. At stall the lift coefficient drops to values around 2. The increase in drag and decrease in pitching moment as depicted in Fig. 7(b) and Fig. 7(c) is evident. Decreasing the moment coefficient reduces the induced pressure gradients at the leading edge which allows for later occurrence of type II and III stall. However parts of the flow detach before reaching the flap trailing edge because of the adverse pressure gradients. Part I stall does not have the severe effects as part II and III as the lift coefficient is only decreased marginally. The increase in drag and the decrease in pitching moment are fitting for the early separation and decreased lift.

4.3 Deformed airfoil

The deformation of the flap section due to the aerodynamic loads leads to a change in the pressure distribution on the section. Exemplary data for one x-z-slices of the profile at a spanwise location of 1.2 meters is shown in Fig. 8(a)–(d). Figure 8(a) displays the
profile contour and Fig. 8(b) shows the respective displacement vector multiplied with the surface normals. The normals are computed on the surface elements and transferred to the grid points. Because of the change in sign of the normal vector, the displacement of the lower skins are depicted negative. Note the relative small displacements compared to the models dimensions, which are due to high stiffness of this section because of the inside position of this section on a real wing. From the leading edge to the spoiler at 2.1 m the model has a uniform displacement. It is slightly rotated around the y-axis. The higher offset points between 2.1 m and 2.55 m are the spoiler displacement and the displacement of the duct sheets due to the large jet outlet pressure. Again note the inversed normal for the lower sheet. The flap displacement coherently offset due to the rotation around the attachments to the wing box. Stray points are due to the interpolated normals on edges and corners of the airfoil.

The corresponding pressure distribution over the flap is shown in Fig. 8(c). Notable are the low pressures at the suction tip and over the Coandă surface. The mean $c_p$ over the top surfaces is close to -4. The lower side pressure coefficient has an almost constant value of 1. Notable again is the very high pressure coefficient of around 70 inside the jet duct.
Figure 8(d) displays the change in $c_p$-distribution due to the deformation. An overall change of -0.1 is recognizable on the lower surface including the flap. While the low values of pressure coefficient are reduced on the Coandă surface, the pressure is lowered at the nose. The relative change in pressure coefficient is about 10% for the lower surface and leading edge, which linearly transitions to 0% at the spoiler. The difference linearly increases to 24% up to 2.9 m on the flap top surface.

It is evident, that the small deformations of a rather stiff section have a noticeable impact on the pressure distribution and thus the aerodynamic performance. A variation in pitching moment is also evident through the distribution differences shown in Fig. 8(d).

**Flexible section characteristics**

The aeroelastic effect on the aerodynamics is investigated for the three highest momentum coefficients for several selected angles of attack. The angles where chosen around the interesting stall points. The lower momentum values in the domain of sub circulation are of low interest since stall type II and III occurs at higher angles than considered.

![Graphs](image)

**Figure 9**: Aerodynamic coefficients $c_L, c_D, c_M$ as a function of $\alpha$ and $c_\mu$ for aeroelastic section
The results of the integrated variables are shown in Fig. 9(a)–(c) including the values of the rigid grid for \( c_\mu = 0.06 \). Examining the lift for \( c_\mu = 0.06 \) in the domain of super circulation as shown in Fig. 9(a) stall occurs at a lower angle of attack which is due to earlier leading edge separation.

This is confirmed by higher lift values of about 3 \%, which induce early leading edge stall. The pitching moment coefficient reflects the stall behavior, though the same values are reached at later angles than on the undeformed model. While the drag before is double the drag of the undeformed airfoil, the drag is about 10 \% higher on the deformed model during stall. This increase can be explained by the expansion of the wake, which is due to the splitting of the attached flow into an attached and detached part. This splitting occurs after the flow bent around the Coand\text{á} surface. One part stays attached and one part separates expanding the wake as shown in Fig. 10.

For the two other configurations which are in the nominal circulation range stall is delayed for one degree. The lift is increased in the same range as for \( c_\mu = 0.06 \) before stall. Stall occurs with a partial separation on the flap leading to higher lift over values of 3. The partial separation results in higher drag values during stall. The pitching moment reflects this behavior by reaching higher negative coefficients due to the existing lift on the Coand\text{á} surface.

It is evident, that even small deformations of the section have an impact on the variation of integrated aerodynamic variables. The overall change in lift is negligible, while the increase in drag due to partial separation over the flap as well as the change in pitching moment needs possibly be taken into account for flight dynamics. The stall point varies around 1 \% illustrating the sensitivity of the circulation control to small airfoil variations. The stability changes depending on the moment coefficient and thus the circulation control domain. It is important to fully understand the change of stall points and phenomena in all three circulation control domains to be able to adapt the flight dynamics strategies and flight envelope accordingly.

5 CONCLUSIONS

In this paper we presented the influence of an elastic Coand\text{á} flap section on aerodynamic performance. A coupled flap section model was used to determine the deformations of the fluid grid for several momentum coefficients and angles of attack. Stall phenomena and aerodynamic characteristics of an rigid airfoil with blown air system were explained.
MATHEMATICAL MODELING OF HYDRODYNAMIC PROCESSES IN GEOTHERMAL PLANT

MARIJONAS BOGDEVIČIUS *, JOLANTA JANUTĖNIENĖ †, SAULIUS RAZMAS †
AND MINDAUGAS DRAKŠAS †

* Mechatronics Science Institute
Klaipeda University
Bijunu st. 17 - 103, LT-91255 Klaipeda, Lithuania
e-mail: mmi@ku.lt, web page: http://www.ku.lt/mmi/

† Department of Mechanical Engineering
Klaipeda University
Bijunu st. 17 - 103, LT-91255 Klaipeda, Lithuania
Email: mik.jtf@ku.lt- Web page: http://www.ku.lt/jtf/mechanika/

Key words: Geothermal Energy, Centrifugal Pump, Hydrodynamic Processes.

Summary. Researches the physical properties of the mixture of water underground and gases is investigated in this paper. Also paper analyzes an existing geothermal energy extraction system, which consists of a long pipe system and a large number of hydraulic and mechanical elements. The mathematical model of the geothermal system “depth centrifugal pump-pipe (pipe system)” was made to assess the characteristics of an extraction depth centrifugal pump and their correlation with impeller frequency, discharge and other system parameters. Based on this mathematical model the instability zones of the system were determined. A universal mathematical model of depth centrifugal pump was made. This model could be used to describe hydrodynamic processes of injection pumps in the pipe systems, where large amounts of released gases influences the productivity of the pump and geothermal energy extraction system overall.

1 INTRODUCTION

One of renewable energy sources is geothermal energy. Scientific research on geothermal energy is aiming for refinement of extraction technology and expansion of the field of use such as centralized heating or generation of electricity. An investigation of the unsteady flow behaviour centrifugal pump is presented in the work by Barrio [1]. The numerical domain was composed of several modules: suction duct, impeller, volute, diffuser and outlet duct. The authors made use of and solved the Navier–Stokes equations for three-dimensional unsteady flow. A numerical simulation of the three-dimensional fluid flow inside a centrifugal pump is performed in the work of Stickland [2]. The simulation was made with an unsteady calculation, using the sliding mesh technique in order to take into account the impeller-volute interaction. The effects of the blade outlet angle and passage width on the performance of a centrifugal pump have been investigated in the work of Shojaeefard et al.[3]. The numerical
and experimental investigation of the 3-D flow in the centrifugal pump along with the volute was numerically simulated. The finite volume method was used for the discretization of the governing equations. The proposal of a physical model to study the heat transfer and oil flow of oil pipeline under normal operation is presented in the work of Yu et al. [4]. Numerical simulations in a wide range of operating conditions were conducted.

The investigated system is relatively complex, when investigating the hydrodynamic processes of the entire system. Therefore, in the first stage of investigation, assessment of the main energy sources (centrifugal pumps) and the influence of the physical properties of the transported fluid on hydrodynamic processes is needed.

2 THE PHYSICAL PROPERTIES OF THE GEOTHERMAL WATER

During experimental studies was determined the dependence between the pressure and the amount of gasses in liquid. The results of experimental investigation is presented in Fig. 1.

![Figure 1](image1.png)

Figure 1. The results of amount of gasses depend on the pressure

The regression relation of the amount of gas and the pressure is deduced. The dependence is described using five degree polynomial function:

\[ V(p) = 0.17\times10^{-3} - 0.14\times10^{-5} \ p + 0.41\times10^{-6} \ p^2 - 0.66\times10^{-7} \ p^3 + 0.51\times10^{-8} \ p^4 - 0.15\times10^{-9} \ p^5 \]  

(1)

here \( p \) – the pressure, Pa.

The bulk modulus of fluid is deduced from the experimental results.

3 INVESTIGATION OF STABILITY OF THE DEPTH CENTRIFUGAL PUMP

The pressure and flow rate dependence on the angular velocity of pump impeller is shown in Fig. 2. The main characteristic of the centrifugal extraction pump dependence of developed pressure on flow rate \( p = \pi (Q) \).

The computational scheme of the analyzed system is shown in Figure 3. This system consists of an extractive well, in which at depth \( H_0 \) the deep well centrifugal pump and vertical pipe are immersed.
Figure 2. The pressure and flow rate dependence on the impeller angular velocity

Figure 3. System “deep well centrifugal pump-pipe” stability determination scheme

Mathematical model of the geothermal system “depth centrifugal pump – pipe (pipe system)” was made. Flow rate $Q_{in}$ change in the suction pipe can be described by the equation:

$$
\dot{Q}_{in} = \frac{1}{m_{in}} \left( p_{in} - p_{inS} - p_{init} - \frac{1}{2} \rho_{in} \xi_{in} \left( \frac{Q_{in}}{A_{in}} \right)^2 \text{sign} \left( Q_{in} \right) \right),
$$

(2)

here $m_{in} = \rho_{in} \frac{L_{in}}{A_{in}}$; $p_{in}, p_{inS}, p_{init}$ – the fluid pressures in the suction pipe, pump suction cavity and hydrostatic pressure.

$$
p_{init} = g \rho_{in} (p_{init}, T)L_{in},
$$

(3)

$\rho_{in}$ - pressure; $L_{in}, A_{in}$ – the length and cross-sectional area of suction pipe; $\xi_{in}$ – pressure loss coefficient.

The density of the fluid $\rho_{in} (p_{init}, T)$ depends on the pressure and the temperature, therefore the hydrostatic pressure $p_{init}$ can be determined from the (3) equation:

$$
p_{init} - g \rho_{in} (p_{init}, T)L_{in} = 0
$$

(4)

The pressure change in the pump suction cavity can be described by the equation:
\[ \dot{p}_{inS} = \frac{1}{C_{inS}} \left( Q_{in} - Q_{s}(p_{inS}) \right), \]  

(5)

here \( C_{inS} = \frac{V_{inS}}{K_{inS}(p_{inS})} \), \( V_{inS} \) – the volume of the pump suction cavity; \( K_{inS}(p_{inS}) \) – the bulk modulus of fluid.

The flow generated by the centrifugal pump is equal to:

\[ \dot{Q}_s = \frac{1}{m_s} \left( p_s(Q_s) - p_s - p_{sH} - \frac{1}{2} \rho_s \xi_s \left( \frac{Q_s}{A_s} \right)^2 \right. \left. \text{sign}(Q_s) \right), \]  

(6)

here \( m_s = \rho_s \frac{L_s}{A_s} \) – acoustic mass of the pipeline; \( L_s, A_s \) – high pressure pipe length and cross-sectional area; \( \xi_s \) – pressure loss coefficient.

Pressure change in cavity No. 1 can be described by the equation:

\[ \dot{p}_1 = \frac{1}{C_1} \left( Q_s - Q_{12}(p_2) \right), \]  

(7)

here

\[ C_1 = \frac{V_1}{K_1(p_1)}, \]

\[ Q_{12}(p_2) = A_{12} \mu_{12} \left( \sqrt{2} p_1 - p_2 \right) \text{sign}(p_1 - p_2), \]  

(8)

\( p_2 \) – System load pressure; \( A_{12} \) – cross-sectional area valve; \( \mu_{12} \) – permeability coefficient.

Pressure generated by deep well centrifugal pump is equal to:

\[ p_s = \pi(Q_s) + p_{in}. \]  

(9)

From the equation (2) we find the pressure \( p_{inS} \). We assume that \( p_s = p_{inS} \) and using the equation (6) we get:

\[ \dot{Q}_s = \frac{1}{m_{inS}} \left( F_1(Q_s) - p_1 \right) \]  

here

\[ F_1(Q_s) = p_{in} + \pi(Q_s) - p_{inh} - p_{sH} - \frac{1}{2} \rho_{in} \xi_{in} \left( \frac{Q_s}{A_{in}} \right)^2 \cdot \text{sign}(Q_s) - \frac{1}{2} \rho_s \xi_s \left( \frac{Q_s}{A_s} \right)^2 \text{sign}(Q_s) \]  

(10)

It was determined that the hydrodynamic processes of the analyzed system “deep well centrifugal pump-pipe” can be described with the two-equation ((9) and (6)) system.
\[
\begin{aligned}
\dot{Q}_s &= \frac{1}{m_{\text{ins}}} (F_1(Q_s) - p_1) \\
\dot{p}_1 &= \frac{1}{C_1} (Q_s - Q_{12}(p_2))
\end{aligned}
\]  

(11)

The parameters of static process that characterize the hydrodynamic process of the system “The deep well centrifugal pump-pipe”, can be determined from the equation system (11):

\[
\begin{aligned}
F_1(Q_s) - p_1 &= 0 \\
Q_s - Q_{12}(p_2) &= 0
\end{aligned}
\]  

(12)

The non-linear equation system was solved using Newton’s method. The solution of the equation system (11) can be written:

\[
\begin{aligned}
Q_s &= Q_{s0} + Q \\
p_1 &= P_{10} + p
\end{aligned}
\]  

(13)

After linearization of the system of equations (11) in respect of point \((Q_{s0}, P_{10})\), the following equation is formed:

\[
\begin{aligned}
\dot{Q} &= \frac{1}{m_{\text{ins}}} \left( \frac{dF_1(Q_{s0})}{dQ} Q - p \right) \\
\dot{p}_1 &= \frac{1}{C_1} \left( Q - \frac{dQ_{12}(P_{10})}{dp} \right)
\end{aligned}
\]  

(14)

By eliminating the pressure \(p\) from the equation (14), the second-degree differential equation was formed:

\[
m_{\text{ins}} \ddot{Q} - \left( \frac{dF_1(Q_{s0})}{dQ} - \frac{1}{C_1} \frac{dQ_{12}(P_{10})}{dp} m_{\text{ins}} \right) \dot{Q} + \frac{1}{C_1} \left( 1 - \frac{dQ_{12}(P_{10})}{dp} \frac{dF_1(Q_{s0})}{dQ} \right) Q = 0
\]  

(15)

The following parameters were introduced into the equation (15):

\[K_1 = C_1 \frac{dF_1(Q_{s0})}{dQ} - \frac{dQ_{12}(P_{10})}{dp} m_{\text{ins}}; \quad K_2 = 1 - \frac{dQ_{12}(P_{10})}{dp} \frac{dF_1(Q_{s0})}{dQ}; \quad K_3 = \omega^2 = \frac{K_2}{m_{\text{ins}} C_1}.\]  

(16)

The parameter \(K_3\) – natural angular frequency of the system. Instability of hydrodynamic processes occurring in the system “deep well centrifugal pump-pipe” is determined from the conditions:

\[K_1 > 0 – \text{the dynamic instability}, \quad K_2 \leq 0 \quad \text{and} \quad K_3 \leq 0 – \text{the static instability.}\]

When \(K_1 > 0\) the vibrations occur in the system “deep well centrifugal pump-pipe”. The system becomes dynamically unstable, because of the increased energy level and the increased vibrations amplitude.
When $K_2 > 0$ system is statically stable. When $K_1 \leq 0$ the energy level of the system decreases and the vibrations become damped (dynamically stable system). When $K_1 > 0$ the energy level of the system increases, vibration amplitude increases (dynamically unstable system). System “deep well centrifugal pump-pipe” stability dependencies ($K_1$, $K_2$ and $K_3$), from fluid flow rate, impeller rpm, when inlet pressure $p_{in} = 2.0$ MPa and $p_2 = 1.6$ MPa are shown in Figure 4.

With $p = \pi(Q_s)$ (Fig. 2) as a main characteristic of the deep well centrifugal pump, with assessment of geometrical parameters of the extraction well and physical properties of transported fluid, the following system instability areas were identified as seen in Fig. 4 in the system “deep well centrifugal pump-pipe” both statically (parameter $K_1 \geq 0$) and dynamic (parameters $K_2 < 0$ and $K_3 < 0$) instability are possible. The system is unstable when RPM of the impeller of the pump changes in the range of 35-40 Hz.

A numerical solution of the mathematical models of geothermal system was made. Conditions: depth 300 m., nominal speed of the motor 43 Hz, resistance of pressure at the end of pipeline 1.4 MPa. The results of numerical analysis is presented in Fig 5-8.
5. CONCLUSIONS

- Dependencies amount of gas on the pressure determined with respect to the results of the experiments.
- It was determined that the amount of released gas directly correlates with increased liquid-gas mixture compressibility and slow down of hydrodynamic processes which changes working properties of centrifugal pump – reduces the natural frequency of the system “subsurface centrifugal pump – pipe line/system”.
- Research showed that the natural frequency of the “depth centrifugal pump – pipe” system depends on frequency of the rotation of impeller, discharge and other parameters. It was determined that a low – 1 Hz natural frequency of the system is a result of the combination of high compressibility of the underground water and the large amount of water inside the pipes/tubes.

This work has been supported by the Research Council of Lithuania within the project “Simulation software and investigation of thermo-hydrodynamic processes in the geothermal loop”, project No. MIP-090/2012

REFERENCES


MODAL CHARACTERISTICS OF A FLEXIBLE TUBE IN TURBULENT AXIAL FLOW: A NUMERICAL APPROACH AND VALIDATION WITH EXPERIMENTAL DATA

J. DE RIDDER*, J. DEGROOTE*, K. VAN TICHELEN†, P. SCHUURMANS† AND J. VIERENDEELS*

*Department of Flow, Heat and Combustion Mechanics
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
e-mail: j.deridder@ugent.be, joris.degroote@ugent.be, jan.vierendeels@ugent.be - web page: http://www.ugent.be/ea/floheacom/en/

†Belgian Nuclear Research Center
Boeretang 200, 2400 Mol, Belgium
e-mail: kvtichel@sckcen.be, pschuurm@sckcen.be - web page: http://www.sckcen.be/

Key words: flow-induced vibration, modal characteristics, turbulent axial flow, numerical method

Abstract. Flow-induced vibration is an important concern in the design of tube bundles. Due to the coupling of fluid motion and structural motion, instabilities such as flutter and divergence can arise. Next to the instabilities caused by the coupling of fluid motion and structural motion, turbulence could cause small amplitude vibrations, which in turn could give rise to long-term damage. Currently, the dynamical behavior of a tube in axial flow is studied by splitting the flow forces into inviscid and viscous components. The inviscid flow forces are determined from potential flow theory while the viscous flow forces come from empirical formulations. In this paper, a computational methodology is proposed to improve the accuracy of the predicted dynamical behaviour. In this methodology partitioned fluid-structure interaction simulations are performed to calculate the free vibration decay of a tube in axial flow. The tube is initially deformed according to an eigenmode in vacuum. Modal characteristics are then derived from the free vibration decay of the tube surrounded by the turbulent water flow. To validate this computational methodology a series of experiments is reproduced. In these experiments the frequency and damping of the fundamental mode of a solid brass cylinder were measured.
1 INTRODUCTION

Flow-induced vibration is an important concern in the design of tube bundles, both in axial-flow as well as cross-flow configurations. Applications with an axial-flow regime are typically found in nuclear reactors while cross-flow regimes are typically found in shell-and-tube heat exchangers. This article will focus on axial flow.

Different types of flow-induced instabilities can arise in tube bundles subjected to axial flow, depending on their fixations. Generally however, due to the coupled fluid-structure motion, centrifugal fluid forces can trigger a static instability while Coriolis forces can trigger a dynamic one [1]. Next to the instabilities caused by the coupled motion, small vibrations are triggered by turbulent fluctuations in the flow. While these vibrations are not as catastrophic as the ones induced by the coupled fluid-structure motion, they can damage the structure in the long term.

In order to predict the vibrational behavior of a tube exposed to axial flow, classical models split the flow forces induced by the motion into an inviscid part and a viscous part [2]. The inviscid forces can be derived from potential flow theory. They result in a force proportional to the acceleration (an added mass), a centrifugal force and a Coriolis force. The viscous (turbulent) forces are based on a linearization of empirically determined turbulent friction forces. Research on the constants required in these expressions is still on-going [3]. In cross-flow configurations some authors use (2D) CFD-simulations to establish the required coefficients [4].

Research based on CFD is mainly concerned with cross-flow regimes, in which vortex shedding is one of the important mechanisms of flow-induced vibration. Regarding axial flow a solver was developed for instabilities in laminar flow conditions and later a linearized solver for turbulent annular flow configurations was developed. The initial solver was based on a staggered approach. To keep the computation stable with higher added fluid masses, an estimation of the added mass was afterwards included in the structural solver [5].

In this paper, modal characteristics of a tube in turbulent water flow will be computed directly from coupled computational fluid mechanics (CFD) and computational solid mechanics (CSM). To assess the accuracy of the proposed methodology, the results are compared to experimental results available in open literature.

2 METHODOLOGY

Essentially, modal characteristics in this paper are determined from unsteady computations of the free decay of initially deformed structures [6]. The computations are split into four steps:

STEP 1: Computation of eigenmodes in vacuum

Initially the eigenmodes in vacuum are computed with a finite element solver, searching eigenmodes of:

\[(K - M\omega^2)x_i = 0\] (1)
with $K$ the stiffness matrix, $M$ the mass matrix and $x_i$ the displacements.

**STEP 2: Initialization of the fluid-structure interaction (FSI) simulation**

The cylinder is then deformed according to the previously determined fundamental mode shape. The steady-state flow around the deformed cylinder is computed, thus solving the steady-state mass balance and Navier-Stokes equations for an incompressible fluid:

$$\nabla \cdot \mathbf{v}_f = 0$$  \hspace{1cm}  (2)

$$\rho_f (v_f \cdot \nabla v_f) = -\nabla p + \mu_f \nabla \cdot \nabla v_f + f_f$$  \hspace{1cm}  (3)

in which $p$ stands for pressure, $v_f$ for fluid velocity, $\rho_f$ for fluid density, $\mu_f$ for fluid viscosity.

**STEP 3: Unsteady FSI-calculation**

The deformed structure and the flow field serve as initial state in an unsteady FSI-simulation. In this simulation both the kinematic and dynamic equations need to be satisfied:

$$d_s = d_f$$  \hspace{1cm}  (4)

$$-\tau_s \cdot n_s = \tau_f \cdot n_f$$  \hspace{1cm}  (5)

with $d_s, d_f$ the displacement of the interface on the structural side and on the fluid side respectively, $\tau_f, \tau_s$ the stress on the interface due to the fluid and due to the structure and $n_s, n_f$ the surface normals on the fluid-structure interface of the structural and the fluid domain. The Newtonian fluid flow itself is governed by the incompressible form of the conservation of mass and the Navier-Stokes equations:

$$\nabla \cdot \mathbf{v}_f = 0$$  \hspace{1cm}  (6)

$$\rho_f \left( \frac{\partial \mathbf{v}_f}{\partial t} + \mathbf{v}_f \cdot \nabla \mathbf{v}_f \right) = -\nabla p + \mu_f \nabla \cdot \nabla \mathbf{v}_f$$  \hspace{1cm}  (7)

The structural displacement is governed by Newton’s second law:

$$\rho_s \frac{\partial^2 d_s}{\partial t^2} = \nabla \cdot \tau_s$$  \hspace{1cm}  (8)

with $\rho_s$ the solid density, $\tau_s$ the stress tensor, which is determined using the constitutive equation of the material.

**STEP 4: Extraction of the modal characteristics**

From the previous step the free vibration decay in a fluid of the original in vacuum mode is available. This vibration can be developed into series of $N$ decaying modes:

$$d_{cl,i}(z,t) \approx d_{cl,i,est}(z,t) = \sum_{i=1}^{N} a_i(z) \exp(-2\pi f_i \zeta_i t) \sin(2\pi \sqrt{1-\zeta_i^2} f_i t + \theta_i)$$  \hspace{1cm}  (9)

with $a_i(z)$ the mode shape, $\zeta_i$ the modal damping ratio, $f_i$ the frequency and $\theta_i$ the phase angle of mode $i$. 
3 SIMULATION DETAILS

The geometry computed in this paper is based on the geometry described in [7]. It consists of a solid brass cylinder mounted in a water-conveying pipe. The geometrical parameters as well as material properties are listed in Table 1. The cylinder was pre-tensioned with 648 N. In the experiments the water speed ranged from 10 m/s to 30 m/s.

Table 1: Geometrical parameters and material properties

<table>
<thead>
<tr>
<th>Geometrical parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cylinder diameter</td>
<td>0.0127 m</td>
</tr>
<tr>
<td>cylinder length</td>
<td>1.19 m</td>
</tr>
<tr>
<td>hydraulic diameter</td>
<td>0.0127 m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Material properties</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>density of brass</td>
<td>8400 kg/m³</td>
</tr>
<tr>
<td>Young’s modulus of brass</td>
<td>107 GPa</td>
</tr>
<tr>
<td>Poisson’s ratio of brass</td>
<td>0.34</td>
</tr>
<tr>
<td>water density</td>
<td>1000 kg/m³</td>
</tr>
<tr>
<td>water viscosity</td>
<td>0.001 Pa.s</td>
</tr>
</tbody>
</table>

The flow is thus turbulent with Reynolds numbers between 124000 and 372000. It is computed by solving URANS-equations, with the $k-\omega$ SST model. The influence of different inlet turbulent intensities is studied. As the cylinder is moving, the flow equations are cast in the ALE (arbitrary Lagrangian-Eulerian) form. The mesh motion is computed by iteratively solving a system of springs. The flow solver uses a 2nd-order discretization, both in space and in time.

The CSM-calculation is a finite element calculation, which uses 2nd-order elements and the Hilber-Hughes-Taylor time-integrator, which is also of 2nd-order accuracy. The fluid-structure coupling is computed with the IQN-ILS algorithm [8].

4 CONVERGENCE STUDY

4.1 Number of modes required in the free-vibration fitting

In the case, studied in this article, the mode shapes of the coupled fluid-structure are very similar to ones of the pure structural problem. As a result, only one mode is required to predict the decay of the fundamental mode. Fitting with one mode already gave $R^2$-values of more than 99.9%.

The goodness of fit can also be seen on Figure 1, which plots the reconstructed displacement as a function of time, together with the error to the original displacement. The error graph shows peaks probably belonging to a small 2nd-mode contribution.
4.2 Influence of time step size

If the time step in the simulations is chosen too large, the modal damping ratio is typically overestimated [6]. However, in the case studied here, neither the modal damping ratio nor the natural frequency were very sensitive to the time step size, as can be seen in Figure 2. In the remainder of the article a time step size of 0.2 ms is used, as it reduces the computational cost.

4.3 Grid convergence

The grid used for the finite-element calculation consists of 400 quadratic 3D-elements. Grid refinements showed no significant change in eigenmodes or eigenvalues. The mesh used in the CFD-calculation consists of 235000 cells. The first grid point is located in the logarithmic region, as its $y^+$-value is 200 for a water velocity of 30 m/s. Modal characteristics were also determined with a mesh consisting of only 62500 cells. Neither the frequency nor the damping showed appreciable difference with the mesh normally used.

5 RESULTS

In this section the influence of inlet conditions, water velocity and molecular viscosity is discussed.
5.1 Influence of inlet conditions

In order to solve the (U)RANS-equations, two additional turbulent quantities are required at the inlet. They will alter the turbulent (eddy) viscosity downstream, which influences the modal damping.

As the ratio of the channel length over its hydraulic diameter (Dh) is relatively high, the flow becomes already fully developed at the beginning of the channel. From Figure 5.1, the flow is fully developed after approximately 20 Dh, while the entire channel length is approximately 95 Dh. Changes in inlet turbulence characteristics only modifies the turbulent viscosity in the beginning of the channel. This is in agreement with experimental work performed by Mulcahy [9]. These changes in turbulent viscosity do not result in significant changes of modal characteristics (on the order of 1 %), as they are only occurring in a limited part of the domain and the modal shape is very small in that part.

5.2 Influence of water velocity

If the flow velocity is high enough, divergence or flutter of the cylinder could occur. Conventional theories predicting flow-induced vibration often recast the flow velocity in a dimensionless form [2]:

\[ v_{f,n} = \left( \frac{EI}{\rho_f A} \right)^{-0.5} v_f L \]  

with \( I \) the second moment of inertia and \( E \) the Young’s modulus of elasticity. For a flow velocity of 30 m/s this dimensionless flow velocity is still only 1.1. Flow instabilities typically occur for non-dimensional flow velocities greater than 2-6 [2].

While the flow velocity is not high enough to trigger flow-induced instabilities, it does change the modal characteristics, as can be seen in Table 2. An increase of flow velocity leads to lower natural frequencies and higher modal damping ratios. The decrease in
Figure 3: Influence of the inlet turbulence length scale (left graph) and turbulence intensity (right graph) on the turbulent viscosity at $Dh/8$ from the cylinder.

Table 2: Natural frequency and modal damping ratio of the ground mode vibration as a function of flow velocity

<table>
<thead>
<tr>
<th>$v_f$ (m/s)</th>
<th>$f_n$ (s$^{-1}$)</th>
<th>$\zeta_n$ ($-$)</th>
<th>$f_n$ (s$^{-1}$)</th>
<th>$\zeta_n$ ($-$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>28.9</td>
<td>0.014</td>
<td>27.9</td>
<td>0.013</td>
</tr>
<tr>
<td>20</td>
<td>28.7</td>
<td>0.022</td>
<td>27.7</td>
<td>0.021</td>
</tr>
<tr>
<td>30</td>
<td>28.4</td>
<td>0.030</td>
<td>27.5</td>
<td>0.030</td>
</tr>
</tbody>
</table>

natural frequency can be explained by an increase of centrifugal forces acting on the cylinder, while the increase in modal damping can be attributed to the increase of turbulent viscosity. Table 2 further shows a good agreement between the computed characteristics and the experimentally determined characteristics.

While current theories have similar reliability on frequency prediction compared to the calculations in this article, they have to include an empirical friction correlation to predict the modal damping ratio.

5.3 Influence of molecular viscosity

The molecular viscosity is not exactly known as the water temperature is not well known. Therefore, different computations with a molecular viscosity between 0.0005 Pa.s and 0.002 Pa.s have been carried out.

For a flow velocity of 30 m/s, the turbulent viscosity is 431 times the molecular one. The molecular viscosity will thus only affect the modal characteristics because it alters its turbulent counterpart.

Table 3 lists the modal characteristics and the average turbulent viscosity for different molecular viscosities. The molecular viscosity has almost no influence on the natural
frequency. However, it influences the modal damping ratio, through changes in turbulent viscosity.

6 CONCLUSION

Modal characteristics are computed purely numerically in this article. It has the advantage over conventional methods that it does not require specific empirical input. The results obtained with the method, presented in this article, have been compared and validated with experiments available in open literature.

The calculations showed that the influence of inlet conditions decayed after 20 Dh. This observation is in accordance with experimental findings. An increase in flow speed gave slightly lower frequencies and significant higher modal damping ratios. Changes in molecular viscosity resulted in slightly different turbulent viscosities and thus slightly different modal damping ratios.

REFERENCES


MULTIDISCIPLINARY AND MULTIPHYSICS COMPUTATION OF SUPersonic FLOW, VIA HYBRID SOLUTIONS FOR COMPRESSIBLE NAVIER-STOKES LAYER

A. NASTASE*

* Aerodynamics of Flight
RWTH, Aachen University
Templergraben 55, 52062 Aachen, Germany
e-mail: nastase@lafaero.rwth-aachen.de

Key words: Meshless, Three-Dimensional, Hybrid Solutions for Compressible Navier-Stokes PDEs, Interfaceless Viscous-Inviscid Interactions, Multidisciplinary Weak Interactions Aerodynamics-Structure, Supersonic Flow.

Abstract. The author proposes hybrid meshless numerical solutions for the computation of the three-dimensional compressible full partial-differential equations (PDEs) of the Navier-Stokes layer (NSL) over flying configurations (FCs) and for the performing of the aerodynamical global optimal design of their shapes. These hybrid solutions have important analytical properties and the derivatives of the velocity's components can be exactly computed. By using a logarithmic density and by applying of the collocation method, a splitting of the NSL's PDEs occurs and all the physical entities are expressed only as functions of the spectral coefficients of the velocity's components, which are obtained by iterative solving of the impulse PDEs. The viscous-inviscid interaction does not need interface and a weak interaction aerodynamics-structure is proposed.

1 INTRODUCTION

The author proposes new hybrid numerical solutions for the three-dimensional, compressible, Navier-Stokes layer (NSL) over a flying configuration (FC), in supersonic flow. The velocity's components of these hybrid NSL's solutions are expressed as products between the corresponding analytical determined potential velocity's components of the same FC and polynomial expansions, with arbitrary coefficients, as in $1^{-3}$. The potential solutions are used also as outer flow, at the NSL's edge.

The absolute temperature and the logarithmic density function are also expressed in form of polynomial expansions with arbitrary coefficients, which are used to satisfy the temperature and, respectively, the continuity PDEs, in some chosen points. By using the logarithmic density instead of the density it was possible to split the NSL's PDEs and to express all the physical entities only as functions of spectral coefficients of the velocity's components. By using the collocation method, these coefficients of velocity's components are determined by iteratively solving of impulse PDEs, which are considered now as a linear algebraic system.
with variable coefficients and which are iteratively solved.

The determined pressure distribution on FC and the load distribution of its structure produce together the deformation of the structure. By cruising flight the pressure of the lower side produces a compression of the structure and, the depression on its upper side, dilates the structure of FC. The deformation of the structure produces the change of the shape of FC and, consequently, the change of the pressure distribution on FC occurs. This iterative process must have a damped character.

2 HYBRID SOLUTIONS FOR THE THREE-DIMENSIONAL COMPRESSIBLE NAVIER-STOKES PDES

The author proposes hybrid numerical solutions for the three-dimensional compressible NSL's PDEs, as in 31−3. Let us firstly introduce the following spectral variable η

$$\eta = \frac{x_1 - Z(x_1, x_2)}{\delta(x_1, x_2)} \quad (0 < \eta < 1)$$

Hereby \(Z(x_1, x_2)\) is the equation of the surface of the flattened FC and \(\delta(x_1, x_2)\) is the NSL’s thickness distribution. The following spectral forms of the axial, lateral and vertical velocity’s components \(u_\delta, v_\delta\) and \(w_\delta\), of the density function \(R = \ln \rho\) (here introduced) and of the absolute temperature \(T\) are here proposed, as in 31−3,

$$u_\delta = u_\epsilon \sum_{i=1}^N u_i \eta^i, \quad v_\delta = v_\epsilon \sum_{i=1}^N v_i \eta^i, \quad w_\delta = w_\epsilon \sum_{i=1}^N w_i \eta^i,$$

$$R = R_\epsilon + (R_\epsilon - R_0) \sum_{i=1}^N r_i \eta^i, \quad T = T_\epsilon + (T_\epsilon - T_0) \sum_{i=1}^N t_i \eta^i. \quad (2a-e)$$

Further, the physical equation of ideal gas for the pressure \(p\) and an exponential law for the viscosity \(\mu\) versus the temperature \(T\) are used:

$$p = R_\epsilon \rho T = R_\epsilon e^{R_\epsilon T}, \quad \mu = \mu \left[ \frac{T}{T_\epsilon} \right]^n. \quad (3a,b)$$

Here \(R_\epsilon\) and \(T_\epsilon\) are the given values of \(R\) and \(T\) at the wall, \(R_\epsilon\) and \(T_\epsilon\) the universal gas constant and the absolute temperature of the undisturbed flow and \(n_1\) the viscosity exponent, \(u_\epsilon, v_\epsilon, w_\epsilon, R_\epsilon\) and \(T_\epsilon\) are the values of \(u, v, w, R\) and \(T\) at the NSL’s edge, obtained from the outer inviscid hyperbolical potential flow and \(u_i, v_i, w_i, r_i\) and \(t_i\) are their free spectral coefficients, which are used to fulfill the NSL’s PDEs.

The non-slip condition on the FC’s surface (\(\eta=0\)) is automatically satisfied by the equations (2a-c) and the boundary conditions for the velocity's components at the NSL’s
edge, written in explicit forms, are the following:

\[ u_{N-2} = \alpha_{0,N-2} + \sum_{i=1}^{N-3} \alpha_{i,N-2} u_i , \quad v_{N-2} = \alpha_{0,N-2} + \sum_{i=1}^{N-3} \alpha_{i,N-2} v_i , \]

\[ u_{N-1} = \alpha_{0,N-1} + \sum_{i=1}^{N-3} \alpha_{i,N-1} u_i , \quad v_{N-1} = \alpha_{0,N-1} + \sum_{i=1}^{N-3} \alpha_{i,N-1} v_i , \]

\[ u_{N} = \alpha_{0,N} + \sum_{i=1}^{N-3} \alpha_{i,N} u_i , \quad v_{N} = \alpha_{0,N} + \sum_{i=1}^{N-3} \alpha_{i,N} v_i , \]

\[ w_N = \gamma_{0,N} + \sum_{i=1}^{N-1} \gamma_{i,N} w_i . \]  \hspace{1cm} (4a-g)

Hereby the coefficients \( \alpha_{i,j} \) and \( \gamma_{i,j} \) are of the forms:

\[ \alpha_{0,N-2} = \frac{N^2 - N}{2} , \quad \alpha_{i,N-2} = -\frac{1}{2} \left[ N^2 - N \left( 1 + 2i \right) + i^2 + i \right] , \]

\[ \alpha_{0,N-1} = -N^2 + 2N , \quad \alpha_{i,N-1} = N^2 - 2N \left( i + 1 \right) + i^2 + 2i , \]

\[ \alpha_{0,N} = \frac{N^2 - 3N + 2}{2} , \quad \alpha_{i,N} = -\frac{1}{2} \left[ N^2 - N \left( 3 + 2i \right) + i^2 + 3i + 2 \right] , \]

\[ \gamma_{0,N} = 1 , \quad \gamma_{i,N} = -1 . \]  \hspace{1cm} (5a-h)

Additionally, the boundary conditions for the absolute temperature and for the density functions, at the NSL's edge, written in implicit forms, are:

\[ \sum_{i=1}^{N} r_i = 1 , \quad \sum_{i=1}^{N} t_i = 1 . \]  \hspace{1cm} (6a,b)

The equations (4a-g) are used for the elimination of the seven corresponding spectral coefficients of the velocity's components from the NSL's PDEs and for the update their values in the different steps of iteration.

The PDE of continuity and the boundary condition (6a) are used for the computation of spectral coefficients \( r_i \) of the here introduced density function \( R = \ln r \)

For this purpose the collocation method is used. These coefficients can be expressed only as functions of the spectral coefficients of the velocity's components by solving of a linear algebraic system, namely:
\[ \sum_{i=1}^{N} g_{ip} r_i = \gamma_p \quad (p = 1, \ldots, N) \quad , \quad \tag{7a} \]

\[ r_p = \frac{\Delta P}{\Delta} \quad , \quad (p = 1, \ldots, N) \quad \tag{7b} \]

Further, the following notations are introduced:

\[ L = u_\delta \frac{\partial R_w^e}{\partial x_1} + v_\delta \frac{\partial R_w^e}{\partial x_2} + w_\delta \frac{\partial R_w^e}{\partial x_3} \quad , \quad Q = - \left[ \frac{\partial u_\delta}{\partial x_1} + \frac{\partial v_\delta}{\partial x_2} + \frac{\partial w_\delta}{\partial x_3} \right] \quad , \quad \tag{7c} \]

\[ M = u_\delta \left[ \frac{\partial R_w^e}{\partial x_1} - \frac{\partial R_w^w}{\partial x_1} \right] + v_\delta \left[ \frac{\partial R_w^e}{\partial x_2} - \frac{\partial R_w^w}{\partial x_2} \right] + w_\delta \left[ \frac{\partial R_w^e}{\partial x_3} - \frac{\partial R_w^w}{\partial x_3} \right] \quad , \quad \tag{7d} \]

\[ S = (R_e - R_w) \left[ u_\delta \frac{\partial \eta}{\partial x_1} + v_\delta \frac{\partial \eta}{\partial x_2} + w_\delta \frac{\partial \eta}{\partial x_3} \right] \quad . \quad \tag{7e} \]

It results in:

\[ g_{ip} = \left( \eta^{-1} (M \eta + i S) \right)_p \quad , \quad \gamma_p = (Q - L)_p \quad \text{if} \quad 1 \leq p \leq N - 1 \quad , \quad \tag{7f} \]

\[ g_{ip} = 1 \quad , \quad \gamma_p = 1 \quad \text{if} \quad p = N \quad . \quad \tag{7g} \]

In this moment the NSL's PDEs are split.

The PDE of the absolute temperature and the boundary condition (6b) are used for the computation of spectral coefficients \( t_i \) of the absolute temperature \( T \). For this purpose the collocation method is also used. These coefficients can be expressed only as functions of the spectral coefficients of the velocity's components by numerical solving of a transcendental algebraic system, it is:

\[ \sum_{i=1}^{N} h_i t_i + h_0 T^n = \theta \quad , \quad \sum_{i=1}^{N} h_{ip} t_i + h_{0p} \left( T^n \right)_p = \theta_p \quad . \quad (p = 1, \ldots, N) \quad \tag{8a,b} \]

\[ h_{ip} = \left( h_i \right)_p \quad , \quad h_{0p} = \left( h_0 \right)_p \quad \theta_p = (\theta)_p \quad \text{if} \quad 1 \leq p \leq N - 1 \quad . \quad \tag{8c,d} \]

\[ h_{ip} = 1 \quad , \quad h_{0p} = 0 \quad \theta_p = 1 \quad \text{if} \quad p = N \quad . \quad \tag{8e,f} \]
The impulse PDEs are used for the determination of the spectral coefficients of the velocity's components. If the spectral forms (2a-c) of velocity's components, (2d,e) of the density function and of absolute temperature are taken into consideration, the seven spectral coefficients, obtained by writing the boundary conditions in explicit form (4a-g), are eliminated and the collocation method is used, the spectral coefficients of the velocity's components are obtained by the iterative solving of a linear algebraic system with variable coefficients:

\[
\sum_{i=1}^{N-3} \left( \widetilde{A}_{ik}^{(1)} u_i + \widetilde{B}_{ik}^{(1)} v_j \right) + \sum_{i=1}^{N-3} \widetilde{C}_{ik}^{(1)} w_i = \sum_{i=1}^{N-3} u_i \left[ \sum_{j=1}^{N-3} \left( \widetilde{A}_{ij}^{(1)} u_j + \widetilde{B}_{ij}^{(1)} v_j \right) + \sum_{j=1}^{N-3} \widetilde{C}_{ij}^{(1)} w_j \right] - \tilde{D}_k^{(1)},
\]

\[
\sum_{i=1}^{N-3} \left( \widetilde{A}_{ik}^{(2)} u_i + \widetilde{B}_{ik}^{(2)} v_j \right) + \sum_{i=1}^{N-3} \widetilde{C}_{ik}^{(2)} w_i = \sum_{i=1}^{N-3} v_i \left[ \sum_{j=1}^{N-3} \left( \widetilde{A}_{ij}^{(2)} u_j + \widetilde{B}_{ij}^{(2)} v_j \right) + \sum_{j=1}^{N-3} \widetilde{C}_{ij}^{(2)} w_j \right] - \tilde{D}_k^{(2)},
\]

\[
\sum_{i=1}^{N-3} \left( \widetilde{A}_{ik}^{(3)} u_i + \widetilde{B}_{ik}^{(3)} v_j \right) + \sum_{i=1}^{N-3} \widetilde{C}_{ik}^{(3)} w_i = \sum_{i=1}^{N-3} w_i \left[ \sum_{j=1}^{N-3} \left( \widetilde{A}_{ij}^{(3)} u_j + \widetilde{B}_{ij}^{(3)} v_j \right) + \sum_{j=1}^{N-3} \widetilde{C}_{ij}^{(3)} w_j \right] - \tilde{D}_k^{(3)}.
\]

(9)

3 COMPUTATION OF AERODYNAMIC CHARACTERISTICS OF A WEDGED DELTA WING MODEL AND THE COMPARISON WITH EXPERIMENTAL RESULTS

Let us consider the wedged delta wing model presented in the (Fig. 1).

**Figure 1:** The Wedged Delta Wing Model

This model has the following geometrical characteristics:
- \( b = 16.703 \ cm \) the maximal span
- \( h_1 = 17.362 \ cm \) the maximal depth
- \( S_0 = 145 \ cm^2 \) the area of its planform
- \( \delta = 25.7^\circ \) the angle of aperture in the planform
- \( \gamma = 5.62^\circ \) the angle of aperture in the vertical symmetry plane \( Ox_1x_3 \)
- \( \ell = 0.481 \) the dimensionless span
- \( \tau = 0.0946 \) the relative volume

The lift and pitching moment coefficients, computed by using an own developed inviscid solver, are in very good agreement with the experimental-correlated results obtained in the trisonic wind tunnel of the DLR Cologne, in the frame of one of the research projects of the author, sponsored by the DFG. These agreements can be seen in the (Figs. 2a,b), for the all ranges of angles of attack \( (\alpha = -16^\circ \div 16^\circ) \) and Mach numbers \( (M_\infty = 1.25 \div 2.2) \) taken here into consideration. For these ranges of \( \alpha \) and \( M_\infty \) the wedged delta wing model has subsonic leading edges. These agreements between theory and experiment lead to the following remarks:

- if the FC is flattened enough and flies at moderate angles of attack, the more economical flight with three-dimensional characteristic surface (instead of the flight with shock waves) occurs and the own developed software, based on it, are confirmed;
- the supersonic flow is laminar, as supposed here and remains attached for larger range of angles of attack than by subsonic flow;
- the influence of friction upon the lift and pitching moment coefficients is neglectable.

**Figures 2a,b**: The Lift- and Pitching Moment Coefficients of the Wedged Delta Wing Model with Subsonic Leading Edges
A. Nastase.

The developed hybrid NSL's solutions are used for the computation of the total drag, which include the inviscid and the friction drag. The inviscid drag coefficients of the thin, of the thick-symmetrical components and of the delta wings are the following:

\[ C_d = 8 \ell \int_{\hat{\delta} \hat{\lambda} \hat{c}} \bar{u} \bar{w} \bar{x}_1 \, d\bar{x}_1 \, d\bar{y} \ , \quad C_d^* = 8 \ell \int_{\hat{\delta} \hat{\lambda} \hat{c}} \bar{u}^* \bar{w}^* \bar{x}_1 \, d\bar{x}_1 \, d\bar{y} \ , \quad C_d^{(t)} = C_d + C_d^* \quad (10a-c) \]

The friction drag and the total drag coefficients are:

\[ C_d^{(f)} = 8 \nu_j u \int_{\hat{\delta} \hat{\lambda} \hat{c}} u \bar{x}_1 \, d\bar{x}_1 \, d\bar{y} \ , \quad C_d^{(i)} = C_d^{(f)} + C_d^{(i)} \quad (11a,b) \]

In the (Fig. 3) are represented the dependences of the inviscid and of the total drag coefficients of the wedged delta wing on angle of attack for the Mach number \(M_\infty = 2\). It is to remark that the viscous drag coefficient has:
- an important contribution in the total drag coefficient of the wedged delta wing and cannot be neglected;
- the magnitudes of all three drag coefficients increase when the absolute value of the angle of attack increases, but the increment of friction drag coefficient is much slowly as those of the inviscid drag coefficient;
- all the dependences of drag coefficients of wedged delta wing versus the angle of attack are symmetrical with respect to the value of the angle of attack \(\alpha = 0^\circ\).

![WEDGED DELTA WING](image)

**Figure 3:** The Dependences of Inviscid and of Total Drag Coefficients \(C_d^{(i)}\) and \(C_d^{(f)}\) upon the Angle of Attack \(\alpha\) at the Mach Number \(M_\infty = 2\)

Due to analytical hybridization the numerical solutions, presented here, have important analytical properties, namely: they have correct last behaviors, they have correct jumps along
the singular lines (like subsonic leading edges, junction lines wing-fuselage and wing-leading edge flaps) according to the minimum singularities (which must fulfill the jumps) and the singularities are balanced. Additionally, for hyperbolic PDE the condition on the characteristic is automatically fulfilled. The viscous-inviscid coupling is realized without interface need, because the hybrid numerical solutions for the NSL's PDEs are prepared for this coupling, due to the hybridization. These hybrid NSL's solutions are also useful for the performing of the viscous global optimal design (GOD) of the shapes of FCs.

4 THE MULTIDISCIPLINARY WEAK INTERACTION AERODYNAMICS-STRUCTURE

A weak interaction is proposed. The influence of structure requests upon the aerodynamics GOD of the FCs shapes with respect to minimum drag at cruise can be transformed in new or modified constraints for the GOD, via iterative optimum-optimorum theory, up its second step of iteration, as proposed in \(^1,^4\). A limitation of the twist can be necessary for FCs, which are flying at higher supersonic Mach number and it can be realized by imposing the constraint of pressure equalization along the subsonic leading edges at a lower supersonic Mach number than the cruising one. A necessary augmentation of thickness distribution can be necessary in the rear part of FC, which can be obtained by artificial prolongation of the depth and, after the optimization of the thickness distribution, the initial depth is restored by cutting and elimination of the prolongation. The influence of aerodynamics upon the structure is obtained by adding the aerodynamic pressure distribution to the load distribution and by computing the new resulting deformation of the structure.

5 CONCLUSIONS

The proposed reinforced hybrid numerical solutions are split, do not need viscous-inviscid interface at the NSL's edge and have important analytical properties, like correct last behaviors, correct jumps along the singular lines (like subsonic leading edges, junction lines wing-fuselage and wing-leading edge flaps), fulfill automatically the non-slip condition on the FC's surface, are meshless, are split, accurate and rapid convergent. Additionally, the derivatives of the velocity's components can be exact computed and the boundary condition on the Mach cone of the apex is automatically fulfilled. A weak interaction aerodynamics-structure is proposed.

REFERENCES

NUMERICAL SIMULATION OF A DEFORMABLE CELL IN MICROCHANNELS

F. NASON*, L. ZHU†, G. DUBINI* AND L. BRANDT†

* Department of Chemistry, Materials and Chemical Engineering "Giulio Natta"
Politecnico di Milano
P.za Leonardo da Vinci, 32 - 20133 Milano, Italy
e-mail: francesca.nason@mail.polimi.it

†Linne Flow Center and SeRC (Swedish e-Science Research Centre)
KTH Mechanics
Stockholm, Sweden
e-mail: luca@mech.kth.se

Key words: FSI, Coupled Problems, Single-cell, Computing Methods

Abstract. The main goal of this work is to numerically investigate the behavior of a cell flowing in a microfluidic system. In particular, we want to model flow-induced deformations of an isolated cell to quantitatively evaluate the cell response when subjected to a representative range of flow rates in a realistic geometry, with specific interest in the case of cell trapping. This research will help optimize operating conditions as well as the design of cell manipulation/culture micro-devices, so as to guarantee cell viability and ultimately improve high-throughput performance.

1 INTRODUCTION

To analyze, describe and also predict the motion of cells in large vessels, microcapillaries and cell separation devices, theoretical and computational models with various degrees of complexity and physiological relevance have been developed. In particular, in the context of particulate microhydrodynamics, cells are treated as capsules, defined as flexible particles formed by a well-defined, distinct, structured elastic membrane containing a liquid in the interior and possibly a stiff nucleus or core [12]. The hydrostatics and hydrodynamics of liquid capsules enclosed by thin elastic shells have therefore received considerable attention in cellular biology, bioengineering and even in microencapsulation technology [11]. The properties of the membrane material play a crucial role in the dynamics of the capsule, therefore the effect of the membrane mechanical properties on capsules shape at equilibrium, deformability and transient motion has been investigated in various types of simple flows. For a review of analytical, numerical and experimen-
tal works of the dynamics of capsules in shear flow the reader is referred to [10]. The effect of the interfacial elasticity on the capsule deformation and on the rheology of dilute suspensions for small deformations of the initial spherical shape were clarified by the pioneering theoretical investigations of Barthes-Biesel and Barthes-Biesel & Rallison [1, 2]. Numerical studies for moderate and large deformations were presented later, often together with experimental analyses. The effects of surface viscosity and incompressibility, relevant to biological membranes consisting of lipid bilayers, were considered more recently [10]. The analysis of the flow in the presence of a capsule is computationally very challenging. The kinematic and dynamic coupling of the flow enclosed in the capsule and the one surrounding it must account for the mechanical properties of the interface; moreover, an outstanding number of non-spherical transient shapes of the capsule can be obtained due to the membrane deformability and these must be computed simultaneously with the variables characterizing the flow [3].

In modern biology, a wealth of tools and new techniques are under development for cell analysis and in particular to investigate stem cells, as they represent a primary source of cells in the context of regenerative medicine [17]. It is believed that differentiated autologous stem cells would be the perfect inhabitants of an engineered tissue surrogate to be assembled and used surgically for bodily repair [14]. Microfluidics, due to its intrinsic characteristics and advantages, is commonly considered to be a powerful and enabling approach for studying cell behavior [15]. A plethora of evidence has shown that cellular heterogeneity commonly exists within an isogenic or clonal population. The most effective approach is therefore to analyze a population at individual cell level. However, a significant number of individual cells is required to obtain statistically meaningful data, and therefore high-throughput analysis is essential. Since cell populations can be very sensitive to stress, this must be taken into account when handling cell samples within high-throughput microfluidic devices, since high velocities may imply high shear stress acting on the cell membrane, potentially resulting in significant cell death and hence in a loss of representative sample across the initial population. [16]

The aim of this work is therefore to develop new state-of-art numerical tools to efficiently investigate the behavior of a cell flowing in a microfluidic system. After presenting our implementation, we simulate flow-induced deformations of an isolated cell run in a realistic microdevice.

2 MATERIALS AND METHODS

We investigate here the deformation of a cell flowing in a microchannel with a 90-degree bend, a configuration most likely to be encountered by cells loaded and run into a microfluidic device for cell culture, analytical purposes or cellular assays in general. To study the evolution of liquid capsules in microchannels we adopted a hybrid boundary integral/immersed boundary method [13, 8], coupled with a spectral discretization of the membrane surface [18].
2.1 Fluid phase

For the fluid part, we solve the steady Stokes equation with boundary conditions given by the external walls limiting our domain and a forcing related to the internal stresses arising on the membranes

\[ -\nabla p(x) + \eta \nabla^2 u(x) + \rho(x) = 0 \]
\[ \nabla \cdot u(x) = 0. \]

(1)

(2)

In the equation above \( \rho(x) \) represents point forces due to the deformable body in the flow. As mentioned above, these are related to the internal deformation of the suspended phase and will be described in detail in the following section. To solve the Stokes equation in a generally complex geometry we follow the General Geometry Ewald Method (GGEM) proposed by Graham [13, 8]. Exploiting the linearity of Stokes problem, we decompose the forcing as

\[ \rho(x) = \rho_l(x) + \rho_g(x) \]

(3)

\( f_i \) is the force density at the discrete points representing the effect of the membrane on the fluid and \( g(x) = \alpha^3/\pi \eta \left( e^{-\alpha^2 r^2} [2.5 - \alpha^2 r^2] \right) \) is a screen function used to decouple local and global solution.

The local solution \( u_l(x) = \sum_{i=1}^{n} G_l(x - x_i) \cdot f_i \) describes the near field effects of the forcing and does not account for any boundary effects or long range interactions. The Green function of the local problem can be computed analytically in free space

\[ G_l(x) = \frac{1}{8\eta \pi} \left[ \frac{\delta + \frac{xx}{r^2}}{r} \text{erfc}(\alpha r) \right] - \frac{1}{8\eta \pi} \left[ \delta - \frac{xx}{r^2} \right] \frac{2\alpha}{\pi^{1/2}} e^{-\alpha^2 r^2} \]

(4)

with \( r \) the distance between the singularity and the evaluation point \( x \). In the numerical code, the solution at each grid point was obtained by summing only over the set of closest points. Indeed, the parameter \( \alpha \) represents the scale of the screen function \( g \) and, as a consequence, it defines the length over which mutual interaction are taken into account (cut-off radius is chosen to be \( 4/\alpha \)) as well as the grid size required for an accurate solution [8]. Note that \( G_l(x) \) is singular as \( r \to 0 \): two approaches can be followed here. i) Use singular integration as in traditional boundary integral methods [12, 13]; ii) regularized Stokeslet [7]. The second approach was followed for this work. The first approach turns out to be crucial for small confinement, small distances between walls and cell.

The global solution was obtained numerically as solution of the Stokes system (1) forced by \( \rho_g(x) \) only. Boundary conditions at the solid boundaries are enforced such that the total solution \( u = u_l + u_g \) satisfies no slip.

The global Stokes problem can be addressed using any solver. In the present work, we adopted the open source Spectral Element Method (SEM) Nek5000 developed by Paul Fischer at Argonne National Laboratory, USA [5]. Nek5000 is a numerical code...
for the simulation of steady and unsteady incompressible fluid flow and heat transfer, as well as optional convective-diffusive passive scalar quantities. The present method appears particularly suited for the numerical simulation of multiphase flow at vanishing Reynolds number in complex geometries. It combines the accuracy of the boundary integral approach and the flexibility of immersed boundary in terms of geometry. Possible extensions are described in [8]. To the best of our knowledge, we present here the first implementation of this method suited for any geometry.

2.2 Elastic membrane

To represent the surface of the cell we adopt spherical harmonics. Several advantages come when using this approach, such as the spectral accuracy and the solution of singularities at the poles when computing derivatives of the basis functions. In addition, this discretization provides uniform resolution over the surface ensuring minimum of the $L_2$ norm of the approximation error and removal of any time step limitation associated with the close spacing of the collocation points near the poles. Finally, dealiasing can be easily implemented to overcome nonlinear instability without degrading the accuracy of the solution.

The method shortly presented here closely follows the formulation in Zhao et al.[18]. The reader is referred to this paper for a more detailed description. The surface of each cell is represented as series of spherical harmonics function of the two angles $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$

$$f(\theta, \phi) = \sum_{n=0}^{N-1} \sum_{m=0}^{n} \hat{P}_n^m (\cos \theta) (a_{nm} \cos m\phi + b_{nm} \sin m\phi)$$

(5)

with the orthonormal Legendre polynomial

$$\hat{P}_n^m(x) = \frac{1}{2^n n!} \sqrt{\frac{(2n + 1)(n - m)!}{2(n + m)!}} \left(1 - x^2\right)^{\frac{m}{2}} \frac{d^{n+m}}{dx^{n+m}} (x^2 - 1)^n$$

(6)

and $N^2$ the total number of spherical harmonics. Transforms are performed using the SPHEREPACK library.

Once the stress-free reference shape of the cell is defined, the surface deformation is described by the tensor $F = a_\alpha A^\alpha$ mapping the curvilinear base vector in the undeformed state $A^\alpha$, into the current tangent vectors $a_\alpha$, $\alpha = 1, 2$. The local deformation is measured by the Cauchy–Green tensor $V^2 = F \cdot F^T$ whose invariants are

$$I_1 = A^{\alpha\beta} a_\alpha a_\beta - 2, \quad I_2 = |A^{\alpha\beta}| a_\alpha a_\beta| - 1,$$

where $|\cdot|$ defines the determinant of a matrix. To model a thin hyper-elastic membrane, two commonly used constitutive laws, the neo-Hookean law and the Skalak law are used. In the neo-Hookean model, the elastic strain energy $W$ is written as

$$W = \frac{E_s}{2} \left(I_1 - 1 + \frac{1}{I_2 + 1}\right)$$

(7)
whereas, in the Skalak model, $W$ is formulated as
\[ W = \frac{E_s}{4} \left( I_1^2 + 2I_1 - 2I_2 + CI_2^2 \right). \] (8)

$E_s$ is the elastic shear modulus and $C$ is the reduced area dilation modulus. The first quantifies resistance to shear, whereas the second coefficient indicates resistance to area variations.

The contravariant expression for the in-plane stress tensor can be obtained as derivative of the elastic energy
\[ \tau^{\alpha\beta} = \frac{2}{J_s} \frac{\partial W}{\partial I_1} A^{\alpha\beta} + 2J_s \frac{\partial W}{\partial I_2} a^{\alpha\beta}, \] (9)

where $J_s = \sqrt{1 + I_2}$ indicates the ratio between deformed and undeformed surface area.

To be able to simulate conditions when negative tension occurs the bending stiffness of the membrane must be also taken into account. The bending moment was assumed to be linear and isotropic
\[ M_{\alpha\beta} = -E_B (b_{\beta} \alpha - B_{\beta} \alpha) \]
where $E_B$ is the bending modulus and $B_{\beta} \alpha$ the curvature tensor for the reference state. The transverse force on the membrane $Q_{\beta}$ can be obtained by a local torque balance
\[ M_{\alpha\beta} \mid \alpha - Q_{\beta} = 0, \] (10)

where $\gamma_{\alpha}$ is the covariant derivative. Finally, we impose force balance to derive an expression for the force density at the membrane
\[ \tau^{\alpha\beta} \mid \alpha - b_{\alpha} Q_{\alpha} + f^{\beta} = 0, (\beta = 1, 2) \] (11)
\[ Q_{\alpha} + \tau^{\alpha\beta} b_{\alpha\beta} + f^3 = 0. \] (12)

The numerical algorithm is as follows, where the Lagrangian mesh indicates the nodes used to discretize the membrane whereas the Eulerian mesh is used to solve the global Stokes problem.

1. Calculate analytically the local velocity $u_l(x)$ on the Lagrangian mesh using the force density $f$ from the previous time step ($f = 0$ at first time step).
2. Calculate the global velocity $u_g(x)$ on the Eulerian mesh and interpolate onto the Lagrangian mesh.
3. Compute the total velocity $u(x) = u_l(x) + u_l(x)$ and update the position of the Lagrangian points.
4. Compute the force density $f$ on the membrane surface.
5. Calculate the global forcing $p_g(x)$ on the Eulerian mesh and boundary conditions for next time step (item 2). The boundary conditions read $u_g = u_w - u_l$, with $u_w$ prescribed at the wall.
6. Repeat from 1.
2.3 Nondimensionalization

We use Capillary number \( Ca \) to quantify the ratio between the viscous force and the elastic force. The definition of \( Ca \) varies for the two flow cases studied, namely, a capsule under constant shear flow and in plane Poiseuille flow for the 90-degree bend. For the first configuration, it is defined as \( Ca = \frac{\eta \dot{\gamma}a}{E_s} \), where \( \dot{\gamma} \), \( a \) and \( E_s \) are the flow shear rate, the cell radius and the surface shear elastic modulus. For the second configuration, \( Ca = \frac{\eta \bar{U}}{E_s} \), where \( \bar{U} \) is the mean bulk velocity of the channel flow. Besides the reduced area dilation modulus \( C \) introduced in equation 8, we also introduce the reduced bending modulus \( C_B \) defined as \( C_B = \frac{E_B a^2}{E_s} \).

2.4 Grid and boundary conditions

To simulate a cell going through an L-bend, we impose the parabolic Poiseuille velocity profile at the inlet. To produce moderate confinement, the channel width is set to \( 3a \). As we focus on the effect of the deformation around the corner, no confinement is added in the spanwise direction: the length of the computational domain in this direction is set to \( 10a \), with periodic boundary conditions. We use uniform spectral elements to discretize the whole domain. Each element is subdivided into arrays of Gauss-Lobatto-Legendre (GLL) nodes for the velocity and Gauss-Legendre (GL) nodes for the pressure field. In our simulation, the size of each element is \( a \) and 4 GLL points are used in each direction. The length of the upstream and downstream channels is equal to \( 20a \) and the total number of velocity grid points is therefore 82560.

3 RESULTS

3.1 Code Validation

The code has been validated first against the data provided in [13]. In this work, the authors report the solution for the motion of liquid capsules under constant shear flow. Fluid viscosities inside and outside the capsule are matched, and a neo-Hookean model of the membrane elasticity is adopted, while the capsule bending stiffness is neglected. The results are presented in Fig. 1 a) in terms of capsule deformation measured by the Taylor parameter

\[
D = \frac{L - B}{L + B}
\]

where \( L \) and \( B \) denote the maximum and the minimum diameters in the shear plane of the ellipsoid of inertia of the deformed capsule. The code was also validated against the data provided by Huang et al. [6]. Here we do not consider pre-inflated capsule, but instead include a finite bending stiffness. Fig. 1b) reports the time evolution of the deformation for a capillary number \( Ca = 0.15 \). The maximum discrepancy between the solid line, our results, and the dotted line, taken from [6], is around 3%. Note however that Huang et al. consider small but finite Reynolds number while we solve the Stokes...
equations for inertialess flow. Finally, we validated against the results presented in Lac et al. [9]. These authors consider the hydrodynamic interaction between two pre-inflated neo-Hookean capsules under shear flow. The two capsules are slightly pre-stressed to avoid those deformation instabilities observed for a single capsule in simple shear flow in the presence of compression and negative tension. The results in Fig. 1c) display the vertical distance versus the horizontal distance of the cell centers during approach and departure. The curve obtained by our simulation perfectly matches the results in [9].

3.2 L-bend results

We report data of the deformation of a capsule passing through an L-bend. Fig. 2 shows the deformation of a capsule with capillary number $Ca = 0.15$, where the color code indicates the magnitude of the stress. The results of the first simulations clearly indicate the importance of bending stiffness as one of the governing parameters. This is very often neglected in numerical simulations of capsules and vesicles. Here we show that when flowing in bends large deformations and negative tensions easily arise. These cause bending of the membrane and can accurately be captured only including bending stiffness in the model. If this is not the case, numerical dissipation may still cure numerical instabilities but the results are then grid-dependent. To give an idea of the stress acting on a typical cell, we report also dimensional values. Assuming a flow rate of around $10 \mu l/h$, and as values for the coefficients describing the cell membrane $E_S^* = 2.1 \cdot 10^{-6} N/m$ and
Figure 2: Capsule position in the microchannel at different instants of the simulation (A-F). Capillary number $Ca = 0.15$.

$E_B^* = 1.8 \cdot 10^{-19} \ N \cdot m$ [4, 18], the maximum stress experienced by the capsule is around 1.5 $Pa$ and it increases to 5 $Pa$ for more flexible capsules with $Ca = 0.45$.

Fig. 3 displays the area change of the capsule for three different capillary numbers: $Ca = 0.15$ (blue line), $Ca = 0.3$ (green line) and $Ca = 0.45$ (red line). The capsules with lower capillary number undergo a larger deformation in the upstream channel (initial dilation) and at the bend before recovering the original shape in the straight channel. In the lower panels in the figure one can see how the initial shape varies from bullet-type at low Capillary numbers to parachute-like for the more floppy capsules.

The same configuration is used to simulate a membrane obeying the Skalak model. Fig. 4 shows the change of total area of a capsule going through the 90-degree bend versus time for $Ca = 0.3$ and three different values of reduced area dilation modulus: $C = 1$, $C = 5$ and $C = 10$, together with a visualization of the cell shape at different positions inside the channel. For large $C$, i.e. large resistance to area changes, we observe lower deformations at the corner as well as a lower propagation velocity (the cell reaches the bend at a slightly later time). For $C = 1$, the cell already shows a larger area (6% higher) in the straight upstream channel that increases to about 15% around the corner.

Next we would like to compare the results for non-Hookean capsule with those for a membrane defined by the Skalak model. These models are the two most commonly used for capsules: neo-Hookean model is typically used for artificial capsules whereas
the Skalak model is used to study red blood cells. The time variation of the total area change at $Ca = 0.3$ for both models is reported in Fig. 4a). The capsule defined by the

Figure 3: a) Area change of a capsule in motion for three cases with different capillary number: blue line $Ca = 0.15$, green line $Ca = 0.3$, red line $Ca = 0.45$; b) Shape change of the capsules for the three capillary numbers.

Figure 4: a) Area change of a capsule in motion for $Ca = 0.3$ and three different values of reduced area dilation modulus: blue line $C = 1$, green line $C = 5$, red line $C = 10$; b) Shape change of the capsules for the three reduced area dilation moduli.
The neo-Hookean model undergoes the largest deformation at the corner, followed by a small compression in the downstream branch \((t \approx 150)\). The presence of the corner is also felt earlier when there is no direct limitation on the admissible variations of the surface area.

4 CONCLUSIONS

To obtain a parametric quantification for the relations between flow rate and maximum stress/strain as well as the details of cell deformation while flowing through microfluidic chip for cell isolation highly accurate numerical simulations are necessary. Here we have presented a novel implementation of the General Geometry Ewald Method (GGEM) proposed in [13, 8] suited for realistic micro-devices.

Our simulations clearly underline the importance of bending stiffness as one of the governing parameters, very often neglected in numerical simulations. Here we show that large deformations and negative tensions easily arise, causing bending of the membrane, which can accurately be captured only including bending stiffness in the model. Therefore, this work allows us to accurately investigate cell behavior in flow in a complex geometry and to compare the behavior of different models. In the future, we plan to use our new numerical tool to optimize the design of effective micro-devices for cell isolation, where the prediction of deformation ans stress acting on cell during processing is essential. Achieving this still represents one of the most significant challenges in the field of stem cells.

REFERENCES


NUMERICAL SIMULATION OF EXPLOSIVE FRACTURING WITH SMOOTHED PARTICLE HYDRODYNAMICS

ZHIMING GUO*†, YONGXING SHEN* AND MOUBIN LIU‡

*Laboratori de Càlcul Numèric (LaCaN)
Universitat Politècnica de Catalunya (UPC BarcelonaTech)
C/ Jordi Girona 1-3, Campus Nord, 08034 Barcelona, Spain
Email: guozhiming@live.cn (Guo), yongxing.shen@upc.edu (Shen)
Web page: http://www.lacan.upc.edu/

†School of Mechatronic Engineering
North University of China
Xueyuan Road 3, 030051 Taiyuan, Shanxi, China
E-mail: guozhiming@live.cn - Web page: http://www.nuc.edu.cn/

‡Key Laboratory for Mechanics in Fluid Solid Coupling Systems
Institute of Mechanics, Chinese Academy of Sciences
#15, Bei Si Huan Xi Lu, Haidian District, Beijing, China 100190
Email: liumoubin@imech.ac.cn - Web page: http://lmfs.imech.cas.cn/

Key words: Explosive Fracture, Smoothed Particle Hydrodynamics, Elasto-damage Model, Rock Fracture, Meshfree Method

Abstract. In this paper we study explosive fracturing with smoothed particle hydrodynamics (SPH). As a particle based Lagrangian method, SPH is particularly suited to the analysis of fracture due to its full Lagrangian frame and capacity to model large deformation. We adopt the Jones-Wilkins-Lee equation as equation of state of the trinitrotoluene (TNT) explosive and a continuum elasto-damage model to predict the fracture of the rock. We predict the evolution of damage using the strain history of each particle. To strengthen the interaction of coupling interfaces we use a penalty function to avoid penetration between different material particles.

1 INTRODUCTION

Explosive fracturing is an important approach in industries such as shale gas exploration, shale oil exploration, rock excavation, and mineral and ore processing. This technique has been effectively used by a few countries. Since the earliest days of blasting with black powder there have been steady developments in explosives, detonation and the
understanding of the mechanics of rock fracture by explosives, but the fundamental mechanisms of rock damage are not yet fully understood in many aspects. Hence, approaches for studying fracture have been heavily dependent on empirical relations based on experimental data of the laboratory scale. Duvall and Fogelson [1], Langefors and Khilstrom [2], and others have published blast damage criteria for buildings and other surface structures. All of these criteria relate blast damage to peak particle velocity resulting from the dynamic stresses induced by the explosion. Both the dynamic stresses induced by the detonation and the expanding gases produced by the explosion play important roles in the fragmentation process, but it is difficult to study the details about this mechanism.

On the other hand, numerical simulations, as an effective mean, can help understand and predict complex exploding and fracture processes. Of the many available methods, mesh-based methods seem inadequate in tracking the motions of the fragments. Instead, most existing methods for this problem fall in one of two categories of methods.

One category of methods, such as the discrete element method [3, 4], explicitly tracks the motion of fragments. Another category of methods models the solid as a damaging body, such as [5, 6, 7, 8, 9]. Damage inhibits the transmission of tensile stress between particles, and once it reaches unity, the particle is unable to transmit tensile stress, resulting in a macro-crack. Connected macro-cracks lead to complete fragmentation. We also mention here damage models for various kinds of solids [10, 11, 12, 13].

One possibility of discretizing the continuous equations of a damage model is to use the smoothed particle hydrodynamics (SPH). SPH was initially developed to solve astrophysical problems [14] Due to its special features and advantages, it has been applied to various areas in engineering and sciences [15].

In this paper, we study the explosive fracturing of rock mass with the SPH method. This process involves the explosion of TNT, the interaction of exploding gas and rock, and the fracture of rock. To model the rock, we employ a modified version of the Grady-Kipp model [16]. The SPH is suited for modeling explosive fracture because it is insensitive to disordered particle distribution, a consequence of the fast expansion of explosive gas and rapid deformation and fast damage on materials.

The rest of this work is organized as follows. In Section 2 we state the problem we want to solve and give the governing equations. In Section 3 we discretize these governing equations with the SPH method. Finally in Section 4 we verify our results by numerical examples.

2 PROBLEM STATEMENT

We study the damage of a piece of rock mass during the detonation and explosion of trinitrotoluene (TNT) explosive in a preexisting crack. For simplicity we will confine ourselves to two-dimensions (2D) within this work. In this section we will introduce the governing equations that model the coupled problem of explosive fracturing. In particular, Section 2.1 states the conservation laws and Section 2.2 gives the constitutive equations and equations of state for both phases: the damaging rock and the exploding gas.
2.1 Conservation laws

In the Langrangian frame, the motion of both the rock mass and the explosive is governed by the following conservation equations:

\[
\begin{align*}
\frac{Dx^\alpha}{Dt} &= v^\alpha, \\
\frac{Dv^\alpha}{Dt} &= \frac{1}{\rho} \frac{\partial \sigma^{\alpha\beta}}{\partial x^\beta}, \\
\frac{D\rho}{Dt} &= -\rho \frac{\partial v^\beta}{\partial x^\beta}, \\
\frac{De}{Dt} &= \sigma^{\alpha\beta}_d \frac{\partial v^\alpha}{\partial x^\beta},
\end{align*}
\]

(1)
of which the first equation is the definition of the velocity, and the second, third, and fourth equations represent the conservation of momentum, mass and energy, respectively. Here superscripts \( \alpha, \beta = 1, 2 \) denote the component index in the Cartesian coordinates, for which repeated indices imply summation from 1 to 2. Quantities \( x, v, \rho, e, \) and \( \sigma_d \) are the position vector, velocity vector, density, energy, and the damaged (Cauchy) stress tensor, respectively. Among these, the damaged stress tensor describes the effect of the material damage in terms of a modification of the undamaged stress tensor \( \sigma \). Derivative \( D/Dt \) denotes the total derivative, or termed material derivative.

2.2 Constitutive equations and equations of state

Below we specify the constitutive equations and equations of state for the rock mass (Section 2.2.1) and the explosive gas (Section 2.2.2). In this part we closely follow [16].

2.2.1 The rock mass

In the sequel we first describe how we model the undamaged stress tensor \( \sigma \), and then detail how we take into account the damage effect.

The undamaged stress tensor. For the rock, the undamaged stress tensor consists of two parts:

\[
\sigma^{\alpha\beta} = -P \delta^{\alpha\beta} + S^{\alpha\beta},
\]

(2)
where \( \delta^{\alpha\beta} \) is the Kronecker delta, \( P \) is the elastic pressure, and \( S \) is the deviatoric stress tensor. We assume the elastic pressure is proportional to the change in density:

\[
P = c^2 (\rho - \rho_0),
\]

(3)
where \( \rho_0 \) is the reference density, \( \rho \) is the current density and \( c \) is the longitudinal wave speed of the solid material.
On the other hand, we can express the deviatoric stress \( S \) with Hooke’s law using the Jaumann rate equation [17]:

\[
\frac{DS^{\alpha\beta}}{Dt} = 2G \left( \dot{\ell}^{\alpha\beta} - \frac{1}{3} \ell^{\gamma\gamma} \delta^{\alpha\beta} \right) + S^{\alpha\gamma} \Omega_{\gamma\beta} - \Omega^{\alpha\gamma} S_{\gamma\beta},
\]

where \( G \) is the shear modulus, \( \ell \) is the strain rate, and \( \Omega \) is the rotation tensor:

\[
\dot{\ell}^{\alpha\beta} = \frac{1}{2} \left( \frac{\partial v^\alpha}{\partial x^\beta} + \frac{\partial v^\beta}{\partial x^\alpha} \right), \quad \Omega^{\alpha\beta} = \frac{1}{2} \left( \frac{\partial v^\alpha}{\partial x^\beta} - \frac{\partial v^\beta}{\partial x^\alpha} \right).
\]

The damaged stress tensor. We adopt a modified form of Grady-Kipp model [16] to predict rock damage based on the local stress history and flaw distribution. This model is based on the Weibull distribution of flaws. As many other models, the modified Grady-Kipp damage model defines a scalar parameter \( D \) such that \( 0 \leq D \leq 1 \) to characterize the volume-averaged micro-fracture of the volume of material based on the strain history. Material with \( D = 0 \) is undamaged and is able to transmit the full tensile load, whereas material with \( D = 1 \) is fully damaged and cannot transmit any tensile load, thus creating a partial macro-crack [10].

Specifically, the damaged stress \( \sigma_d \) and the undamaged stress \( \sigma \) are related as follows. Let the principal components of \( \sigma \) be \( \sigma^1 \) and \( \sigma^2 \) and the corresponding principal directions be \( n^1 \) and \( n^2 \). Then \( \sigma \) can be written as

\[
\sigma = \sum_{a=1}^{2} \sigma^a n^a \otimes n^a.
\]

The principal components of \( \sigma_d \) are then obtained from

\[
\sigma^a_d = \begin{cases} 
(1 - D)\sigma^a, & \text{if } \sigma^a \geq 0, \\
\sigma^a, & \text{otherwise,}
\end{cases}
\]

for \( a = 1, 2 \). In other words, tensile stresses are only partially transmitted through damaged materials while compressive stresses are totally transmitted.

As a result,

\[
\sigma_d = \sum_{a=1}^{2} \sigma^a_d n^a \otimes n^a.
\]

At each point, the damage \( D(t) \) evolution is given by

\[
\frac{d D^{1/3}}{dt} = \begin{cases} 
C \varepsilon_{eff}^{m/3}, & \text{if } t > t_{onset}, \\
0, & \text{otherwise,}
\end{cases}
\]

(6)
where we have absorbed the material parameters in [16, 10] into a single one, \( C \). The effective tensile strain \( \varepsilon_{\text{eff}} \) is given by Melosh et al. [18]:

\[
\varepsilon_{\text{eff}} = \frac{\sigma_{\text{max}}}{\left( K + \frac{4}{3} G \right)},
\]

where \( \sigma_{\text{max}} = \max\{0, \sigma^1, \sigma^2\} \) is the maximum positive principal stress, and \( K \) is the bulk modulus of the material.

In the subsequent calculations, the onset time \( t_{\text{onset}} \) is defined as the first instant when \( \varepsilon_{\text{eff}} \geq \varepsilon_{\text{min}} \), where

\[
\varepsilon_{\text{min}} := \left( \frac{m_j}{\rho_j} \right)^{-1/m},
\]

where \( m_j \) and \( \rho_j \) are the particle mass and density to be introduced later, and \( k \) and \( m \) are parameters of the Grady-Kipp model.

### 2.2.2 The explosive gas

For the explosive gas, as the isotropic pressure is much larger than components of viscous shear stress, the viscous shear stress can be neglected, hence

\[
\sigma^{\alpha\beta} = -p\delta^{\alpha\beta},
\]

where the hydrodynamic pressure \( p \) is related to the internal energy \( e \) according to the standard Jones-Wilkins-Lee (JWL) equation of state [19]:

\[
p = A \left( 1 - \frac{\omega \eta}{R_1} \right) e^{\frac{R_1}{\eta}} + B \left( 1 - \frac{\omega \eta}{R_2} \right) e^{\frac{R_2}{\eta}} + \omega \eta \rho_0 E,
\]

where \( \eta = \rho/\rho_0 \) is the ratio of the density of the explosive gas to the initial density of the original explosive. Parameters \( A, B, R_1, R_2 \) and \( \omega \) are coefficients obtained by fitting experimental data, and \( E \) is the initial internal energy of the high explosive per unit mass. Values of the corresponding coefficients are listed in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \rho_0 )</th>
<th>( A )</th>
<th>( B )</th>
<th>( R_1 )</th>
<th>( R_2 )</th>
<th>( \omega )</th>
<th>( E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity</td>
<td>1630</td>
<td>317.2</td>
<td>3.21</td>
<td>4.15</td>
<td>0.95</td>
<td>0.30</td>
<td>4290</td>
</tr>
<tr>
<td>Unit</td>
<td>kg \cdot m^{-3}</td>
<td>GPa</td>
<td>GPa</td>
<td></td>
<td></td>
<td></td>
<td>kJ \cdot kg^{-1}</td>
</tr>
</tbody>
</table>
2.3 Coupling conditions

At the interface, the velocity and pressure of the two phases are continuous, i.e.,

\[ v_{\text{gas}}^\alpha = v_{\text{rock}}^\alpha, \quad p_{\text{gas}} = p_{\text{rock}}. \]

3 SMOOTHED PARTICLE HYDRODYNAMICS METHOD

3.1 Basic concept

We start by representing an arbitrary function \( f(x) \) with discrete values at particles located at \( x_1, \ldots, x_N \), denoted \( f_1, \ldots, f_N \):

\[ f(x) = \sum_{j=1}^{N} \frac{m_j}{\rho_j} f_j W(x - x_j, h), \quad (9) \]

where \( m_j \) and \( \rho_j \) denote the mass and density of particle \( j \), respectively, \( h \) is the smooth length, and \( W \), usually called a kernel or kernel function, is a smoothing function representing a weighted contribution of the particles. This smoothing function should satisfy some basic requirements. Detailed discussions on the smoothing function, its basic requirements and constructing conditions can found in [15]. In this work, we choose the Gaussian kernel

\[ W(r,h) = W(r,h) = \alpha_d \exp \left(-\frac{r}{h}\right)^2, \]

where \( r := |r| \) and \( \alpha_d \) is a dimension-dependent constant related to the smoothing length. In 2D, \( \alpha_d = (\pi h^2)^{-1} \).

Any partial derivative of \( f(x) \) can then be represented as:

\[ \frac{\partial f(x)}{\partial x^\beta} = \sum_{j=1}^{N} \frac{m_j}{\rho_j} f_j \frac{\partial W(x - x_j, h)}{\partial x^\beta}, \quad (10) \]

where \( \frac{\partial c}{\partial x^\beta} \) is an approximation to \( \partial / \partial x^\beta \). This approximation involves a kernel correction, which we will elaborate in Section 3.2.

Specializing (9) and (10) to \( x = x_i \) and using \( f_i = f(x_i) \), we have, for \( i = 1, \ldots, N \),

\[ f_i \doteq \sum_{j=1}^{N} \frac{m_j}{\rho_j} f_j W_{ij}, \quad \left( \frac{\partial f}{\partial x^\beta} \right)_i \doteq \sum_{j=1}^{N} \frac{m_j}{\rho_j} f_j \frac{\partial W_{ij}}{\partial x_i^\beta}, \quad (11) \]

where

\[ W_{ij} := W(|x_i - x_j|) = W(|x_i - x_j|, h). \]

Ideally, by applying (11) to \( f(x) = 1 \), we have the approximate consistency equations,

\[ \sum_{j=1}^{N} \frac{m_j}{\rho_j} W_{ij} \doteq 1, \quad \sum_{j=1}^{N} \frac{m_j}{\rho_j} \frac{\partial W_{ij}}{\partial x_i^\beta} \doteq 0. \quad (12) \]

where \( i = 1, \ldots, N \) and \( \beta = 1, 2 \).
### 3.2 Kernel gradient correction

In this paper, we improve the kernel gradient in SPH approximations with a kernel gradient correction (KGC) technique [20]. In the KGC technique, a modified or corrected kernel gradient is obtained by multiplying the original kernel gradient with a local reversible matrix $L_i$, which is obtained from Taylor series expansion method. In 2D, the new kernel gradient of the smoothing function $\nabla_i^c W_{ij}$ can be obtained as follows

$$\nabla_i^c W_{ij} := L_i \nabla_i W_{ij},$$

where

$$L_i := \left( \sum_j m_j \frac{1}{\rho_j} \begin{bmatrix} x_{ji}^1 \frac{\partial W_{ij}}{\partial x_i^1} & x_{ji}^2 \frac{\partial W_{ij}}{\partial x_i^2} \\ x_{ji}^1 \frac{\partial W_{ij}}{\partial x_i^2} & x_{ji}^2 \frac{\partial W_{ij}}{\partial x_i^1} \end{bmatrix} \right)^{-1},$$

(13)

where

$$\nabla_i^c W_{ij} := \left\{ \frac{\partial W_{ij}}{\partial x_i^1}, \frac{\partial W_{ij}}{\partial x_i^2} \right\}, \quad \nabla_i W_{ij} := \left\{ \frac{\partial W_{ij}}{\partial x_i^1}, \frac{\partial W_{ij}}{\partial x_i^2} \right\}.$$

### 3.3 SPH discretization of motion equations

We will obtain SPH discretization equations by applying SPH approximations (11) along with the approximate consistency equations (12) to (1). For stability concerns, we introduce artificial viscosity $\Pi_{ij}$ and artificial heat $H_i$ [15]. With these, the SPH discretization equations become

$$\begin{cases} \frac{Dx_i^\alpha}{Dt} = v_i^\alpha, \\
\frac{Dv_i^\alpha}{Dt} = \sum_{j=1}^N m_j \left( \frac{\sigma_{d,i}^{\alpha\beta}}{\rho_i^2} + \frac{\sigma_{d,j}^{\alpha\beta}}{\rho_j^2} + \Pi_{ij} \right) \frac{\partial W_{ij}}{\partial x_i^\beta}, \\
\frac{D\rho_i}{Dt} = \rho_i \sum_{j=1}^N m_j \left( \frac{v_i^\beta - v_j^\beta}{\rho_i} \right) \frac{\partial W_{ij}}{\partial x_i^\beta}, \\
\frac{De_i^\alpha}{Dt} = \frac{1}{2} \sum_{j=1}^N m_j \left( \frac{p_i}{\rho_i} + \frac{p_j}{\rho_j} + \Pi_{ij} \right) \left( v_i^\beta - v_j^\beta \right) \frac{\partial W_{ij}}{\partial x_i^\beta} + \frac{1}{\rho_i} S_i^{\alpha\beta} \ell_i^{\alpha\beta} + H_i, \end{cases}$$

(14)

where subscripts $i$ and $j$ denote particle numbers, and all approximate equal signs “≈” are replaced with equal signs for brevity. The derivation of all equations of (14) is straightforward except for the last one, for which we refer the reader to [15, Chapter 4].

For particle $i$, the SPH equations for the strain rate $\ell_i$ and the rotation tensor $\Omega_i$
defined in (5) are:

\[
\ell_{i}^{\alpha\beta} = \frac{1}{2} \sum_{j} \frac{m_{j}}{\rho_{j}} \left[ (v_{j}^{\alpha} - v_{i}^{\alpha}) \frac{\partial W_{ij}}{\partial x_{i}^{\alpha}} + (v_{j}^{\beta} - v_{i}^{\beta}) \frac{\partial W_{ij}}{\partial x_{i}^{\beta}} \right],
\]

(15)

\[
\Omega_{i}^{\alpha\beta} = \frac{1}{2} \sum_{j} \frac{m_{j}}{\rho_{j}} \left[ (v_{j}^{\alpha} - v_{i}^{\alpha}) \frac{\partial W_{ij}}{\partial x_{i}^{\alpha}} - (v_{j}^{\beta} - v_{i}^{\beta}) \frac{\partial W_{ij}}{\partial x_{i}^{\beta}} \right].
\]

(16)

### 3.4 Artificial viscosity and artificial heat

The artificial viscosity is used in SPH method to stabilize the numerical scheme, prevent particle penetration and capture shock waves. In this paper, we employ the standard artificial viscosity [15]:

\[
\Pi_{ij} = \begin{cases} 
-\alpha_{\Pi} \phi_{ij} + \beta_{\Pi} \phi_{ij}^{2} & \text{if } \mathbf{v}_{ij} \cdot \mathbf{x}_{ij} < 0, \\
0 & \text{otherwise},
\end{cases}
\]

(17)

where

\[
\phi_{ij} := \frac{\mathbf{h}_{ij} \mathbf{v}_{ij} \cdot \mathbf{x}_{ij}}{|\mathbf{x}_{ij}|^2 + (0.1 \mathbf{h}_{ij})^2}, \quad \mathbf{v}_{ij} := \mathbf{v}_{i} - \mathbf{v}_{j}, \quad \mathbf{x}_{ij} := \mathbf{x}_{i} - \mathbf{x}_{j},
\]

\[
\mathbf{c}_{ij} := \frac{1}{2} (c_{i} + c_{j}), \quad \rho_{ij} := \frac{1}{2} (\rho_{i} + \rho_{j}), \quad \mathbf{h}_{ij} := \frac{1}{2} (h_{i} + h_{j}).
\]

In the above equations, \(\alpha_{\Pi}\) and \(\beta_{\Pi}\) are constants that are typically set around 1.0. Symbol \(c_{i}\) denotes the longitudinal wave speed of particle \(i\). The viscosity term associated with \(\alpha_{\Pi}\) produces a bulk viscosity, while the second term associated with \(\beta_{\Pi}\), which is intended to suppress particle interpenetration at high velocity, is similar to the von Neumann-Richtmyer artificial viscosity.

To compensate the heat due to artificial viscosity, we put in the system the following artificial heat expression:

\[
H_{i} = 2 \sum_{j=1}^{N} \frac{\bar{q}_{ij}}{\mathbf{p}_{ij} |\mathbf{x}_{ij}|^2 + (0.1 \mathbf{h}_{ij})^2} \mathbf{v}_{ij} \cdot \nabla_{ij} W_{ij},
\]

where

\[
\bar{q}_{ij} := q_{i} + q_{j}, \quad q_{i} := \alpha_{\Pi} h_{i} \rho_{i} c_{i} |\nabla \cdot \mathbf{v}_{i}| + \beta_{\Pi} h_{i}^{2} \rho_{i} |\nabla \cdot \mathbf{v}_{i}|^2.
\]

### 3.5 Interaction between phases

The summation in (14) only takes into account particles of the same species. Between different particles, in this case between a solid particle and a gas particle, a force is
applied to model their interaction. The force on particle \(i\) by particle \(j\) is expressed as a Lenard-Jones type interaction:

\[
f_{ij} = \begin{cases} 
D_{\text{int}} \left[ (r_0/r_{ij})^4 - (r_0/r_{ij})^2 \right] (x_{ij}/r_{ij}^2), & \text{if } r_{ij} < r_0, \\
0, & \text{otherwise,}
\end{cases}
\]

where \(D_{\text{int}}\) is a parameter proportional to the square of the largest velocity, and \(r_0\) is on the scale of the initial spacing of the particles.

### 3.6 Time integration

We use the leap-frog (LF) method to advance in time because it is efficient and needs to store only a small amount of memory [21]. As a typical explicit method, it is subject to the Courant-Friedrich-Levy (CFL) condition for stability, which typically results in a time step proportional to the smoothing lengths \(h\). In this work, the time step is taken as,

\[
\Delta t = \min_i \xi h_i \left( \nabla \cdot \mathbf{v}_i + c_i + 1.2(\alpha_\Pi c_i + \beta_\Pi |\nabla \cdot \mathbf{v}_i|) \right),
\]

where \(\xi = 0.3\) is the Courant number.

### 4 NUMERICAL EXAMPLES

As a preliminary study, we model a typical sandstone of a rectangular shape with an elliptical void. The size of the rectangle is \(2m \times 1m\) and the major and minor axes of the void are \(1.6m\) and \(0.2m\), respectively. The sandstone has a bulk modulus of \(12.2\) GPa, a shear modulus of \(2.67\) GPa, and a density of \(2300\) kg/m\(^3\). Here the Weibull damage parameters involved in the damage model are \(k = 6.35 \times 10^{46}\) and \(m = 12.8\), respectively. As a result, the parameter \(C\) in the Grady-Kipp model takes the value of \(4.604 \times 10^{18}\). The geometry and material properties of the rock are from [22]. The boundaries of the rock are free, which will be changed to a fixed boundary in the near future. The specimen domain was discretized using particles in an initial rectangular grid pattern with a spacing of \(0.02m\).

Figure 1 illustrates the initial configuration, Figure 2 and Figure 3 show the snapshots of the damage parameter at different times with free boundary and fixed boundary respectively. We can see that damage originates from the part of the rock closest to the
Figure 2: Snapshots of the damage parameter $D$ at various times with free boundary. We can see that damage originates from the part of the rock closest to the explosive and then the tips of the void due to expansion of explosive gas. There is spalling at upper and down boundary.

explosive and then the tips of the void due to expansion of explosive gas. In the example with free boundary, the spalling is accompanied at top and bottom of rock mass. In the example with fixed boundary, obviously we can see spalling and collapse of inner domain of the rock mass and the artifact due to the fixed boundary, which we will change to a non-reflected boundary in the near future.

REFERENCES


Figure 3: Snapshots of the damage parameter \( D \) at various times with fixed boundary. We can see that damage originates from the part of the rock closest to the explosive and then the tips of the void, as expected. Nevertheless, we also see the artifact due to the fixed boundary, which we will change to a non-reflected boundary in the near future.


NUMERICAL SIMULATION OF LANDMINE EXPLOSIONS: COMPARISON BETWEEN DIFFERENT MODELLING APPROACHES

CLAUDIO FICHERA*, LORENZO PERONI*, MARTINA SCAPIN*

* Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Turin, Italy
e-mail: claudio.fichera@polito.it, web page: http://www.polito.it

Key words: Explosions, Landmine, ALE, FSI, CONWEP, THUMS.

Abstract. Until decade ago the design of mechanical structures, having to resist to explosive events, was mainly performed using experimental tests with explosive materials. In the last years, numerical methods are assuming importance thanks to the following advantages: high cost reduction, flexibility in investigating different scenarios and the chance to study explosive phenomena without risks. An explosion is a complex and multidisciplinary subject. It involves a large number of physical parameters which influence the amount of energy transferred to the target above the detonation. The aim of this paper is to describe numerical models to simulate landmine explosion and blast loading on structures, using different approaches: an Arbitrary Lagrangian Eulerian (ALE) mesh and a pure Lagrangian mesh. For what concerns the ALE simulations, three different cases are analyzed. First of all, the numerical model of the landmine explosion is validated through the comparison with experimental data. The same model is then used to evaluate the effect of detonations against two structures, using a fluid-structure interaction algorithm: a steel plate and a human leg. For this type of simulations, an Eulerian approach is needed, in order to reproduce the expansion of the mix of sand, air and gas against the target. When the gas encounters the target a fluid structure interaction algorithm (FSI) determines the pressure values, which are transferred from the Eulerian parts to the Lagrangian ones. The main disadvantage of an ALE approach is the large computational time, which is further aggravated by the need to use quite fine mesh resolution to adequately reproduce the air shock. For this reason it is interesting to use 2D modeling. The second approach is based on empirical airblast equations developed by Kingery and Bulmash, for the application of pressure loads due to explosives in conventional weapons, and was implemented in LS-DYNA by Randers-Pehrson and Bannister. This methodology is applied to simulate the detonation against the plate and the results are compared with the corresponding results obtained using an ALE approach.

1 INTRODUCTION

Recently, in the explosive phenomena field, the numerical simulations are assuming a relevant importance for structure design. The numerical approach brings significant advantages compared to experimental tests: no risks, high cost reduction and great opportunity to study different scenarios just changing the model parameters. Usually, the
experimental repeatability is quite difficult to achieve, so it is complicated to carry out right and univocal considerations. Moreover, the explosion tests are very complex to perform and require high experience and high instrumentations cost. On the other hand, the numerical simulation allows performing a first stage of Design of Experiment, since it make possible the study of many different scenarios, changing the parameters involved. By the comparison and the analysis of the numerical results, it is possible to better understand the phenomena evolution, so it is possible to focus the attention on the critical situations, which in turn are useful to validate the numerical results. In any case, frequently, the experimental tests cannot be realized, because restrictive laws on explosive materials are in force, so the numerical simulation represents the unique tool available for the design.

In general, the numerical simulations of explosions are very complicated, because many factors have to be taken into account: the explosive material properties, the properties of the medium in which the shock-wave is transmitted and the target type. Besides, the numerical solution is further complicated by the simultaneous presence of fluids and solids. It is recommended to use a combination of pure Eulerian mesh, used for modeling fluids, and pure Lagrangian mesh, used for modeling solid structures. In order to allow the expansion and compression of neighboring fluids in the same region, it is necessary to define an appropriate algorithm for simulating the mixing of different fluids. The fluid-structure interaction is another feature to take into account, which is very expensive from a computational point of view. The fluid-structure interaction algorithm allows transferring the pressure values, generated in the fluid, to the target structure. Finally, the numerical model needs an equation of state to represent the detonation expansion of the gas produced by the detonation.

In this work, a benchmark numerical model has been realized in LS-DYNA [1] in order to reproduce a landmine explosion and the results have been validated through the comparison with experimental data obtained from Canadian Defense Department [2] with a 3D and 2D approaches. Following, the same landmine model has been used to study the effects of the explosion against two different structures: a steel square plate [3] and the human leg extracted from THUMS [4].

The results obtained for the plate are then compared with the results obtained for the same case using the airblast model, implemented in LS-DYNA, with a pure Lagrangian mesh. The empirical blast equations were developed by Kingery and Bulmash [5] for the application of pressure loads due to explosive in conventional weapons. Kingery and Bulmash performed a series of tests, varying the charge weights, and used curve fitting techniques to represent the data with polynomial equations. The equations were then implemented in the computer program CONWEP, which was introduced in LS-DYNA by Randers-Pehrson and Bannister [6]. In the present work, this model is applied to simulate the detonation against the square plate and the results are compared with the corresponding results obtained using the ALE approach.

2 EXPLOSIVE MATERIAL EQUATION OF STATE

In the numerical models of such events, the equation of state (EOS) of the explosive, which expresses the pressure as a function of density and energy, is a crucial aspect. In the past, different theoretical and empirical approaches have been developed to describe the explosions and the behavior of the gas produced during the detonation. The Jones-Wilkins-
Lee (JWL) equation of state, which is implemented in LS-DYNA, is the most commonly used thanks to its simplicity. Moreover, a relevant number of high explosive materials have a good representation using this equation, which defines the pressure as:

\[
P = A \left(1 - \frac{\sigma}{R_1 v}\right) e^{-R_1 v} + B \left(1 - \frac{\sigma}{R_2 v}\right) e^{-R_2 v} + \frac{\sigma E}{v}
\]

where \( v \) is the relative volume, \( E \) is the internal energy and \( A, B, \omega, R_1 \) and \( R_2 \) the input constants.

### 3 MULTI-MATERIAL GROUP

During an explosion different materials are mixed together generating the expansion of some fluids inside other ones. Generally, two kinds of fluids take part in the explosion: the gas produced by the detonation and the fluids in which the explosion propagates (air, water, sand, etc.). For these reasons, the Eulerian mesh is recommended, since it is appropriated to describe the fluids behavior. Moreover, it is necessary to define a common space where the fluids can interact each others, which is represented by the Multi-Material Groups (MMG) and presents a common mesh for all the material belonging the same MMG. The definition of MMG is such that each element of the discretized volume can include, at the actual timestep a fluid different from that of the previous timestep, simulating the expansion of a fluid inside the another ones.

In LS-DYNA, when an Eulerian mesh is adopted, the solution is obtained in two steps. At the beginning, the problem is solved from a Lagrangian point of view, in which the mesh deforms following the material flow. In the second step, the nodes are considered to be in the initial position and the solution is mapped from the deformed mesh.

### 4 FLUID-STRUCTURE INTERACTION

The interaction between fluid materials, which are modeled with an Arbitrary Lagrange-Euler (ALE) mesh, and solids, for which the Lagrangian mesh is recommended, is a relevant factor in the simulation of explosive phenomena, in which high pressure can be generated in a very short time. To solve this problem a fluid-structure interaction algorithm (FSI) is necessary. The FSI is a multi-physic phenomenon where a fluid, acting against a structure, generates the shape structure modification due to pressure and shear loads.

Sometimes, the FSI could be stationary and this happens if the loads applied by the fluid are exactly balanced from the reaction force of the structure, so the fluid reaches strain equilibrium. In the explosion field, the FSI is a transient phenomenon and the structure deformation is dynamic and changes with time. Since the fluid-structure interaction algorithm is very time consuming, it is recommended to use it only in problems where high pressure impacts the target very quickly. When the FSI is more stationary, it is suggested to use simple contact algorithms.

The FSI algorithm is based on a soft coupling between Eulerian and Lagrangian solvers, which are dedicated, respectively, to obtain the fluid and structural solutions. The Lagrangian structure imposes the interface boundary location, the displacements and the velocity. This
information represent the interface conditions used by the fluid solver to compute the pressure to apply to the structural interface as exterior forces, which, in turn, represent the input for the structural solver. This means that at each timestep, the fluid and structural response are separately solved and then coupled together before starting the calculation for the next timestep (see Fig. 1).

![Fluid Solver EULERIAN Structural Solver LAGRANGIAN](image)

*Figure 1*: Scheme of the FSI algorithm.

The interaction between the pressure wave, generated by the explosion, and the structure, invested by it, produces the reflection of the pressure wave itself with the same sign. This makes the overpressure value to be increased with respect to the incident one. In the case of perfectly rigid structure the effective overpressure should be twice the incident one. If the structure intercepted by the pressure wave is deformable, this produces an immediate decreasing of the fluid pressure. As a matter of fact, the deformation produces the presence of some void zones close to the deformed surfaces, in which the fluid can expand reducing its pressure. This phenomenon becomes more relevant in case of rigid fluids, as e.g. for undersea explosions. The previous considerations make clear the advantages of using a FSI algorithm instead of simulating in a decoupled manner the fluid expansion and the structure deformation.

5 BENCHMARK MODEL: LANDMINE

In the technical-scientific literature, about explosion scenarios, the most part of the available experimental data regards landmine explosions tests. Therefore, a benchmark landmine model is realized in LS-DYNA and the numerical results are compared with the experimental ones, obtained by Bergon, Walker and Coffey in [2].

The test setup is shown in Fig. 2. A 100 g cylindrical charge, with a diameter of 64 mm and a height of 20 mm, of C-4 was used. A steel cylindrical container (inner diameter 889 mm, height 698 mm and thick 12.7 mm) is filled with dry sand (Silica 20), where the mine was buried. The sieve analysis of the sand showed that the diameter of the majority of the particles was between 160 and 630 microns, so the mean density is 1.8 kg/dm$^3$. The mine was buried at different depths of burial (DOB): 0 mm, 30 mm and 80 mm and for each case about six tests were performed. Pressure transducers were located above the soil surface at different heights (see Fig. 2) and at different depths in sand, with the aim of recording the trend of the pressure history vs. time and the time at which the pressure wave reaches the transducer. Other measurements regard the pit dimensions, the height and the diameter of the clouds of the produced gas and the height of the cloud of sand ejected in the atmosphere.
The numerical model (see Fig. 2) reproduces the geometry and the materials properties used for the material strength model and the equation of state are taken from [8]. The explosive is modeled combining the JWL equation of state with a model, which controls the detonation characteristics of the explosive. The steel pipe is modeled using an elastic-linear plastic strength model. The soil is modeled using an ad-hoc formulation, introduced in LS-DYNA for the description of soil and foam behavior, which allows defining the plastic yield function and the pressure vs. volumetric strain curve. Finally, the air is modeled from a pure hydrodynamic point of view using a polynomial equation of state, equivalent to the ideal gas law. Some of the models used are probably too simple, but the choice has been made on the basis of the available data in scientific literature.

As first attempt, the explosive, the air and the sand are modeled using 3D solid elements, while the steel pipe is modeled with 2D shell elements. The results show that, as expected, the explosion event is axisymmetric (at least as long as the pressure wave does not reach any boundaries). For this reason, the same case is also modeled using a 2D axysymmetric geometry (the ALE 2D axysymmetric option has recently been introduced in LS-DYNA). The aim is both reducing the computational time and having the possibility to increase the number of elements for studying the mesh influence on the solution. As a matter of fact, the main goal of this paper is to build reliable and stable numerical models of explosion events. The meaningfulness of the results is demonstrated through the comparison with experimental data. In any case, since a lot of data of the experimental tests are unknown, it is difficult to perfectly reproduce the same event. On the other hand, the perfect match between numerical and experimental results is not the main goal, since it would imply to optimize the materials parameters to obtain the best fit.

The comparison between numerical results and experimental data in terms of relative pressure vs. time curves are reported in Fig. 3 for the 3D case and the coarser 2D case. The comparison is made for the DOB 0 and DOB 8 in correspondence of the measuring point placed at 30 cm. The EOS introduced for the air is such that the air is at -1 bar at the initial condition, so in the diagram, the numerical pressure history starts when the pressure become positive.
Figure 3: comparison between numerical results and experimental data [2] in terms of relative pressure vs. time curves, obtained at 30 cm above the soil surface: (a) DOB=0 cm; (b) DOB=8 cm.

Figure 4: comparison between numerical results obtained for the 2D axysimmetric case varying the mesh dimension. The results are in terms of overpressure vs. time curves, obtained at 30 cm above the soil surface and in terms of spatial fluid distribution (at two different times).

Moreover, the numerical history is shifted in time in order to synchronize the time at which the peak of overpressure arrives. Looking at the results it is possible to conclude that the model is able to reproduce the same evolution obtained experimentally with a sufficient level of accuracy, especially for what concerns the case at DOB=0. Increasing the depth of burial
the differences between numerical and experimental results become more significant. This could also be correlated to the fact that it is quite difficult to correctly model the soil, since it is a very heterogeneous material. In Fig. 4 the results obtained for the 2D axisymmetric case are reported for two different mesh dimensions (coarse: 10×10 mm; fine: 5×5 mm). The results are reported both in terms of pressure vs. time histories and in terms of fluids distribution for the case at 0 DOB at 30 cm. Looking at the results it is possible to notice that decreasing the mesh size both the shape of the fluids distribution and the speed of propagation changes with respect to the coarser case.

![Figure 4](image1.jpg)

**Figure 4:** Comparison between numerical and experimental results. The left column shows the numerical results, while the right column shows the experimental results for different mesh sizes and time steps.

In Fig. 5 a qualitative comparison in terms of sand volume fraction is made between numerical results (case 2D axisymmetric - fine) and the images taken by the high speed camera for the three depths of burial at two times (for each experimental test two images are available). In each case, the synchronization is qualitatively performed on the first image and then the second one is taken after a time equal to the time interval between the experimental images. By the comparison, it is possible to conclude that the numerical models are, in general, able to reproduce the phenomenon evolution and the shock-wave propagation.

6 **SQUARE STEEL PLATE**

The numerical model of the landmine is then used in order to evaluate the effects induced by the detonation on a steel square plate, placed at a certain distance above the explosive charge. The results are compared with the experimental data obtained by the Australian Department of Defense [3]. The square plate used for the test was made is AS3678-250 steel.
and had 1200 mm of edge and 5 mm of thickness. The explosive charge used was a sphere of Pentolite with a mass of about 250. Four different tests were performed varying the distance between the charge and the plate between 200 and 500 (two tests at 500 mm, one test at 400 mm and one test at 250 mm). The experimental data were obtained by a LVDT displacement gage, accelerometers and pressure gages placed on the plate. The experimental setup is reported in Fig. 6, in which also the scheme of the numerical model is shown. As first approach the 3D model of a quarter of the system is built, in which the plate is modeled with shell elements. In order to achieve a good accuracy of the results it is needed to increase a lot the number of the elements. For this reason, also if the geometry is not axysimmetric, a 2D model is also analyzed, since it allows refining the mesh.

The models described in the previous paragraph for a DOB=0 cm are adapted to correctly reproduce the geometry of this type of tests. In [3] there is not any specification of the typology of the soil on which the charge is positioned. Anyway this could strongly influence the goodness of the results, since it affects the amount of energy transmitted to the target. As first approximation, the same properties of the previous analysis are used. The strength model and EOS parameters for Pentolite are taken from [8] and those of plate from [3].

In Fig. 7 the comparison between experimental and numerical data in term of pressure vs. time curves is reported for the measuring point indicated as P1 and P2 in Fig. 6. Similarly, also the history plot of the acceleration in correspondence of the measuring points A1 and A2 are reported. As mentioned before, since the explosion can be considered axysimmetric, both 3D and 2D case are analyzed, also if this implies an error in the geometry of the plate and the mine. Different 2D models are built varying the mesh dimension: the coarsest one presents a mesh dimension comparable with the 3D one. The results are shown for the 2D (coarse and fine) and 3D cases solved with the FSI algorithm and for the 3D case solved using the CONWEP algorithm [6]. The case study regards the plate placed at 250 mm above the soil.

As it is possible to notice, considering the same mesh dimension (2D coarsest and 3D, both with FSI and CONWEP), the results are comparable. Otherwise, decreasing the elements dimension, especially for the measuring point 1, greater values both for pressure and acceleration can be obtained. Maybe, in order to better appreciate the comparison with the experimental data it would be necessary to consider an average value obtained in correspondence to the area of the sensor. From the results, it is possible to asses that, also if the CONWEP model is very simple, it produces acceptable results, especially when the
explosion is originated in air, which is the case used for the model calibration. The reliability of this model decreases when the landmine is buried in sand and, therefore, there is the interaction between sand and explosive. Comparing the experimental and numerical results, reported in Fig. 7, it seems that the frequency response of the sensors used during the tests (especially for the accelerometers) is too limited.

In Fig. 8, the time evolution of the fluids distribution is reported for two 2D cases obtained varying the mesh dimension. As mentioned before, increasing the number of elements, both the shape of the clouds of fluids and the speed of propagation change. In fig. 9 there is the comparison in case of 2D models at DOB 0, as before, and DOB 10, in a fully saturated wet sand (2200 kg/m$^3$). The qualitative comparison regards the shape of the deformed plate in the two cases obtained at the same time (2 ms) after the detonation. The quantitative comparison is made in terms of internal and kinetic energies of the plate (per unit length,
circumferentially): if the explosive is buried in a dense soil the impulse is longer and a greater amount of energy is transferred to the plate. The different level of danger between the air-bust and the explosion with debris projection is it well known from a phenomenological point of view and justifies the effort made for the developing of reliable methods for the fluids-structure interaction.

Figure 8: Time evolution of the fluids distribution is reported for two 2D cases obtained varying the mesh dimension (coarse: 10×10 mm and fine 2.5×2.5 mm, plate at 250 mm).

8 HUMAN LEG: THUMS MODEL

The objective of this paragraph is the description of the problem concerning the numerical simulation of the explosion against a complex structure, such as a human body. The USA Defence Department published a report on the number of incidents during the mine clearing operations, from which it appears evident the importance in the protection system for the lower limbs. The model represents a human leg extracted from THUMS (Total HUman Model for Safety). It is a sophisticated FE model developed by Toyota [4] for the prediction of the results in case of the numerical simulation of crush tests, so it is calibrated for the prediction in case of impact events. The complete model represents a sitting average-sized American men. For the evaluation of the consequences of the shock-wave propagation, it is sufficient to take into account only the leg, since the shock amplitude decays very quickly and no effects are produced in the remaining part of the body. The numerical model of the leg is assembled with the numerical model validated for the case of the landmine, adding the FSI algorithm with the fluids and the leg.
Figure 9: Comparison between DOB 0 and DOB 10 (plate at 250 mm): shape of the deformed plate at 2 ms and time history of internal and kinetic energies (per unit length, circumferentially).

Figure 10: Comparison between numerical and experimental results in terms of damage on the foot.
The numerical results are qualitatively compared with the data obtained experimentally on an artificial leg (FSL, Frangible Surrogate Leg) by Bergeron, Coley and Fall in [9].

In Fig. 10, the image taken from an experimental test is compared with the LS-DYNA result, for evaluating the damage of the foot. The test was performed on an unprotected combat boot and the flash x-ray demonstrates that: the calcaneus is pulverized, a high compression acts on the heel, fracture and dislocation of foot phalanges and bones occur, the crack propagates up to the tibia and finally lacerations are provoked. Looking at the numerical results, it is possible to notice that the model is able to reproduce these effects, so it could be used in the military protection development.

9 CONCLUSIONS

This paper described numerical models to simulate landmine explosion and blast loading on structures, using different approaches: Arbitrary Lagrangian Eulerian (ALE) mesh and a pure Lagrangian mesh. For what concerns the ALE simulations, three different cases was analyzed: the numerical model of the landmine explosion varying the DOB and the detonations against two structures (a steel plate and a human leg), using a fluid-structure algorithm. For this type of simulations, an Eulerian approach was needed, in order to reproduce the expansion of the mix of sand, air and gas against the target. When the gas encounters the target a fluid structure interaction algorithm (FSI) determines the pressure values, which are transferred from the Eulerian parts to the Lagrangian ones. The results showed that the numerical model realized are able to reproduce with a good level of accuracy the detonation event and the consequences on different structures. The second approach is based on empirical airblast equations and was applied to simulate the detonation against the plate and the results were compared with the corresponding results obtained using an ALE approach.

REFERENCES

PREDICTION OF SOUND FIELDS IN ACOUSTIC CAVITIES COUPLED TO ABSORPTIVE STRUCTURES DUE TO VIBRATING SURFACES

MARTIN BUCHSCHMID*, GERHARD MÜLLER* AND MARTINA POSPIECH†

*Chair of Structural Mechanics
Technische Universität München
Arcisstrasse 21, 80333 Munich, Germany
e-mail: martin.buchschild@tum.de, web page: http://www.bm.bv.tum.de

†Bertrandt Ingenieurbüro GmbH
Anton-Ditt-Bogen 16, 80939 Munich, Germany
e-mail: martina.pospiech@partner.bmw.de, web page: http://www.bertrandt.com

Key words: Room Acoustics, Fluid Structure Interaction, Component Mode Synthesis, Integral Transform Methods

Abstract. In the scope of this contribution a model for the Fluid Structure Interaction (FSI) is presented, where absorptive structures can be considered. A special focus is laid on the interface coupling modes between absorber and fluid. For the simulation of the spatial resolution of the sound field within acoustic cavities techniques based on Finite Element formulations are used. To reduce the number of degrees of freedom a model reduction method, based on a Component Mode Synthesis (CMS), is applied. The cavity boundary conditions, e.g. compound absorbers made of homogenous plates and porous foams, are modeled using Integral Transform Methods (ITM) and appropriate material formulations. Wavenumber dependent impedances are computed for the absorptive structure and used for the coupling with the acoustic cavity adding interface coupling modes for the fluid and applying Hamilton’s principle.

1 INTRODUCTION

Due to increasing requirements of comfort, acoustic design has become more important during the last years, especially in the field of civil engineering and automotive design. The sound field within rooms or vehicles has to be predicted in the scope of the design process for the specific use.

The calculation of the sound pressure level inside of acoustic cavities is usually done with the help of the Statistical Energy Analysis (SEA). This method is robust for systems...
with a high modal density and it is based on an averaging over frequency bands, points of 
excitation and points of observation. However, its performance is limited if a description 
of the spatial resolution of the response is necessary and if the influence of boundary 
conditions has to be described in detail.

In the scope of the acoustic design of rooms or vehicles elements like reflectors or 
absorbers are placed into the sound field. Therefore a robust method for the phase 
correct modeling of interior sound fields with sufficient spatial resolution is needed for 
the relevant mechanisms of excitation, where the absorptive behavior of the delimiting 
surfaces can be considered. Methods, based on Finite Element formulations are used for 
this purpose.

Finite Element models for absorptive boundary conditions in acoustical calculations 
lead to a huge number of degrees of freedom. In order to reduce this number of unknowns 
an impedance approach considering a wavenumber dependent impedance is used for plate-
like compound absorbers to introduce varying angles of incidence for the sound wave. The 
coupling modes used for the FSI of acoustic fluid and absorber are discussed in detail, 
where a method for the estimation of the number of modes, which have to be considered, 
is presented. Finally a numerical example for the coupled system is presented.

2 FLUID STRUCTURE INTERACTION

The derivation of the FSI method is carried out in the frequency domain. Therefore 
only excitations, harmonically oscillating in time, with the circular frequency of excitation 
Ω are considered. Consequently the steady state solution for both state variables pressure 
p_A and sound velocity v_A is harmonic in time.

2.1 Hamilton’s Principle and Ritz Approach

For the vibro-acoustical problem discussed in this contribution a description of Hamil-
ton’s principle, which is based on velocities, is applied. The structure is divided into 
substructures (see figure 1), where the acoustic fluid and the boundary conditions are de-

fined as subsystems respectively. According to Hamilton’s principle equilibrium is fulfilled 
by the velocity field, which meets the kinematic boundary conditions, the conditions at 
t = t_1 and t = t_2 and, in addition to that, satisfies

\[
\int_{t_1}^{t_2} \delta \left( L_A(t) + L_{BC}(t, Z) + R^T \lambda(t) \right) + \delta W_{BC}^{nc}(t, Z) + \delta W_{\text{Load}}^{nc}(t) \, dt = 0. 
\]

The Lagrangian function \( L_A \) for the acoustic fluid results from the kinetic energy \( T_A \) and 
the potential energy \( U_A \)

\[
L_A(t) = T_A(t) - U_A(t),
\]
where the energies are computed out of

\[ T_A(t) = \frac{\rho_A}{2} \int_V |v_A(x, t)|^2 dV \quad \text{and} \quad U_A(t) = \frac{1}{2\rho_A c^2} \int_V |p_A(x, t)|^2 dV \]

respectively. Harmonically oscillating loads or excitations via surfaces are considered in Hamilton’s Principle as a non-conservative forces by their virtual work \( \delta W_{nc} \)

\[ \delta W_{nc}^{Load}(t) = \int_{A_{Load}} p_{Load}(x, t) n_{Load}(x) \delta w(x, t) dA. \]  \( (3) \)

The formulation of \( L_{BC} \) and \( \delta W_{nc}^{BC} \) will be given for a wavenumber-dependent impedance in the following section. In the scope of a Ritz approach a linear equation system is obtained to compute the unknown coefficients.

### 2.2 Component Mode Synthesis

The Component Mode Synthesis (CMS) is a substructuring technique for large coupled problems, which was introduced by Hurty [5] to reduce the number of unknowns while keeping the physical characteristics of the structure. In contrast to Hurty, the CMS is used based on a modal description in the scope of this method. In order to model arbitrary geometries for the acoustic fluid a numerical approach based on the Spectral Finite Element Method (SFEM) [6, 4] is used. In the frame of the CMS the superscript \( N \) stands for normal modes and the superscript \( C \) for coupling modes. Normal modes are the eigenmodes of the air volume enclosed by totally reflecting boundaries and coupling modes are additionally introduced to provide the coupling to other boundary conditions, like a deformable structure, an absorber or an open interface to another acoustic volume.
For the velocity $v_A$ in the acoustic fluid the approach (4) is applied.

$$v_A(x, t) = \sum_m v_{m}^N(x) \left( A_m e^{i\Omega t} + \overline{A}_m e^{-i\Omega t} \right) + \sum_n v_{n}^C(x) \left( B_n e^{i\Omega t} + \overline{B}_n e^{-i\Omega t} \right)$$

Assuming an acoustic fluid, the irrotational behavior of the sound velocity allows the use of a velocity potential $\Phi_A(x)e^{i\Omega t}$.

$$v_A(x, t) = \text{grad} \Phi_A(x, t)$$

Considering the steady state problem after applying a Fourier transformation from the time- to the frequency-domain, the velocity potential solves the Helmholtz equation (5), where $c_A$ denotes the constant speed of sound.

$$\Delta \hat{\Phi}_A(x, \omega) + \frac{\omega^2}{c_A^2} \hat{\Phi}_A(x, \omega) = 0$$

In the Fourier-transformed domain the velocity $\hat{v}_A$ and the pressure $\hat{p}_A$ read as follows:

$$\hat{v}_A(x, \omega) = \text{grad} \hat{\Phi}_A(x, \omega)$$

$$\hat{p}_A(x, \omega) = -\frac{\rho_A c_A^2}{i\omega} \text{div} \hat{v}_A(x, \omega)$$

The normal modes for the acoustic fluid are defined in terms of the velocity potential $\hat{\Phi}_N$ assuming fixed interfaces, which means reflective wall conditions for all boundaries of the fluid:

$$\text{grad} \hat{\Phi}_m^N(x, \omega_m) \cdot n_{BC} = 0$$

The normal modes are supplemented by coupling modes in order to define a valid set of trial functions for (4). These coupling modes enable velocities perpendicular to the coupling interface and also at the location of a surface-excitation. They fulfill the reflective boundary conditions at all surfaces of the room, except for the interface defined as $x_{BC}$, where modal trial functions $g(x_{BC})$ are prescribed

$$\text{grad} \hat{\Phi}_n^C(x, \Omega) \cdot n_{BC} = g(x_{BC})$$

leading to an inhomogeneous Helmholtz equation for each coupling mode. For the specification of the function $g(x_{BC})$ a multi-index $n = (n_1, n_2)$ is defined with respect to the prescribed vibration pattern. The coupling modes are calculated as solutions of the dynamic problem in a harmonic analysis. Thus the number of coupling modes, which is considered in the calculation, can be chosen with respect to the physical properties of the system for reasons of efficiency.
In order to exemplify the influence of the wavenumber, a rectangular room with reflective walls $[0, L_x = 6 \text{ m}] \times [0, L_y = 3 \text{ m}] \times [0, L_z = 2 \text{ m}]$ is considered and the coupling modes are calculated with the Spectral Finite Element formulation.

Figure 2 shows the velocity potential $\Phi^C$ of the holohedral coupling modes, which cover the whole wall and the subregional coupling modes, when a sinusoidal vibration pattern is prescribed at the interface with a circular frequency of excitation $\Omega = 459 \text{ rad/s}$.

Comparing the potential-fields for different multi-indices $n = (n_1, n_2)$ one observes far-fields for small wavenumbers in the acoustic fluid, whereas with rising wavenumbers near-field solutions are obtained. They are characterized by an exponential decaying behavior perpendicular to the interface. The impact of the subregional modes on the acoustic fluid is smaller than in case of holohedral coupling modes. Also concerning the decay characteristics one observes differences.

These near-field effects can be used to reduce the number of unknowns in the CMS approach. Focusing on the sound field in the cavity, it is sufficient to consider just the far-field coupling modes. To define a sufficient number of coupling modes for an efficient numerical computation, this effect has to be predicted with low numerical effort (a calculation of the coupling modes with the SFEM in advance in order to investigate the decay characteristics would be too expensive). Applying Integral Transform Methods and filtering techniques in the spatial domain one can estimate these decay characteristics with negligible numerical effort. Starting from the wave equation in terms of displacements

$$\frac{\partial^2 u(x, y, z, t)}{\partial x^2} + \frac{\partial^2 u(x, y, z, t)}{\partial y^2} + \frac{\partial^2 u(x, y, z, t)}{\partial z^2} = \frac{1}{c_A^2} \frac{\partial^2 u}{\partial t^2}$$

a Fourier transformation is applied assuming infinite dimensions of the interface. The spatial coordinates $y$ and $z$, defining the plane of the interface, are transformed in the wavenumber domain and a transformation into the frequency domain is carried out, considering vibrations in the steady state with a circular frequency of excitation $\Omega$:

$$\frac{\partial^2 \hat{u}(x, k_y, k_z, \Omega)}{\partial x^2} + \left[\left(\frac{\Omega}{c_A}\right)^2 - k_y^2 - k_z^2\right] \hat{u}(x, k_y, k_z, \Omega) = 0$$

The ordinary differential equation (11) is solved with the exponential approach obtaining the solution for the displacement field of the homogeneous problem:

$$\hat{u}(x, k_y, k_z, \Omega) = \begin{cases} \hat{u}_0(k_y, k_z) \left( \frac{e^{-\lambda L_x}}{e^{\lambda L_x} - e^{-\lambda L_x}} \right) e^{\lambda x} + \left( \frac{e^{\lambda L_x}}{e^{\lambda L_x} - e^{-\lambda L_x}} \right) e^{-\lambda x}, & \lambda \neq 0 \\ \hat{u}_0(k_y, k_z) \left( 1 - \frac{x}{L_x} \right), & \lambda = 0 \end{cases}$$
Figure 2: Holohedral and subregional coupling modes for the velocity potential $\Phi_n^C(x)$ of the rectangular room with reflective walls $[0, L_x = 6 m] \times [0, L_y = 3 m] \times [0, L_z = 2 m]$ for different multi-indices $n = (n_1, n_2)$ with $\Omega = 459 \frac{rad}{s}$. 
The ranges for the far-field, the near-field and the transition zone, where a linear decay is observed, are listed in (13).

\[
k_y^2 + k_z^2 \begin{cases} < \frac{\Omega^2}{c_A^2}, & \text{far-field} \\ > \frac{\Omega^2}{c_A^2}, & \text{near-field} \\ = \frac{\Omega^2}{c_A^2}, & \text{linear decay} \end{cases}
\] (13)

In the \( k_y, k_z \)-domain the transition zone marks a circle with the radius \( \frac{\Omega}{c_A} \). In the practical problem finite absorbers with dimensions \( L_{BC}^x \times L_{BC}^z \) have to be applied. Therefore the infinite vibration pattern \( u_0(y, z) \) is multiplied with a rectangular filter function \( \Theta(y, z) \) in the spatial domain, which is equals a convolution with a 2d sinc-function in the wavenumber domain.

\[
\Theta(y, z) \circ \frac{4}{k_y k_z} \sin \left( \frac{L_{BC}^x}{2} k_y \right) \sin \left( \frac{L_{BC}^z}{2} k_z \right)
\] (14)

Applying a Fourier transformation of the velocity pattern and evaluating the results according to condition (13) the necessary number of coupling modes can be specified. In figure 3 the results are depicted for the holohedral and the subregional coupling modes, where the magenta colored circle specifies the transition zone. The wavenumbers within this circle refer to far-field solutions.

Figure 3 shows, that for the subregional coupling modes the transition to near-fields is linked with smaller multi-index combinations than for the holohedral coupling modes, because a fixed multi-index combination results in higher wavenumbers for the subregional than for the holohedral modes. Also the effect caused by the spatial limitation of infinite vibration patterns is clearly recognized. The number of normal modes, which has to be considered, depends on the frequency of excitation \( \Omega \) and on the load pattern. The number of coupling modes can be limited efficiently selecting the far-fields in the wavenumber domain.

In the following a rectangular geometry is considered for the absorptive boundary condition and \( g(x_{BC}) \) is expressed with \( \psi_n(y, z) \) for each mode, where \( y \) and \( z \) mark the local coordinates in the reference coordinate system of the absorber.

Considering the procedural method in the next steps, in especially the computation of the Lagrangian \( L_{BC} \) for the absorber out of impedances, it is advantageous to express \( \psi_n(y, z) \) with its Fourier Series.

\[
\hat{\psi}_n(y, z) = \sum_r \sum_s E_{nrs} e^{i(r\Delta k_y y + s\Delta k_z z)}
\] (15)
Figure 3: Near-field effects of the holohedral and subregional coupling modes
Thus the trial function at the interface is specified for an absorptive boundary condition as
\[ \hat{v}_{BC}(y, z, t) = \sum_n \hat{\psi}_n(y, z) \left( c_n e^{i\omega t} + \overline{c}_n e^{-i\omega t} \right). \] (16)

Carrying out the integration required in equation (1) over one period of the steady state vibration one obtains the Lagrangian \( L_{BC} \) and the virtual work of the non conservative forces \( \delta W_{BC} \). In this short essay we focus on trial functions, where due to reasons of orthogonality the off diagonal terms vanish.

\[
\int_0^T L_{BC} \, dt = \frac{T}{i\Omega} L_y L_z \left[ \sum_n C_n \overline{C}_n \sum_r \sum_s \text{Im} \left( Z(r, s, \Omega) \right) |E_{nrs}|^2 \right] \] (17)

\[
\int_0^T \delta W_{BC} \, dt = -\frac{T}{i\Omega} L_y L_z \sum_n \left( \overline{C}_n \delta C_n - C_n \delta \overline{C}_n \right) \sum_r \sum_s \text{Re} \left( Z(r, s, \Omega) \right) |E_{nrs}|^2 \] (18)

In case of sinusoidal functions also the Fourier approximation \( \hat{\psi}_n(y, z) \) can be omitted. A detailed discussion as well as the expressions for a general definition of the trial functions are presented in [2]. With the imaginary part of the impedance \( \text{Im} \left( Z(r, s, \Omega) \right) \) the flexible characteristics of the absorber can be modeled, as shown above. The absorptive characteristics are expressed by the real part of impedance \( \text{Re} \left( Z(r, s, \Omega) \right) \).

A description of porous layers based on the Theory of Porous Media [1] as well as the formulation of the system of differential equations and the solution for the fundamental system in the wavenumber-frequency domain for compound absorbers is presented in detail in [2, 3].

3 COUPLING THE SUBSYSTEMS AND ASSEMBLING THE EQUATION SYSTEM

The Normal and coupling modes, which are specified in the CMS approach in equation (4), are computed for the acoustic fluid as trial functions in the scope of a Ritz approach and the Lagrangian of the fluid as well as the virtual work of the external loads are computed with equations (2) and (3) respectively. The Lagrangian of the compound absorber and the virtual work of the non-conservative damping forces are computed with (17) and (18) for instance.

The unknown complex coefficients \( A_i \) and \( B_i \) refer to the normal and the coupling modes in the acoustic volume, whereas \( C_i \) are the coefficients of the trial functions of the compound absorber. The coupling condition of the fluid and the absorber at the interface, which is defined in equation (1) with the help of the vector of Lagrange multipliers \( \lambda \), simply results in \( B_i = C_i \) and \( \overline{B}_i = \overline{C}_i \), if the same velocity pattern is chosen for the trial function of the absorber and for the boundary condition of the fluid at the absorber-interface. Thus the vectors for the unknown coefficients \( x \) and the corresponding conjugate
complex values \( \bar{x} \) read:

\[
x = \begin{bmatrix} A_1 & \cdots & A_{m_{\text{max}}} & B_1 & \cdots & B_{n_{\text{max}}} \end{bmatrix}^T
\]

\( \bar{x} = \begin{bmatrix} \bar{A}_1 & \cdots & \bar{A}_{m_{\text{max}}} & \bar{B}_1 & \cdots & \bar{B}_{n_{\text{max}}} \end{bmatrix}^T \) (19)

The solution of the variational problem is reduced to a problem of minimization because of the Ritz approach. It is advantageous to express the conjugate complex coefficients with real and imaginary values in order to formulate the extremal problem:

\[
x = \begin{bmatrix} x^R + i x^I \end{bmatrix}^T
\]

\( \bar{x} = \begin{bmatrix} x^R - i x^I \end{bmatrix}^T \) (20)

The real and imaginary parts of the complex coefficients mark the new set of unknowns \( y = [x^R \ x^I]^T \), where the total number of real valued unknowns is \( 2(m_{\text{max}} + n_{\text{max}}) \). Here \( m_{\text{max}} \) and \( n_{\text{max}} \) are the maximum numbers of normal and coupling modes respectively.

\[
x^R = \begin{bmatrix} A^R_1 & \cdots & A^R_{m_{\text{max}}} & B^R_1 & \cdots & B^R_{n_{\text{max}}} \end{bmatrix}^T
\]

\[
x^I = \begin{bmatrix} A^I_1 & \cdots & A^I_{m_{\text{max}}} & B^I_1 & \cdots & B^I_{n_{\text{max}}} \end{bmatrix}^T \) (21)

For the consideration of the virtual work a vector \( \delta y \) is specified analogously. Carrying out the minimization of the Lagrangian one obtains a system of real valued linear equations

\[
K y = F, \tag{22}
\]

where the matrix of coefficients \( K_{ij} \) reads

\[
K_{ij} = \frac{\partial^2}{\partial y_i \partial y_j} \int L_A dt + \frac{\partial^2}{\partial y_i \partial y_j} \int L_{BC}(Z) dt + \frac{\partial^2}{\partial y_i \partial y_j} \int \delta W_{\text{nc}}^{BC}(Z) dt \tag{23}
\]

and the load vector \( F = [F^R \ F^I]^T \) considers the external forces:

\[
F_i = -\frac{\partial}{\partial \delta y_i} \int \delta W_{\text{Load}}^\text{nc} dt \tag{24}
\]

Defining submatrices \( K_{rs} \) for the matrix of coefficients, equation (22) reads:

\[
\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} x^R \\ x^I \end{bmatrix} = \begin{bmatrix} F^R \\ F^I \end{bmatrix} \tag{25}
\]

In consequence of the complex property of the unknown coefficients the relations

\[
K_{11} = K_{22} \tag{26}
\]

\[
K_{21} = -K_{12} \tag{27}
\]

hold for the submatrices.
4 NUMERICAL EXAMPLE

For arbitrary geometries the normal and coupling modes can be computed with the SFEM. The only restriction, given by the application of the ITM for the Lagrangian of the absorber, is that the interface has to be plain. In the following a 2d acoustic volume with an inclining rear-wall is considered. The wall containing the interface could be inclined as well. The geometry of the system is sketched in figure 4, where \( L_1^x = 6 \, \text{m} \), \( L_2^x = 1.5 \, \text{m} \) and \( L_y = 2 \, \text{m} \). The model is set up with 192 spectral finite elements. The interface is covered with a 7.2 cm layer of Melamine Foam. A unit point source is located at \( x = 1.15 \, \text{m} \) and \( y = 0.77 \, \text{m} \). The location is chosen under the premise of exciting nearly all modeshapes.

\[
\begin{align*}
Z(k_x, k_y, \Omega) & \quad \text{Load} \\
L_y & \quad d_{TPM} \\
L_1^x & \quad Z(k_x, k_y, \Omega) \\
L_2^x & \quad \text{Load}
\end{align*}
\]

**Figure 4:** 2d structure with with inclined wall and porous absorber (holohedral and subregional coupling)

In figure 5 the steady state response for the sound pressure level is sketched. The different interface-specifications are compared for a frequency of excitation of 163 Hz. Due to the fact, that the frequency of excitation is near to a natural frequency, primarily one specific modeshape is excited, which would lead to very high sound pressures in case of an undamped system. A significant reduction is achieved due to the application of the absorptive layer at the boundary. Comparing both results in figure 5 one observes lower sound pressure levels for the holohedral case, because here the absorptive area, which is introduced into the system by the boundary condition, and therefore the dissipation of energy is higher than for subregional coupling.

5 CONCLUSION

In this contribution a method is presented to compute acoustic cavities under harmonically oscillating loads in order to get phase correct results with a spatial resolution for the sound field using a CMS approach. The normal and constraint modes for the acoustic cavity are calculated with the SFEM, where the number of necessary coupling modes is estimated in the wavenumber domain. The SFEM formulation is implemented in order to model arbitrary geometries for the acoustic cavity. Because of the Fourier transforms in the scope of the ITM however the interface has to be plane. Layered boundary structures as compound absorbers, consisting of homogeneous and porous materials, are modeled efficiently using the Fourier transform, where the number of unknowns can be reduced significantly.
Figure 5: Sound-pressure $p(x, y)$ [Pa] for a frequency of excitation of 163 Hz

REFERENCES


SEIZMIC ANALYSIS OF THE LEŠĆE DAM INCLUDING WATER-DAM-SOIL DYNAMIC INTERACTION

JURE RADNIĆ, ALEN HARAPIN AND MARINA SUNARA*

University of Split, Faculty of Civil Engineering, Architecture and Geodesy
Matice hrvatske 15, 21 000 Split, Croatia
e-mail: marina.sunara@gradst.hr

Key words: Lešće dam, seismic analysis, water – dam – soil dynamic interaction.

Abstract. Structures that are in direct contact with fluid (for example: dams, water tanks, offshore structures, pipelines, water towers, etc.) are often present in engineering practice. Numerical models for real simulations of these structures have to include the simulation of the fluid-structure dynamic interaction. The results of seismic analysis of Lešće dam are presented, which include the water-dam-soil dynamic interaction. The main nonlinear effects of concrete, reinforcement, soil and water, as well as change of the system geometry (large displacements), are modeled. Three different scaled earthquakes and harmonic base excitation, with a maximum acceleration of 0.15g, are considered. Performed analysis confirms sufficient safety of the dam, compared to current regulations and seismic activity of the area.

1 INTRODUCTION

The Lešće dam, as part of namesake hydropower plant, was the first dam built in Croatia (Fig. 1). It is situated on Gojačka Dobra river, a tributary of Kupa, at the end of the canyon near the town of Generalski stol, downstream of Hydropower plant Gojak. The dam is 52.5 meters high and the crest length is 176.4 meters. Its bottom thickness is 35 meters and its top thickness 4 meters. The dam created an accumulation lake length of about 12.61 km and surface area of 146 ha.

Figure 1: Hydropower plant and dam Lešće
The dam was constructed of 12 concrete blocks, with an average width of 14.7 m. Blocks are founded on the solid rock, with different and relatively large founding depth at the bottom of the existing river bed (Fig. 2). The dam was designed [1,2] in accordance with regulations applicable at that time and calculated on the appropriate seismic actions: (i) for the operating period of 200 years, earthquake type "e_1" with maximum acceleration 0.08g, and (ii) for the operating period of 1000 years, earthquake type "e_2" with a maximum acceleration of 0.1g.

This paper discusses only the dam block no.7, 14.5 m wide, with cross section shown in Figure 3. Block is differently buried into the ground and it’s assumed that it has the extreme stresses and strains compared to all others blocks. It is also assumed that, as the most unfavorable condition, the height of the block along its entire length is equal to maximum height of dam, and direct contact of water with the dam base is possible (the maximum possible water pressure at the bottom of the dam). According to the current map of seismic areas of Croatia, HPP Lešće site is located in earthquake area with the maximum expected peak ground acceleration of 0.141g for return period of 475 years [3].

Acting of three real earthquakes on the dam is analyzed: “Ston”, “Banja Luka” and “Petrovac” [4], as well as the harmonic base acceleration. The influence of the simultaneous horizontal and vertical components of base acceleration was included. All excitations are scaled to a maximum acceleration of 0.15g. It should be noted that the adopted computational base acceleration is significantly higher than the one on which the dam was originally calculated. The period of harmonic base accelerations is calculated to correspond to the first period of free oscillations of water-dam-soil coupled system.

2 BASIC CHARACTERISTICS OF THE USED NUMERICAL MODEL

Previously developed numerical model, described in detail in papers [5-9], is used for static and dynamic analysis of Lešće dam. The model can simulate fluid-structure dynamic interaction, as well as all main non-linear effects of concrete and ground rock (yielding in compression, cracks occurrence and propagation in tension, tensile and shear stiffness of cracked materials), reinforcement (yielding in compression and tension) and water
(cavitation). Also, the effect of hydrostatic and hydrodynamic water pressure in structure cracks, as well as geometric non-linearity (large displacements), is simulated.

In displacement formulation for the structure and displacement potential formulation for the fluid, behavior of the fluid-structure system (structure in this context includes the dam and surrounding soil) in dynamic load conditions can be expressed with two second order differential equations:

\[
\begin{align*}
\mathbf{M}_s \ddot{\mathbf{u}} + \mathbf{C}_s \dot{\mathbf{u}} + \mathbf{K}_s \mathbf{u} &= \mathbf{f}_s \\
\mathbf{M}_f \dot{\mathbf{\psi}} + \mathbf{C}_f \mathbf{\psi} + \mathbf{K}_f \mathbf{\psi} &= \mathbf{f}_f + \mathbf{f}_{cf}
\end{align*}
\]

which define dynamic equilibrium of system, where:

\[
f_{cs} = \mathbf{Q} \mathbf{\psi}
\]

\[
f_{cf} = \rho_f \mathbf{Q}^T (\ddot{\mathbf{u}} + \dot{\mathbf{u}})
\]

In the above equations \(\mathbf{M}_s, \mathbf{C}_s\) and \(\mathbf{K}_s\) represent mass, damping and stiffness matrices for structure, whilst \(\mathbf{M}_f, \mathbf{C}_f\) and \(\mathbf{K}_f\) represent same matrices for fluid. Vectors \(\mathbf{u}, \dot{\mathbf{u}}, \ddot{\mathbf{u}}\) represent structure’s displacements, velocities and accelerations, and \(\mathbf{\psi}, \dot{\mathbf{\psi}}, \ddot{\mathbf{\psi}}\) are displacement potential and its associated derivations. \(\mathbf{Q}\) is the interaction matrix between structure and fluid. System of equations (1) in matrix form can be expressed as:

\[
\begin{bmatrix}
\mathbf{M}_s & 0 \\
\rho_f \mathbf{Q}^T & \mathbf{M}_f
\end{bmatrix}
\begin{bmatrix}
\ddot{\mathbf{u}} \\
\dot{\mathbf{\psi}}
\end{bmatrix}
+
\begin{bmatrix}
\mathbf{C}_s & 0 \\
0 & \mathbf{C}_f
\end{bmatrix}
\begin{bmatrix}
\dot{\mathbf{u}} \\
\dot{\mathbf{\psi}}
\end{bmatrix}
+
\begin{bmatrix}
\mathbf{K}_s - \mathbf{Q} \\
0 & \mathbf{K}_f
\end{bmatrix}
\begin{bmatrix}
\mathbf{u} \\
\mathbf{\psi}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{f}_s - \mathbf{M}_f \ddot{\mathbf{u}} \mathbf{d} \\
\mathbf{f}_f - \rho_f \mathbf{Q}^T \dot{\mathbf{u}} \mathbf{d}
\end{bmatrix}
+
\begin{bmatrix}
f_{cs} \\
f_{cf}
\end{bmatrix}
\]

Fluid-structure interaction contact interface with fluid and structure elements is shown in Fig. 4. Interaction matrix \(\mathbf{Q}\) includes only the contact interface integration and is defined as:

\[
(Q)_{ij} = \int_{\Gamma_i} \mathbf{N}_{ui}^T \mathbf{n} \mathbf{N}_{pj} d\Gamma_i
\]

Matrices \(\mathbf{N}_{ui}\) and \(\mathbf{N}_{pj}\) represent shape functions matrices for structure and water (fluid), where \(\mathbf{n}\) represents normal unit vector on interaction contact interface. It should be noted that the second derivation of displacement potential \(\ddot{\mathbf{\psi}}\) represents dynamic water pressure on that position.

Figure 4: Water-structure contact interface and unit normal vector

The system of equations (1) and (3) is solved by so-called partitioned scheme approach. Each field is solved separately, including interaction forces on the interaction contact interface between fluid and structure for every increment of the imposed load and every iteration step. The process is repeated until convergence criterion is obtained.
This approach allows the use of previously developed models for each separate field, with additional calculations of the interaction forces only. Thus, in the fluid-structure interaction model, all non-linear effects of material and geometry, that are present in a particular field, can also be simulated in the coupled problem. Flow chart for the solution of the fluid-structure coupled problem is given in Fig. 5.

Soil-structure interaction is modeled indirectly by contact elements on the contact interface. In fact, by applying the appropriate material model for contact elements, various effects in the contact surface can be simulated, such as: separating, embedment and sliding.

**Figure 5:** Flow chart for the solution of the fluid-structure coupled problem
7 THE RESULTS OF ANALYSIS

As previously mentioned, this paper considers only the dynamic analysis of the dam block no. 7 (Fig. 3). Numerical discretization of the analyzed coupled system is shown in Fig. 6. Adopted values of the basic material parameters are shown in Table 1, and adopted base ground acceleration is shown in Figure 7. Reinforcement is embedded around opening (gallery) in the body of the dam (rods $\Phi 16$ mm on spacing of 25 cm).

Figure 6: Spatial discretization of water-dam-soil system

Figure 7: Adopted base acceleration
The 2D model for spatial discretization is adopted, which is proved for reliable modeling of gravity dams. First, the initial stress-state and displacement state of the system for gravitational loads is solved. Then, a dynamic analysis is performed. Computational accelerations according to Fig. 7 are applied to the model (Fig. 6) as prescribed base accelerations on the bottom of the dam. In the conventional analysis, excitation is applied as acceleration of boundary restrained nodes. Therefore, according to this approach, ground acceleration on model shown on Fig. 7 should be applied to the bottom of the considered part of ground, which will result in significantly more unfavorable state of the dam. The difference in calculated dam’s top horizontal displacement, according to on which level earthquake (“Ston”) is applied, is shown in Fig. 8. It can be concluded that, if “Ston” earthquake acts on the considered part of ground the dam will be collapsed at the end of earthquake.

Some results of the analysis are presented in Fig.9-Fig.20. It is clearly shown that earthquake “Ston” has the most unfavorable effects on the dam. The presented displacements and stresses under dynamic load oscillate around the initial static values.

Horizontal displacement of the dam’s crest is shown in Fig. 9. They are relatively small (due to the high stiffness of the dam) and tend to initial displacement from gravitational loads.

Vertical stresses at the bottom of the upstream side of the dam are shown in Fig. 10. Tensile stresses caused by dynamic excitation do not exceed the tensile strength of concrete, so the cracks do not occur. When dynamic excitation ends, stresses in concrete tend to the initial stress value.

Fig. 11 shows vertical stresses in soil below the dam on the upstream side (contact between dam and foundation rock). Cracks on the joint appear after vertical tensile stresses occurred. When dynamic excitation ends, stresses in the soil tend to initial compressive stress value.

### Table 1. Adopted main material characteristics

<table>
<thead>
<tr>
<th>Material</th>
<th>Modulus of elasticity (E) (MPa)</th>
<th>Shear modulus (G) (MPa)</th>
<th>Poisson’s coefficient (\nu)</th>
<th>Compression strength in (f) (MPa)</th>
<th>Tensile strength in (f) (MPa)</th>
<th>Limit compression strain (\varepsilon)</th>
<th>Limit tension strain (\varepsilon)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dam</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>32 000</td>
<td>13 333</td>
<td>0.2</td>
<td>30</td>
<td>3.0</td>
<td>-0.0035</td>
<td>0.00198</td>
</tr>
<tr>
<td><strong>Reinforcement</strong></td>
<td>210 000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Es</td>
<td>210 000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fs</td>
<td>560</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Soil (basal rock)</strong></td>
<td>25 000</td>
<td>10 416</td>
<td>0.2</td>
<td>3.0</td>
<td>0.3</td>
<td>-0.0035</td>
<td>0.00024</td>
</tr>
<tr>
<td>Et</td>
<td>25 000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gt</td>
<td>10 416</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Water</strong></td>
<td>2045</td>
<td>1 000</td>
<td>1.430</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: The dam’s top horizontal displacement for a “Ston” earthquake in function of earthquake level application.
Vertical stresses at the bottom of the downstream side of the dam are shown in Fig. 12, and vertical stresses in soil below the dam on the downstream side in Fig. 13. It is obvious that vertical stresses are always compressive and trend to initial compressive stress value after the excitation ends.
Vertical stresses at the bottom of the downstream side of the dam (MPa)

Figure 12: Vertical stresses in concrete at the bottom of the downstream side of the dam

Vertical stresses in soil below the dam on the downstream side (MPa)

Figure 13: Vertical stresses in soil below the dam on the downstream side

Fig. 14 shows the stresses in reinforcement on the bottom of the gallery, which are relatively low. The maximal principal stresses ($\sigma_{11}, \sigma_{22}$) in the dam, registered in time steps of maximal displacement, are shown on Fig. 15. It can be noted that they are also relatively low.

Figure 14: Stresses in reinforcement on the bottom of the gallery
Figure 15: Principal stresses $\sigma_{11}$ and $\sigma_{22}$ in dam

Deformations of the dam at some time steps for “Ston” earthquake, with the cracks on the surface of the joint between dam and the base rock, are shown in Figure 16 (red line indicates the length of the open horizontal crack).
Fig. 17 shows the hydrodynamic forces on the dam. It is evident that its value is about 25% of hydrostatic force for the harmonic base acceleration. Position of hydrodynamic force in relation to the dam height is shown in Fig. 18. It oscillates around a point which is positioned about 40% of the dam height, which is close to the result for the rigid structures [5].

Hydrodynamic pressures at the bottom of the dam are shown in Fig. 19. The maximal hydrodynamic pressure is about 25% of hydrostatic pressure. Cavitation in water is not registered under the considered excitations.
4 CONCLUSIONS

Previously developed, tested and reliable numerical model for static and dynamic analysis of water-soil-structure coupled system [5-9] was used for the analysis of Lešće dam. The model simulates all main nonlinear effects of water, concrete, reinforcement and soil. The dam block no. 7, which is probably the least favorable of all blocks of the dam, was analyzed. The analysis determines sufficient security and stability of the block, even for possible earthquake actions with a maximum acceleration of 0.15 g. Of all possible non-linearity, only the elevation of the dam bottom of the foundation rock is present. It is expected that other dam blocks also have sufficient security and stability for the same maximum base acceleration. With respect to the all considered dynamic excitations, earthquake “Ston” showed the most unfavorable effects.

REFERENCES

ADAPTIVE FINITE ELEMENT LIMIT ANALYSIS INCORPORATING STEADY STATE PORE PRESSURES

SCOTT W. SLOAN, ANDREI V. LYAMIN AND KRISTIAN KRABBENHØFT

ARC Centre of Excellence for Geotechnical Science and Engineering (CGSE)
The University of Newcastle
NSW 2308, Callaghan, Australia
e-mail: scott.sloan@newcastle.edu.au
web page: http://www.cgse.edu.au

Key words: Finite Element, Limit Analysis, Pore Pressures, Coupled Problems.

Abstract. This paper describes an adaptive method for incorporating the pore pressures associated with steady-state seepage in finite element limit analysis. The formulation can model both general seepage conditions, and locating the phreatic surface in unconfined flow presents no special difficulty [1]. Since the proposed method employs the same mesh for the upper analysis, the lower-bound analysis and the pore pressure computation, there is no need to import and interpolate the pore-pressures from another grid (or program). This is a significant practical advantage and leads to an efficient solution process.

In the iterative solution method, a Hessian(curvature)-based error estimator is applied to the pore pressure field to generate a mesh which gives accurate pore pressures. Simultaneously, a variant of the ‘bounds gap’ error estimator of [2],[3] is employed to identify a separate mesh which gives accurate upper and lower bounds on the limit load. By combining these two strategies, a hybrid refinement strategy is developed which minimises both the bounds gap and the error in the computed pore pressures.

1 INTRODUCTION

Finite element limit analysis has become an important tool for predicting the load capacity of geostructures such as foundations, retaining walls, tunnels and slopes. These methods are based on discrete formulations of the limit theorems [4],[5] and lead to large sparse optimisation problems which can be solved very efficiently. In this paper, the pore pressures are incorporated using the upper bound procedures described in [6],[7], the lower bound procedure described in [8], and the algorithms described in [1]. After briefly describing the discrete bound formulations, we outline a variant of the scheme proposed in [2],[3] which measures the contribution of each element to the ‘bounds gap’. This ‘bounds gap’ estimator relies on using an identical discretization for the upper and lower bound analyses, and provides a very effective means for refining the mesh to give accurate solutions. Next, we describe a procedure for computing steady-state pore pressures. This method uses the same mesh as the bound procedures, and permits the effect of pore pressures on the limit load to be modelled accurately. Finally, the formulation is used to study the stability of a slope with a weak layer.
2 UPPER BOUND FORMULATION

The upper bound formulation follows that described in [6], with the enhancements described in [7] to model velocity discontinuities. In an upper bound calculation we search for a velocity distribution \( \mathbf{u} = \{u_x, u_y\}^T \) which satisfies compatibility, the flow rule, the velocity boundary conditions \( \mathbf{w} \) on the surface area \( A_w \), and minimises the internal power dissipation less the rate of work done by the fixed external loads. The latter quantity can be written as

\[
\dot{W} = \int \sigma^T \dot{\varepsilon}^p dV - \int A_t \mathbf{t} \cdot dA - \int g^T \mathbf{u} dV
\]  

(1)

where \( \sigma = \{\sigma_{xx}, \sigma_{yy}, \tau_{xy}\}^T \) is a vector of stresses, \( \dot{\varepsilon}^p = \{\dot{\varepsilon}_{xx}^p, \dot{\varepsilon}_{yy}^p, \gamma_{xy}^p\}^T \) is a vector of plastic strain rates, \( \mathbf{t} \) are fixed surface tractions acting on the area \( A_t \), \( g \) are fixed body forces, and \( V \) is the volume of the body. The first integral on the right hand side of (1) corresponds to the plastic dissipation \( \int P \). Once the velocity field which minimizes \( \dot{W} \) has been found, the optimized value of \( \dot{W} \) can be equated to the power expended by the external loads

\[
P_{ext} = \int A_t \mathbf{q} \cdot \mathbf{u} dA + \int h \mathbf{u} dV
\]  

(2)

to find an upper bound on the unknown surface tractions \( \mathbf{q} \) (acting over the area \( A_t \)) or the unknown body forces \( h \) (acting over the volume \( V \)). Note that the latter are included to permit optimisation of body forces, such as the unit weight, which is very useful in modelling slopes, excavations, and tunnels. The 2D continuum element used in the upper bound formulation is shown in Fig. 1. Over each triangle, the unknown velocities \( \mathbf{u} \) are assumed to vary linearly while the unknown element stresses \( \sigma^e \) are assumed to be constant. Although it is possible to develop upper bound formulations without the stresses \( \sigma^e \), they are included because they permit general types of yield surfaces to be modelled. Moreover, they also enable a simple error estimate to be formulated which measures the ‘bounds gap’ contribution of each element directly.

Following the approach proposed in [7], kinematically admissible velocity discontinuities are modelled by using a patch of zero-thickness continuum upper bound elements, as shown in Fig. 2. The elements \( D_1 \) and \( D_2 \) are standard upper bound elements, except that two of the nodal co-ordinates are set to be identical, and the flow rule equations are multiplied by the
element areas. This strategy ensures the velocity jumps across the discontinuity are kinematically admissible [7], leads to an efficient formulation in 2D and 3D, and does not require special algorithms when computing the ‘bounds gap’ error estimator [1]. The discrete formulation of the upper bound theorem leads to a large, sparse, nonlinear optimisation problem where $W$ in equation (1) is the quantity that is minimised over the field of unknowns $(\boldsymbol{\sigma}, \mathbf{u})$. After assembling the objective function coefficients and constraints for a 2D mesh, this optimisation problem can be expressed as

\[
\begin{align*}
\text{Minimise} & \quad \sigma^T \mathbf{B} \mathbf{u} - \mathbf{c}^T \mathbf{u} & \text{power dissipation - rate of work done by fixed external forces} \\
\text{Subject to} & \quad \mathbf{B}^e \mathbf{u}^e = \alpha^e \nabla f(\mathbf{\sigma}^e) & \text{flow rule conditions for each element } e \\
& \quad \alpha^e \geq 0 & \text{plastic multiplier times for each element } e \\
& \quad \alpha^e f(\mathbf{\sigma}^e) = 0 & \text{consistency condition for each element } e \\
& \quad \mathbf{A} \mathbf{u} = \mathbf{b} & \text{velocity boundary conditions, load constraints} \\
& \quad f(\mathbf{\sigma}^e) \leq 0 & \text{yield condition for each element } e
\end{align*}
\]

where $\boldsymbol{\sigma}$ is a global vector of unknown element stresses, $\mathbf{u}$ is a global vector of unknown nodal velocities, $\mathbf{B}^e$ is the element compatibility matrix multiplied by the element area, $\mathbf{B} = \sum \mathbf{B}^e$ is a global compatibility matrix, $\sigma^T \mathbf{B} \mathbf{u}$ is the power dissipated by plastic shearing in the continuum and discontinuities, $\mathbf{c}^T \mathbf{u}$ is the rate of work done by fixed tractions and body forces, $\alpha^e$ is the plastic multiplier times the area for element $e$, $f(\mathbf{\sigma}^e)$ is the yield function for element $e$, $\mathbf{A}$ is a matrix of equality constraint coefficients, and $\mathbf{b}$ is a known vector of coefficients. Rather than solve the primal problem (3) directly, it is convenient to consider its dual as discussed in [7]. For 2D Tresca and Mohr-Coulomb yield criteria, which are widely used in geomechanics, the dual optimisation problem can be solved very efficiently using second order cone programming.

3 LOWER BOUND FORMULATION

The discrete lower bound formulation [8] seeks to find a statically admissible stress field $\boldsymbol{\sigma} = \{\sigma_{xx}, \sigma_{yy}, \tau_{xy}\}^T$ which satisfies equilibrium throughout $V$, balances the prescribed tractions $\mathbf{t}$ on $A$, nowhere violates the yield criterion $f$ so that $f(\mathbf{\sigma}) \leq 0$, and maximises the collapse load

\[
Q = \int_{A} Q_1(\mathbf{q}) \, dA + \int_{V} Q_2(\mathbf{h}) \, dV
\]

where the functions $Q_1$ and $Q_2$ depend on the case at hand. In a footing bearing capacity problem, for example, $Q_2 = 0$ and we wish to maximise the load carried by the tractions normal to a boundary edge, $q_n$, so that $Q_1 = q_n$. In contrast, for a slope or excavation stability we frequently wish to maximise a dimensionless stability parameter which is a function of the unit weight $\gamma$ so that $Q_1 = 0$ and $Q_2 = \gamma$. The 2D continuum element used in the lower bound formulation is shown in Fig. 3. Over each triangle, the unknown stresses $\boldsymbol{\sigma}$ are assumed to vary linearly while the unknown element body forces $\mathbf{h}^e$ are assumed to be constant.
Following the approach proposed in [9], stress discontinuities are modelled by using a patch of zero-thickness continuum lower bound elements, as shown in Fig. 4. The elements $D_1$ and $D_2$ are standard lower bound elements, except that pairs of the nodal co-ordinates are set to be identical (as in the modelling of the velocity discontinuities described in the previous section). To ensure that the normal and shear stresses are continuous everywhere along the discontinuity, so that the solution is statically admissible, the standard equilibrium equations are multiplied by the element areas. This type of formulation is again efficient in both 2D and 3D and is particularly convenient when computing the ‘bounds gap’ error estimator [1].

![Figure 3: Lower bound element](image1)

![Figure 4: Stress discontinuity](image2)

After assembling the objective function coefficients and constraints for a mesh, the load carried by the unknown stresses and body forces, denoted by $T_1^c \sigma$ and $T_2^c h$ respectively, can be maximised by solving the following nonlinear optimisation problem

$$\text{maximise } \begin{vmatrix} c_1^\top \sigma + c_2^\top h \\ -f(\sigma') \leq 0 \end{vmatrix}$$

subject to

$$A_{11} \sigma + A_{12} h = b_1$$

continuum equilibrium

$$A_{21} \sigma = b_2$$

discontinuity equilibrium, stress boundary conditions

$$f(\sigma') \leq 0$$

yield conditions for each node $i$

where $c_1$, $c_2$, $b_1$, and $b_2$ are vectors of constants, $A_{11}, A_{12}$ and $A_{21}$ are matrices of constants, $f$ is the nonlinear yield criterion, $\sigma'$ is a local vector of Cartesian stresses at node $i$, $\sigma$ is a global vector of unknown Cartesian stresses, and $h$ is a global vector of unknown body forces acting on each element. As discussed in [10],[11], second order cone programming algorithms have proved to be very efficient for solving (5) when 2D Tresca and Mohr-Coulomb criteria are used.

4 ADAPTIVE MESH REFINEMENT USING BOUNDS GAP MEASURE

To develop an exact measure of the contribution of each element to the bounds gap, we consider the case where identical meshes are used for the upper and lower bound analyses. From the principle of virtual equilibrium, the total plastic dissipation for an upper bound analysis can be written as
\[ \int_{V} \mathbf{\sigma}_{UB}^{T} \mathbf{\dot{e}}^{p} \, dV = \int_{A} \mathbf{q}_{UB}^{T} \mathbf{u} \, dA + \int_{V} \mathbf{h}_{UB}^{T} \mathbf{u} \, dV + \int_{A} \mathbf{t}^{T} \mathbf{u} \, dA + \int_{V} \mathbf{g}^{T} \mathbf{u} \, dV \]  

(6)

where the subscript \( UB \) denotes upper bounds for the unknown stresses, surface tractions and body forces. Since the velocities \( \mathbf{u} \) and plastic strain rates \( \mathbf{\dot{e}}^{p} \) are kinematically admissible throughout the domain, the principle of virtual equilibrium can be invoked again to give

\[ \int_{V} \mathbf{\sigma}_{LB}^{T} \mathbf{\dot{e}}^{p} \, dV = \int_{A} \mathbf{q}_{LB}^{T} \mathbf{u} \, dA + \int_{V} \mathbf{h}_{LB}^{T} \mathbf{u} \, dV + \int_{A} \mathbf{t}^{T} \mathbf{u} \, dA + \int_{V} \mathbf{g}^{T} \mathbf{u} \, dV \]  

(7)

where \( \mathbf{\sigma}_{LB}, \mathbf{q}_{LB} \) and \( \mathbf{h}_{LB} \) are the statically admissible lower bound stresses, tractions, and body forces respectively. Subtracting (7) from (6), the ‘dissipation gap’ for the mesh, \( \Delta \), can be expressed as

\[ \Delta = \int_{V} (\mathbf{\sigma}_{UB} - \mathbf{\sigma}_{LB})^{T} \mathbf{\dot{e}}^{p} \, dV = \int_{A} (\mathbf{q}_{UB} - \mathbf{q}_{LB})^{T} \mathbf{u} \, dA + \int_{V} (\mathbf{h}_{UB} - \mathbf{h}_{LB})^{T} \mathbf{u} \, dV \]

Assuming proportional loading, with the upper and lower bound load multipliers \( (\lambda_{UB}, \lambda_{LB}) \) defined so that \( \mathbf{q}_{UB} = \lambda_{UB}^{q} \mathbf{q}, \mathbf{h}_{UB} = \lambda_{UB}^{h} \mathbf{h} \) and \( \mathbf{h}_{LB} = \lambda_{LB}^{h} \mathbf{h} \), the equation above becomes

\[ \Delta = \int_{V} (\mathbf{\sigma}_{UB} - \mathbf{\sigma}_{LB})^{T} \mathbf{\dot{e}}^{p} \, dV = (\lambda_{UB}^{q} - \lambda_{LB}^{q}) \int_{A} \mathbf{q}^{T} \mathbf{u} \, dA + (\lambda_{UB}^{h} - \lambda_{LB}^{h}) \int_{V} \mathbf{h}^{T} \mathbf{u} \, dV \]  

(8)

In practical calculations, it is usual to optimise either the tractions or the body forces. For cases where the tractions and body forces are optimised simultaneously, the multipliers must be linked (typically through a relation of the form \( \lambda_{UB}^{q} = \beta \lambda_{LB}^{h} \), with \( \beta \) being a prescribed constant). Equation (8) shows that the dissipation gap gives an exact measure of the gap between the upper and lower bounds for a mesh. The contribution of each element to the ‘bounds gap’, \( \Delta^{e} \), can thus be computed using the expression

\[ \Delta^{e} = \int_{V} (\mathbf{\sigma}_{UB}^{e} - \mathbf{\sigma}_{LB}^{e})^{T} (\mathbf{\dot{e}}^{p})^{e} \, dV = (\mathbf{\sigma}_{UB}^{e} - \mathbf{\sigma}_{LB}^{e})^{T} \mathbf{B}^{e} \mathbf{u}^{e} \]  

(9)

where \( \mathbf{B}^{e} \) is the standard compatibility matrix times the element volume and

\[ \Delta = \sum_{\text{elements}} \Delta^{e} \]

Since the quantities defined by equation (9) are always positive [12], their values can be used to govern an adaptive mesh refinement process, with each new element size being chosen to be inversely proportional to the magnitude of \( \Delta^{e} \). For each discontinuity element, the bounds gap contribution can be added to the neighbouring continuum element with which it shares the most nodes. A summary of an efficient algorithm for refining the mesh, based on the bounds gap measure \( \Delta^{e} \), can be found in [1].

5 COMPUTING STEADY STATE PORE PRESSURES DUE TO SEEPAGE FLOW

At a point with an elevation head \( z \), the steady-state pore pressure \( p \) is related to the total head \( H \) by the equation \( p = (H - z)\gamma_{w} \), where \( \gamma_{w} \) is the unit weight of water. To compute the pore
pressure field, we thus need to solve for the total head which is governed by the well-known equation

\[ k(x, y) \nabla^2 H = k(x, y) \frac{\partial^2 H}{\partial x^2} + k(x, y) \frac{\partial^2 H}{\partial y^2} = 0 \]  

(10)

where \( k(x, y) \) is the soil permeability. Using standard variational methods, it can be shown that the solution to (10) is defined by the optimisation problem

\[
\text{minimize } \frac{1}{2} \int_V (\nabla H)^T k \nabla H \, dV
\]

subject to appropriate boundary conditions on \( H \). Assuming a linear variation of \( H \) over a 3-noded triangular element, the discrete form of this optimisation problem may be written as

\[
\text{minimise } \frac{1}{2} H^T K H
\]

subject to \( A H = H_0 \)  

(12)

where \( H \) is a global vector of unknown nodal heads, \( H_0 \) is a vector of prescribed values, \( A \) is a matrix specifying the constant head boundary conditions, and \( K \) is a flow matrix defined by

\[
K = \sum_{e} \int_{A^e} \nabla N_e^T k(x, y) \nabla N_e \, dA
\]

(13)

In the equation above, \( E \) denotes the number of triangular elements and \( \nabla N_e \) is the gradient of the shape function matrix for element \( e \). The solution to the quadratic programming problem (12) is defined by the equations

\[
KH = 0
\]

(14)

with the boundary conditions \( AH = H_0 \). To avoid the need to transfer the pore pressures from one grid to another, it is convenient to employ the same mesh for the pore pressure and limit analysis calculations. This means, however, that the matrix \( A \) must also contain appropriate terms to ensure that the pore pressure field is continuous across the discontinuities between adjacent triangles \[1\]. Once (14) has been solved for the nodal heads \( H \), the pore pressures at each node can be found from the relation \( p = (H - z) \gamma_w \) where \( z \) is a vector of the nodal elevation heads.

If the seepage flow is unconfined, then the phreatic surface must be located and the quadratic optimisation problem (12) is subject to the additional constraint that \( p = (H - z) \gamma_w \geq 0 \). This supplementary condition can be used to compute the pore pressures using the following algorithm described in \[1\]:

1. Solve (14) to give the nodal heads \( H \).

2. Find the pore pressures at each node from \( p = (H - z) \gamma_w \).
3. If the change in $H^T K H \leq TOL$, where TOL is a suitably small tolerance, exit with the
solution for the pore pressures.

4. For all nodes $i$ where the pore pressure $p_i < 0$, adjust the nodal permeability using the
relation $k_i = s(p) k_i$, where $s(p) = 0.5(1 + \tanh(50p))$ is a smoothed step function which
lies between 0 and 1.

5. Using (13), recompute $K$ for each element with the adjusted nodal permeabilities $k_i$,
then go to step 1.

The smoothed step function in step 4 reduces pore pressure oscillations in the vicinity of the
phreatic surface, and the overall process typically requires less than 10 iterations with a
stringent convergence tolerance. When computing the contributions to $K$ in step 5, the
permeability is assumed to vary linearly over each element. This gives a ‘weighted’ permeability for elements that are bisected by the phreatic surface, and further enhances the
rate of convergence of the iteration scheme.

6 **UPPER BOUND FORMULATION WITH PORE PRESSURE**

To include pore pressure in the upper bound formulation, the effective stress (i.e. the total
stress minus the pore pressure) must be used in the yield condition and flow rule equations.
The pore pressure field is treated as an auxiliary variable, which varies linearly over each
element but is fixed during each limit analysis calculation. Referring to Fig. 1, the upper
bound nodal variables are thus $u' = \{u'_x, u'_y, p\}^T$ while the element effective stresses are
$\sigma' = \{\sigma'_{xx} - \bar{p}, \sigma'_{yy} - \bar{p}, \sigma'_{xy}\}^T$ where $\bar{p}$ is the pore pressure at the triangle centroid. In addition
to these changes, the pore pressures do extra external work so that equation (1) becomes

$$W' = \int V \sigma'T \varepsilon dV - \int A \tau^T u dA - \int V g^T u dV - \int V \nabla p^T u dV$$

(15)

where $\nabla p = \{\partial p / \partial x, \partial p / \partial y\}^T$ is the gradient of the pore pressure field. Since the pore
pressure is assumed to vary linearly, this gradient is constant and the last term in equation
(15) is simple to evaluate.

7 **LOWER BOUND FORMULATION WITH PORE PRESSURE**

To include pore water pressure in the lower bound formulation, effective stresses are used in
the yield constraints, while total stresses are employed when imposing the equilibrium and
stress boundary conditions. No other changes are needed.

8 **ADAPTIVE MESH REFINEMENT WITH PORE PRESSURES**

To minimise the error in both the pore pressures and the bound solutions, a hybrid mesh
refinement strategy is employed [1]. This first uses a Hessian (curvature)-based error
estimator [13] to predict good element sizes for the pore pressures, and then applies the
bounds gap scheme (as described in Section 4) to predict good element sizes for the discrete
limit analysis calculations. Where the element sizes from these two separate methods differ,
the hybrid strategy simply chooses the smallest one. A Hessian-based scheme for selecting element sizes in lower bound limit analysis can be found in [14]. Exactly the same approach is adopted here, with the ‘isotropic’ form of the method being implemented which omits element stretching.

9 STABILITY OF SLOPE IN A WEAK LAYER

To illustrate the performance of the discrete limit analysis formulations with pore pressures and adaptive mesh refinement, we consider an example taken from the benchmark prediction exercise documented in [15]. The problem, defined in Fig. 5, has a non-circular failure surface which propagates along the weak zone, and is a challenging test for conventional slope stability methods as well as finite element limit analysis. In order to compare the new stability solutions with those reported in [15], the following strength reduction process was used to estimate the safety factor [1]:

1. Start with a trial safety factor $F_0$
2. Compute the available strengths $c'_a = c'/F_0$ and $\phi'_a = \tan^{-1}(\tan \phi'/F_0)$.
3. Using the available strengths $(c'_a, \phi'_a)$ and the adaptive discrete limit analysis algorithm with the bounds gap error indicator, compute upper and lower bounds on the unit weight that can be supported by the slope $(\gamma_{LB}, \gamma_{UB})$. Then compute the mean of the these bounds according to $\gamma = (\gamma_{UB} + \gamma_{LB})/2$ and the multiplier $m_0 = \gamma / \gamma$, where $\gamma$ is the actual unit weight.
4. If $m_0 < 1$ set $\Delta F = -\delta$, else set $\Delta F = \delta$.
5. Compute $F_1 = F_0 + \Delta F$
6. Update the available strengths according to $c'_a = c'/F_1$ and $\phi'_a = \tan^{-1}(\tan \phi'/F_1)$.
7. Using the available strengths $(c'_a, \phi'_a)$, find new bounds on the unit weight $(\gamma_{LB}, \gamma_{UB})$. Then compute $\gamma = (\gamma_{UB} + \gamma_{LB})/2$ and the multiplier $m_1 = \gamma / \gamma$.
8. If $(m_1 - 1)(m_0 - 1) > 0$, then set $m_0 = m_1$ and $F_0 = F_1$ and go to 5.
9. Linearly interpolate the factor of safety using $F = F_0 + (F_1 - F_0)(m_0 - 1)/(m_0 - m_1)$.

In step 4, $\delta \in [0.01, 0.1]$ is a constant that defines the increment in the search for the critical safety factor. The strength reduction process starts by assuming a trial estimate of the safety factor and continues with a simple marching scheme until a value is found which gives a gravity multiplier on the unit weight, $m$, of unity. Instead of taking the average of the bounds on $\gamma$ to compute $m$ in steps 3 and 7, it is possible to use the actual lower or upper bounds, and hence compute an upper or lower bound on the safety factor $F$. This feature may be useful for difficult problems, but it is often unnecessary due to the very tight bounds that are generated by the discrete limit analysis methods.
Fig 6 shows the progress of the strength reduction process described above. Starting with an initial estimate of the safety factor of $F_0=0.80$ and employing a step size of $\delta=0.01$, the algorithm gives the safety factor of $F=0.822$ after four iterations. This analysis, which used a maximum of 4,000 elements, required a total of about 30 seconds of CPU time on a standard desktop machine. In the vicinity of the final solution, the upper and lower bounds bracket the computed safety factor to within $\pm 2.9\%$. If desired, more accurate solutions could be obtained by using more elements.

Fig. 7 shows the optimised mesh at the completion of the strength reduction process. This indicates that the bounds gap error measure has concentrated elements along the failure surface while, simultaneously, the Hessian-based error estimator has clustered elements around the phreatic surface. The corresponding plots of the plastic multipliers (strains) and...
velocity vectors, shown in Figs 8 and 9 respectively, confirm that the mode of failure is dominated by intense shear deformation in the weak layer of cohesionless material. In addition, wedge-type mechanisms occur at the toe of the slope and behind its crest.

Figure 7: Optimised mesh

Figure 8: Plastic multiplier (strain) contours at collapse

Figure 9: Velocity vectors at collapse

Fig. 10 compares the factors of safety computed from discrete limit analysis and a variety of conventional limit equilibrium methods. The latter, taken from [15], indicate a significant scatter in the estimated safety factor, even for analyses with the same technique. These variations reflect the fact that locating the critical limit-equilibrium failure surface is a
difficult unconstrained optimisation problem. Indeed, highly sophisticated search strategies are needed to obtain a reliable solution, especially if the failure surface is permitted to be non-circular. The solutions from the discrete limit analysis method, on the other hand, are guaranteed to give the best possible bounds for a specified mesh, since the governing constrained optimisation problem is convex (provided the yield surface is convex). Different limit equilibrium procedures make different assumptions in order to obtain a solution, some of which are physically more justified than others [16]. Although the Morgenstern-Price [17], Spencer [18] and Sarma [19] procedures satisfy both force and moment equilibrium for each slice, and may be viewed as ‘complete equilibrium’ techniques, they also give results with considerable scatter. The simplified Bishop method over-estimates the safety factor since it is better suited to cases where the failure surface can be approximated by a circle. In the study described in [15], the best estimate of the safety factor was taken to be \( F = 0.78 \), which is 5% below the limit analysis predictions.

**Figure 9**: Comparison of safety factors

10 **CONCLUSIONS**

An adaptive formulation for incorporating pore pressures in discrete limit analysis has been described. The method can model the effects of both confined and unconfined seepage flow, and is a powerful new tool for practical stability analysis in geotechnical engineering.

11 **REFERENCES**


A PFEM APPROACH TO THE SIMULATION OF LANDSLIDE GENERATED WATER-WAVES

MASSIMILIANO CREMONESI*, CLAUDIO DI PRISCO* AND UMBERTO PEREGO*

*Department of Civil and Environmental Engineering
Politecnico di Milano
Piazza Leonardo Da Vinci, 32, 20133 Milan, Italy
e-mail:cremonesi@stru.polimi.it, claudio.diprisco@polimi.it, umberto.perego@polimi.it

Key words: landslide simulation, PFEM, Lagrangian approach

Abstract. A Particle Finite Element Method is here applied to the simulation of landslide-water interaction. An elastic-visco-plastic non-Newtonian, Bingham-like constitutive model has been used to describe the landslide material. Two examples are shown to show the potential of the approach.

1 INTRODUCTION

Catastrophic landslides impinging into water reservoirs may generate impulsive waves whose propagation can cause considerable damages. This is an exceptional natural hazard, usually associated with erosion, fault movements, earthquakes, heavy rainfalls or storms. The prediction of landslides velocity, runout distance and travelling path is useful for preventing and mitigating the consequences of these events. Recent developments in the simulation techniques for coupled problems have led to efficient analysis procedures allowing for the accurate reproduction of landslide-reservoir interactions (see for example [1, 2]). The numerical analysis of these events requires capabilities for tracking interfaces and free surfaces undergoing large displacements, and accounting for the mixing of different constituents, for complex constitutive behaviours and for multi-physics processes. A recently developed Lagrangian finite element approach formulated in the spirit of the Particle Finite Element Method [3, 4, 5] is here reconsidered and adapted to the specific case of landslide-reservoir interaction.

Owing to its capability of automatically tracking free-surfaces and interfaces, the proposed method is particularly suitable for the simulation of landslide-water interaction problems, which are dominated by fast propagating waves and interfaces.
2 NUMERICAL TECHNIQUE

The Particle Finite Element Method (PFEM) was originally developed [5, 6, 7] for solving problems involving free surfaces fluid flows and fluid-structure interaction. The method is here revisited and applied to the simulation of landslides, their interaction with a basin and the generation and propagation of water waves.

Both landslide and water motions are governed by Lagrangian Navier-Stokes equations:

\[ \rho_0 \frac{Du}{Dt} = \text{Div}\Pi + \rho_0 \mathbf{b} \quad \text{in } \Omega_0 \times (0, T) \quad (1) \]
\[ \text{Div} (JF^{-1}\mathbf{u}) = 0 \quad \text{in } \Omega_0 \times (0, T) \quad (2) \]

In the first equation, expressing momentum balance, \( \rho_0 \) is the density of the fluid, \( \mathbf{u} \) is the velocity, \( \Pi = J\sigma F^{-T} \) is the first Piola-Kirchhoff stress tensor, \( \sigma \) is the Cauchy stress tensor, \( F \) is the deformation gradient, \( J \) is the determinant of \( F \) and \( \Omega_0 \) represents the initial (and reference) configuration. In the second equation, expressing mass conservation in view of the assumed incompressibility, \( \mathbf{u} \) is the velocity vector.

A classical Finite Element procedure is used to discretize the problem in space while a backward Euler scheme is employed for the time integration. In the spirit of the Particle Finite Element Method, to avoid excessive mesh distortion due to the Lagrangian nature of the equations, the domain is frequently remeshed. An index of the element distortion is used to check whether the mesh should be regenerated or not. When a new mesh is to be created a Delaunay triangulation technique is used to redefine the nodal connectivity starting from the current node position. Moreover, an "alpha shape" technique is introduced to identify the free-surfaces and the interacting surfaces between water and landslide. Details on the numerical procedure can be found in [2, 3, 4, 5].

3 CONSTITUTIVE LAW

Both the landslide and the reservoir water have been modelled as viscous fluids. The Cauchy stress tensor \( \sigma = \sigma(x, t) \) is decomposed into its hydrostatic \( p \) and deviatoric \( \tau \) components as \( \sigma = -p\mathbf{I} + \tau \) where \( \mathbf{I} \) the identity tensor.

Water is assumed to be a Newtonian isotropic incompressible fluid. Focusing on a one-dimensional case, the constitutive law can be expressed as:

\[ \tau = \mu\dot{\gamma} \quad (3) \]

where \( \mu \) is the dynamic viscosity and \( \dot{\gamma} \) is the one-dimensional shear rate.

Unlike in standard Navier-Stokes formulations, the landslide material is assumed to obey an elastic-visco-plastic non-Newtonian, Bingham-like constitutive model to be able to consider also the initial phase of static equilibrium which precedes the activation of the landslide motion. The main assumptions are as follows. The landslide material is incompressible. Only small strains take place in the initial static equilibrium phase, so that linear compatibility can be assumed. In this phase, viscous strains are also small
since nodal velocities are vanishing and the deviatoric effective stress is in general below
the yield limit. When external actions trigger the landslide motion and the elastic limit is
exceeded, large viscoplastic deformations take place, so that the elastic part of the strain
can be neglected. From now onward, the running landslide behaves as a viscoplastic
Bingham fluid. To be able to deal with the static phase, the balance equation (1) contains
a stiffness dependent internal force contribution, in addition to the viscous term. The
primary variables are as usual nodal velocities and pressures, but nodal displacements are
also computed in the static phase through time integration, to allow for the computation
of the stiffness contribution.

In the assumed model the deviatoric stress $\tau$ can be expressed as:

$$
\tau = \begin{cases} 
\tilde{\mu} \dot{\gamma} + G \gamma^e & \text{per } \tau < \tau_y \\
\tau_y + \tilde{\mu} \dot{\gamma} & \text{per } \tau \geq \tau_y 
\end{cases}
$$

(4)

where $\gamma^e$ is the elastic part of the deviatoric strain, $\dot{\gamma} = \dot{\gamma}^e + \dot{\gamma}^p$ is the deviatoric strain
rate and $\tau_y$ a yield shear stress. $\tilde{\mu}$ is an apparent viscosity defined as:

$$
\tilde{\mu} = \mu + \frac{p \cdot \tan(\varphi)}{|\dot{\gamma}|} (1 - e^{-n|\dot{\gamma}|})
$$

(5)

where $\varphi$ is the friction angle. When $\tau < \tau_y$ the behaviour is viscoelastic and dominated
by the elastic term $G \gamma^e$, conversely when the yield stress is reached ($\tau \geq \tau_y$) a viscoplastic
behaviour is obtained. The exponential term in (5) has only a regularization purpose
[2, 8], and has not to be given a constitutive interpretation. The extension to the 3D is
straightforward.

This model can be easily used to describe landslides originated from layered slopes.
Furthermore, the soil transition from an initial static equilibrium state to an unstable
landslide, due to an imposed ground acceleration, can be also accounted for.

4 NUMERICAL EXAMPLE

4.1 Granular flow on a rigid obstruction

The estimation of the impact force of a flowing landslide against a rigid wall is critical
for the safety assessment of protection structures such as earth retaining walls. In [9],
small-scale tests have been conducted to measure the impact force on a rigid wall of a sand flow. In the same paper, numerical tests have also been performed in an Eulerian
framework to analyze and reproduce the laboratory results. The previously described
approach has been used to simulate these tests and its results have been validated against
both the experimental and numerical results in [9].

Figure 1 depicts a schematic representation of the problem geometry. As suggested in
[9], the following physical parameters are used:

$$
\rho = 1379 \text{Kg/m}^3 \quad \mu = 1 \text{Pa s} \quad \varphi = 35^\circ
$$
Other details about the geometry of the problem as well as about the calibration of the parameters can be found in [9].

Four different tests have been performed varying the flume inclination $\theta$. Figure 2 shows the impact force time histories for the different flume inclinations, compared with experimental and numerical outputs of [9]. In all cases, good agreement is obtained.

Finally, figure 3 shows snapshots of the simulation at different time steps for the case of $\theta = 55^\circ$.

### 4.2 Landslide interaction with water reservoir

Water waves generated by fast landslides impinging in water basins can be very dangerous for the safety of the surrounding area. To study this phenomenon, the simplified 2D geometry of the Gilbert Inlet, at the head of the Lituya bay, Alaska, considered in [10] and reproducing the experimental setup in [11], has been used to simulate the motion of a landslide along the slope and the formation and propagation of the water waves on the opposite side.

In Figure 4 different snapshots of the simulation are shown. In [11], an experimental landslide run-up on the opposite side of 152 m has been measured, which compares well with the value of 160 m obtained with the present simulation (a run-up height of 226 m was obtained in [10]).
Figure 2: Granular flow on a rigid obstruction: impact force time histories for different flume inclinations.

REFERENCES


Figure 3: Granular flow on a rigid obstruction: snapshots of numerical flow.


Figure 4: Landslide interaction with water reservoir. From left to right and from top to bottom, different phases of numerical simulation.


ADVANCES ON MESH GENERATION IN NUMERICAL SIMULATION OF FRACTURED NETWORK SYSTEM

SHULI SUN*, JIE SUI* AND MINGWU YUAN†

* LTCS, Department of Mechanics and Engineering Sciences, College of Engineering Peking University, Beijing 100871, China  
e-mail: sunsl@mech.pku.edu.cn, suijie2005@163.com

† Department of Mechanics and Engineering Sciences Peking University, Beijing 100871, China  
e-mail: yuanmw@pku.edu.cn

Key words: Discrete Fractured Network (DFN), Mesh Generation, Mesh Self-adaptation, Moving Grid, Mesh Deformation, Mesh Improvement.

Abstract. Generally, meshing a finite model is often a significant portion of the time for obtaining results from finite element solution. The meshing quality of the discrete model for simulating geological problems influences the accuracy, convergence and efficiency of the solution. This paper focuses on mesh generation in numerical simulation of fractured network system in geological problems. Usually initial cracks exist in such system, and new cracks will generate during deformation of geological media. In this paper, a meshing approach that takes mesh self-adaptation with crack extension into account is proposed, which includes: geometric data cleaning of cracks, construction of cracks, fast node inserting and triangulation, mesh deformation, local re-meshing, and mesh quality improvement. The proposed approach is successfully applied to modeling of two and three-dimensional discrete fractured network (DFN) in geological problems.

1 INTRODUCTION

Modeling and simulation of geological problems are usually tedious and time-consuming. DFNs have a rather complex geometry, especially in 3D case. Usually initial cracks exist in such fractured network system, and new cracks will generate during deformation of geological media. The existence of cracks, especially the extension of the cracks, brings lots of troubles for mesh self-adaptation. The meshing quality of the discrete model influences the accuracy, convergence and efficiency of the solution. As a result a high quality of finite element mesh is desired especially in the neighborhood of fractures. Another problem must be faced is the efficiency of meshing algorithm. Usually a geological media in practical engineering problems includes a lot of cracks. If more than 100 cracks are considered in a practical three dimensional engineering problem, a hundred millions of mesh may be needed to work on. This could cause that the scale of the problem will be massive and have to use a high performance parallel computer.

Many methods are developed to discretize complex fractured media; however, due to the
complexity of the problem, most of them have been used for simple fractured configurations [1]. For example, Blessent et al. [2] have developed a new method to discretize non-planar fractures. Their approach generates fine tetrahedrons close to fractures and identifies certain tetrahedron faces as fracture faces. That method efficiently approximates single-fracture configuration. However, the method becomes computationally costly and difficult to implement for complex fractured media [1].

This paper proposes a meshing approach that takes mesh self-adaptation with crack extension into account, which includes: geometric data cleaning of cracks, construction of cracks, fast node inserting and triangulation, mesh deformation, local re-meshing, and mesh quality improvement. The initial cracks are constructed by inserting pre-defined nodes and clearing up the mesh around the cracks. For new created cracks or crack extension, mesh deformation and local re-meshing technique are applicable. Mesh quality optimization techniques includes smoothing and topological optimization are performed to raise the mesh quality around cracks. The proposed approach is successfully applied to modeling of two and three-dimensional fractured network system in geological problems. So far we still could not find these functions in commercial software packages.

2 CONSTRUCTION OF CRACKS

2.1 Initial construction of cracks

First, the geometric data of cracks must be analyzed and modified if necessary to ensure the applicability for mesh generation and the FE simulation [3]. Some tiny cracks are removed or merged into big ones. Then nodes along cracks are inserted in pairs in advance according to pre-defined element size. Thus the cracks are confined by these pre-setting nodes and described or constructed in some sense (see Figure 1 for 2D illustration).

![Figure 1: Inserting nodes in pairs along crack](image)

With the retained pre-defined nodes around the cracks, next step is to generate mesh in the whole area using fast node-by-node Delaunay inserting techniques (at this moment do not consider the recovery of the boundary to maintain cracks, see Figure 2).
Figure 2: Triangulation by fast node-by-node Delaunay inserting

By clearing up and merging the mesh around cracks, cracks can be exactly constructed, and the whole mesh is completed (see Figure 3). This approach can also be used to generate inner-boundaries inside the region.

Figure 3: Mesh clearing up for crack construction

2.2 Re-meshing for crack extension

The above approach can be used to construct initial cracks. For new created cracks or crack extension, local re-meshing technique is applicable. A local region is first taken out, and then re-generate local mesh of this region with new cracks constructed by the proposed pre-defined nodes method (Figure 4).

Figure 4: Local re-meshing for new created cracks or crack extension

3 LOCAL CORRECTION OF CRACKS

Generally, the above approach works well in construction of cracks for 2D case; however,
it is not always to ensure the topological geometry for each boundary or intersection. Such undesired configurations may happen quite often in random 3D DFNs. It is mandatory to apply a geometrical local correction, which modifies slightly the surface of the crack. In this paper, mesh deformation algorithm described in the following is applied to local correction of cracks.

3.1 Mesh deformation algorithm

Dynamic grid algorithm is one of the key techniques in numerical simulation of fluid-structure interaction problems, where the mesh has to dynamically change to cover the changed solution domain. Due to the efficiency, the mesh deformation approach is one of the most desirable categories of the dynamic grid method.

The widely used mesh deformation methods include spring analogy [4,5] and elastic solid based methods [6,7]. Liu and Qin et al. [8] proposed a simple and efficient method for dynamic grid deformation. This method is based on Delaunay graph mapping of the original mesh. A Delaunay graph of the solution domain is first generated. The geometric movement is carried out using the Delaunay graph with ease and efficiency. The original grid is then mapped back onto the deformed Delaunay graph to provide the new mesh. The main steps of the method are summarized in the following:

1) Generate the Delaunay graph of the original mesh;
2) Locate the mesh points in the graph;
3) Move the Delaunay graph according to the specified geometric change;
4) Relocate/map the mesh points in the new graph.

In the whole procedure, most calculations concentrate in step 2) and 4).

Facing to the most time consuming parts of the Liu’s approach (mesh points locating and mapping/relocating), Sun et al. [9] presented a high efficient algorithm and implementation scheme to speed up the method. Two main improvements were proposed. First, a fast locating technique is developed to locate the background element for the mesh points. Second, an efficient scheme is proposed that avoids most of repeated calculations in relocating the mesh points in the graph. Both time complexity analysis and testing results indicate that the proposed algorithm and implementation scheme has gained substantial speedup compared with the Liu’s original approach, while the memory requirement is even decreased.

To preserve the quality of mesh, Sun et al. [10,11] presented a high quality mesh deformation approach with the aid of barycentric coordinates. The deformation procedure is divided into the following steps:

1) Search for points on each boundary and re-order them anticlockwise;
2) For each grid point in computational domain, evaluate its barycentric coordinates for each boundary;
3) Interpolate displacement of each grid point according to the motion of boundary points;
4) Calculate new location for each grid point;
5) Check illegality of the deformed grid. If found, reduce the motion of boundary points and return to step 3).

This non-iterative algebraic approach is efficient, easy to implement, and works for any type of mesh. Since the interpolation of barycentric coordinates is global, the transition of deformation for the grid is relatively smooth. Thus maintaining the primary qualities of the
grid is expected. Testing results show that the proposed approach can preserve grid quality in original grid even for relatively large deformation cases.

3.2 Local correction of cracks

In this paper, a practical moving grid or mesh deformation procedure which combines the algebraic moving grid approach based on the barycentric coordinate interpolation and background graph mapping is applied to local correction of cracks.

The main steps of the proposed procedure are summarized in the following:
1) Insert the points inside of the deformation area based on the feature of the mesh;
2) Find the points near the crack that will be corrected, and insert these points to set of boundary points;
3) Generate the background mesh using boundary points and the inserted points;
4) Locate the mesh points and calculate the related location parameters;
5) Compute the deformation of the inserted points by using the algebraic method based on the interpolation of barycentric coordinates;
6) Move the background mesh based on the displacements of boundary points and the inserted points;
7) Relocate the mesh points: calculate the new location of the mesh points based on the relative location of the mesh points in old background mesh and deformed background mesh.

This local correction procedure can also be used to deal with crack extension.

4 MESH QUALITY IMPROVEMENT

After the corrections, the modified fracture boundary is well defined; however, the quality of mesh in particular local region is probably unsatisfactory. Performing extra mesh quality improvement is a good choice. Mesh quality smoothing technique is performed to raise the mesh quality around cracks.

Sun et al. [12] proposed a high efficient node repositioning procedure based on chaos search algorithm. The node-by-node manner is employed due to the consideration of computational cost. The process begins with searching the free vertex which needs to be relocated or updated. First, the worst element according to a specified shape criterion is found. Then for each vertex of this element, calculate the mean value of the quality for all elements connected to this vertex. If the calculated mean value of the quality for a vertex is the smallest one, then choose this vertex as the free vertex. Next, find the optimal position of the free vertex by maximizing the minimum quality of the elements connected to this vertex with size and boundary constraints:

$$\text{Maximum } f(x,y,z) = \min \limits_{\text{element}} f_e(x,y,z)$$

Subject to

$$x_L \leq x \leq x_H$$
$$y_L \leq y \leq y_H$$
$$z_L \leq z \leq z_H$$
$$b(x,y,z) \leq 0$$
$$v(x,y,z) > 0$$

(1)
where \( n \) is the total number of elements connected to the free vertex, \( f_i \) represents a given element shape measure for element \( i \), the design variables \( x, y, z \) are coordinates of the free vertex, \( x_L, x_H, y_L, y_H, z_L, z_H \) are size constraints of design variables which can be easily obtained by positions of other vertices of these elements, \( b \) is boundary restraint function and \( v \) is the volume (for 3D) or area (for 2D) of the element \( i \). After obtaining the optimal position of the free vertex by solving Eq. (1), upgrade the mesh and set the current free vertex immovable. Then find the next free vertex, and repeat the local optimization process defined in Eq. (1). Please note that the free vertex is only selected among movable nodes.

Based on above algorithm, a revised smoothing procedure that simultaneously updates all vertices of the worst element in each optimization step is proposed in this paper to raise the mesh quality around cracks. First, locate the worst element according to a specified shape criterion. Next, find all elements that connected to the vertices of this worst element. Then, formulate and solve the following local optimization problem to find the optimal position of the vertices of the worst element (assuming each element has four vertices):

\[
\text{Maximum } f(x, y, z) = \min_{i \in \text{mesh}} f_i(x_i, y_i, z_i, x, y, z) \\
\text{Subject to } \\
x_i^L \leq x_i \leq x_i^H, \; y_i^L \leq y_i \leq y_i^H, \; z_i^L \leq z_i \leq z_i^H \quad (k = 1, 2, 3, 4) \\
b(x_i, y_i, z_i, x, y, z) \leq 0 \\
v(x_i, y_i, z_i, x, y, z) > 0
\]

The mutative scale chaos optimization algorithm [13] is employed to solve Eq. (2). The definition of the non-differentiable objective function in Eq. (2) has an advantage to speed up the optimal searching process: if an element with quality value lower than the existing worst element is detected, then we can immediately discard the current position of the vertices and skip the current search step.

5 TESTING EXAMPLES AND SUMMARY

Two examples are given to demonstrate the effectiveness of the proposed approach. The meshes in examples are generated by fast node-by-node Delaunay inserting techniques. The first example is a 2D discrete fractured network system with individual and intersected cracks (red lines in Figure 5). By the proposed techniques of inserting pre-defined nodes and clearing up, the cracks are constructed precisely.
The second example is a 3D discrete fractured network system in a cubic box with 36 circular disk fractures (Figure 6–8). The mesh includes 114,605 vertexes and 733,798 tetrahedrons. Inserting pre-defined nodes and mesh deformation techniques are used to construct the fractures.

**Figure 5:** Mesh of 2D fractured network system

**Figure 6:** Circular disks (left) and tetrahedral elements around circular disks (right)
Figure 7: Mesh of 3D fractured network system

Figure 8: Mesh of 3D fractured network system
This paper proposes a feasible meshing approach for discrete fractured network (DFN). The initial cracks are constructed by inserting pre-defined nodes and clearing up the mesh around the cracks. For new created cracks or crack extension, mesh deformation and local re-meshing technique are applicable. In order to improve the mesh quality around cracks smoothing techniques are performed. The proposed approach is successfully applied to modeling of two and three-dimensional fractured network system in geological problems and good results is obtained.

ACKNOWLEDGEMENTS

The authors would like to thank the supports of National Basic Research Program of China (2010CB731503), the National Natural Science Foundation of China (11172004, 10772004) and Beijing Municipal Natural Science Foundation (1102020).

REFERENCES


FINITE ELEMENT MODELING OF THERMO-HYDRO-MECHANICALLY (THM) COUPLED PROBLEMS IN FROZEN GROUND ENGINEERING: STATE-OF-THE-ART

YARED W. BEKELE*  

*Sustainable Arctic Marine and Coastal Technology (SAMCoT),  
Centre for Research-based Innovation - CRI  
Norwegian University of Science and Technology, Trondheim, Norway  
E-mail: yared.bekele@ntnu.no, www.ntnu.no

Key words: Thermo-Hydro-Mechanical, Coupled Problems, Finite Element, Frozen Ground Engineering.

Abstract. Fully coupled Thermo-Hydro-Mechanical (THM) modeling has been widely studied in various areas of geomechanics, owing to the multiphase nature of geomaterials. Several researches have dealt with THM coupled modeling of geomaterials in high temperature regimes, but a limited work is available for geomaterials in low temperature regimes. A review and summary of existing work in the literature on THM coupled modeling of frozen soils is presented here. THM coupled modeling in general and its applications are pointed out. The basic governing equations of a coupled THM model in general form, namely mass, momentum and energy balance equations, are discussed. A review of fully coupled models is made and the numerical aspects of THM modeling are briefly discussed. A mechanical constitutive model makes up an important component of a fully coupled THM model and a brief review of existing constitutive models for frozen soils is presented. The models reviewed range from elastoplastic models to viscoplastic or creep and damage coupled models. Some models that consider different approaches from the plasticity framework are briefly reviewed. The state-of-the-art is summarized by pointing out the main aspects of THM coupled modeling and directions for future work.

1 INTRODUCTION

Coupled Thermo-Hydro-Mechanical (THM) modeling is essential in several areas of geomechanics where the multiphysics nature and response of the porous medium needs to be well understood. It has been mostly applied in geomechanics of the high temperature regime environment as in [1], [2], [3], [4], [5], [6], [7], [8] and [9]. Some application areas of THM coupled modeling in high temperature regime geomechanics are geothermal energy extraction, safety assessment of nuclear waste repositories, oil and gas reservoir engineering, underground energy storage and CO2 sequestration. Nowadays, application of THM coupled modeling is gaining popularity in low temperature regime geomechanics, or frozen ground engineering, and has been studied, for instance, by [10], [11], [12], [13] and [14]. Some
specific application areas in frozen ground engineering include frost heave prediction, pipelines buried in permafrost, artificial ground freezing and foundations in cold regions.

In both the high and low temperature regime geomechanics of THM coupled modeling, the main components of the coupled model are the main governing or balance equations and the constitutive equations. The basic governing equations of a coupled model are the mass balance equation, the energy balance equation and the momentum (linear or angular) balance equation. The specific forms of these governing equations strongly depend on the assumptions made for the porous medium in question. The constitutive equations include some known laws for materials, such as Darcy’s law for hydraulics and Fourier’s law for heat transfer, and the equations of the mechanical constitutive model for deformation. The numerical implementation considers the solutions of both the governing and constitutive equations using, for example, the finite element method.

The state-of-the-art in coupled THM modeling of frozen soils is presented here. The governing equations of a coupled model are first presented in general forms. A summary of the literature review from coupled models for frozen soils is discussed. A brief discussion on the numerical aspects of THM modeling, including solution methods and difficulties, is made referring to the studies by [15], [16], [17], [18], [19], [20], and [21].

The mechanical part of a coupled THM model is completed by the constitutive model. The frozen soil models in the literature vary from simple thermo-elastic models to advanced elastoplastic, viscoplastic/creep and damage coupled models. Some of the elastoplastic models that are reviewed here include [10], [22] and [23]. Other researchers studied viscoplastic/creep modeling of frozen soils such as [24], [25], [26], [27], [28], [29], [30] and [31]. Damage coupled modeling is proposed by a number of studies such as by [32], [33], [34], [35], [36] and [37]. Other models for frozen soils in different frameworks, such as hypoplasticity and fracture, have also been proposed by some studies such as [26], [38], [39], [40], [41] and [42].

The coupled THM models and some of the selected mechanical constitutive models are reviewed and discussed in the following sections to show the state-of-the-art status in coupled THM modeling of frozen soils. A summary and conclusion regarding the state-of-the-art status and the observations made are presented at the end.

2 THM COUPLED MODELING

The present section focuses on the general formulations of THM coupled modeling in geomechanics. Several researches have dealt with THM coupled modeling of geomaterials in high temperature regimes in the different application of THM modeling, but a limited work has been done for low temperature regimes. For both temperature regimes in general, thermo-hydraulic, thermo—mechanical and hydro-mechanical interactions are involved, as depicted in Figure 1. The specific types of interaction depend on the assumptions made for the porous medium under consideration.
The volume fraction concept and formulation of the governing equations play an important role in the coupled modeling of a porous medium as frozen soil. These are discussed in the next section, in their general forms.

2.1 Volume Fraction Concept

The volume fraction concept is used to create a homogenized media for a multiphase porous medium. In the volume fraction concept, it is assumed that the porous solid always models a control space and that only the liquids and/or gases contained in the pores can leave the control space, \([44]\). The basis of the description of porous media, using elements of the theory of mixtures restricted by the volume fraction concept, is the model of a macroscopic body, where neither a geometrical interpretation of the pore-structure nor the exact location of the individual components of the body or constituents are considered.

The formulation of the volume fraction concept from \([44]\) is presented here. The porous medium in a representative volume element is assumed to consist of constituents \(\varphi^\alpha\), with real volumes \(\nu^\alpha\), where index \(\alpha\) denotes \(\kappa\) individual constituents. The concept of volume fractions can be formulated as follows:

\[
n^\alpha (\mathbf{x}, t) = \frac{d\nu^\alpha}{dv}, \tag{1}
\]

where \(\mathbf{x}\) is the position vector of the actual placement and \(t\) the time. The volume elements of the real materials and the bulk volume are denoted by \(dv\) and \(d\nu^\alpha\). The volume fractions \(n^\alpha\) in Eq. (1) satisfy the volume fraction condition for \(\kappa\) constituents \(\varphi^\alpha\),

\[
\sum_{\alpha=1}^{\kappa} n^\alpha = 1 \tag{2}
\]

The volume fraction concept assumes that each constituent occupies the whole representative element. Let the real density and the partial density of the constituent materials

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Thermo-Hydro-Mechanical interactions (after \([43]\))}
\end{figure}
be denoted as:

\[ \rho^{\alpha_R} = \rho^{\alpha_R}(x,t), \rho^{\alpha} = \rho^{\alpha}(x,t) \]  

(3)

The above equation indicates that:

\[ \rho^{\alpha}(x,t) = n^{\alpha}(x,t)\rho^{\alpha_R}(x,t) \]  

(4)

The volume fraction concept is important in formulating the governing equations of a THM coupled model. Due to this concept, all geometric and physical quantities, such as motion, deformation, and stress, are defined in the total representative element, and thus, they can be interpreted as the statistical average values of the real quantities.

2.2 Governing Equations in General Form

For THM coupled modeling of a porous medium like a frozen soil, balance equations – balance of mass, balance of linear momentum and moment of momentum, as well as balance of energy – have to be established for each constituent \( \varphi^\alpha \) in consideration of all interaction and external agencies. The interaction effects in the sum have to be equal to zero. Such governing equations in general form are included here as given in [44].

**Balance of Mass**

The principle of mass balance can be formulated in two ways: for the bulk mixture body as a whole and for each individual constituent in such a way that the superposition of the mass balance can be applied. The balance of mass for the individual constituents \( \varphi^\alpha \) requires that the rate of mass \( \varphi^\alpha \) equal a mass term caused by other constituents:

\[ \frac{\partial M^\alpha}{\partial t} = \frac{\partial}{\partial t} \left( \int_{B_\alpha} \rho^\alpha dv \right) = \int_{B_\alpha} \dot{\rho}^\alpha dv \]  

(5)

where \( \dot{\rho}^\alpha \) is the mass supply per volume element. The integration in Eq. (5) covers the domain \( B_\alpha \) of each individual constituent. Using the transport theorem with \( v^\alpha \) as the velocity of phase \( \alpha \), we can write:

\[ \frac{\partial \rho^\alpha}{\partial t} + \nabla \cdot (\rho^\alpha v^\alpha) = \dot{\rho}^\alpha \quad \text{and} \quad \sum_{\alpha=1}^{\hat{\alpha}} \left[ \frac{\partial \rho^\alpha}{\partial t} + \nabla \cdot (\rho^\alpha v^\alpha) \right] = \sum_{\alpha=1}^{\hat{\alpha}} \dot{\rho}^\alpha = 0 \]  

(6)

**Balance of Linear Momentum**

The balance principle of momentum states that the material time derivative of the momentum is equal to the sum of external forces. Thus,

\[ \frac{\partial I^\alpha}{\partial t} = \frac{\partial}{\partial t} \left( \int_{B_\alpha} \rho^\alpha v^\alpha dv \right) = f^\alpha \]  

(7)

The external force vector can be written as the sum of body forces \( \rho^\alpha b^\alpha \) and surface tractions \( t^\alpha \), as well as the interaction forces \( \dot{\rho}^\alpha \) which belong to the volume forces. Using Cauchy’s theorem, the divergence theorem and the mass balance principle, the momentum
balance equation for \( \varphi^\alpha \) with velocity \( \mathbf{v}^\alpha \) and acceleration \( \mathbf{a}^\alpha \) and a summation for the mixture body can be written as:

\[
\nabla \cdot \mathbf{\sigma}^\alpha + \rho^\alpha \mathbf{b}^\alpha + \mathbf{\mathbf{\mathbf{p}}^\alpha} = \rho^\alpha \mathbf{a}^\alpha + \hat{\rho}^\alpha \mathbf{v}^\alpha
\]

\[
\sum_{\alpha=1}^{K} (\nabla \cdot \mathbf{\sigma}^\alpha + \rho^\alpha \mathbf{b} + \mathbf{\mathbf{p}}^\alpha) = \sum_{\alpha=1}^{K} (\rho^\alpha \mathbf{a}^\alpha + \hat{\rho}^\alpha \mathbf{v}^\alpha)
\]

(8)

**Balance of Energy**

The first law of thermodynamics (balance of energy) is the most fundamental relation in the thermodynamics of one-component materials. It states that the sum of the material time derivatives of the internal and kinetic energies equals the rates of the mechanical work and the heat. This balance principle is transferred to the individual constituents. Applying the above statement to the constituents, the following balance principle is obtained:

\[
\frac{\partial E^\alpha}{\partial t} + \frac{\partial K^\alpha}{\partial t} = W^\alpha + Q^\alpha + \int_{B_\alpha} \hat{e}^\alpha dv
\]

(9)

where \( E^\alpha \), \( K^\alpha \), \( W^\alpha \), \( Q^\alpha \) and \( \hat{e}^\alpha \) are the internal energy, the kinetic energy, the rate of the mechanical energy, the rate of the heat of the constituent \( \varphi^\alpha \) and the energy supply to \( \varphi^\alpha \) caused by all other constituents, respectively. The general expressions for \( E^\alpha \), \( K^\alpha \), \( W^\alpha \) and \( Q^\alpha \) are given by:

\[
E^\alpha = \int_{B_\alpha} \rho^\alpha e^\alpha dv
\]

\[
K^\alpha = \int_{B_\alpha} \frac{1}{2} \rho^\alpha \mathbf{v}^\alpha \cdot \mathbf{v}^\alpha dv
\]

\[
W^\alpha = \int_{B_\alpha} \mathbf{v}^\alpha \cdot \rho^\alpha \mathbf{b}^\alpha dv + \int_{\partial B_\alpha} \mathbf{v}^\alpha \cdot \mathbf{t}^\alpha da
\]

\[
Q^\alpha = \int_{B_\alpha} \rho^\alpha r^\alpha dv - \int_{\partial B_\alpha} \mathbf{q}^\alpha \cdot \mathbf{n} da
\]

(10)

Here, \( e^\alpha = e^\alpha(x,t) \) is the specific internal energy, \( r^\alpha = r^\alpha(x,t) \) is the partial energy source and \( \mathbf{q}^\alpha = \mathbf{q}^\alpha(x,t) \) is the partial heat flux vector.

### 3 THM COUPLED MODELS FOR FROZEN SOILS

Coupled THM models for frozen soils have been proposed by some researchers such as [10], [11], [12], [13] and [14]. The basic framework almost in all cases is to define the governing balance equations based on varying assumptions and to propose a mechanical constitutive model. Selection of the coupling parameters also makes an important aspect of the fully-coupled model. A full review of some of the fully coupled THM models is presented below.
One of the earliest studies in the fully coupled THM modeling of frozen soils was presented by [13]. The most important, or simplifying, assumptions they made include: 1) the volume of soil particles remains constant in the freezing process, 2) both unfrozen and frozen soil are isotropic, 3) unfrozen soil is an elastic body with a constant Young’s modulus and 4) Young’s modulus and yield point are independent of the strain rate and confining pressure. The basic moisture transport, heat transport and phase change equations make up the theoretical framework. The stresses and deformations to complete the mechanical part of the fully coupled model are presented as functions of several state variables. Specifically, the creep strain was described according to Prandtl-Reuss law, and Mises equivalent stress was used. The Young’s modulus was expressed as a function of temperature. The finite difference method was used to solve the heat and moisture equations and the finite element method for the mechanical equations. Experimental frost penetration and heave were simulated by the proposed model.

[10] presented the formulation and application of the THM coupled finite element analysis of frozen soil. The basic governing equations of mass, momentum and energy balance make up the theoretical formulation. The thermodynamic equilibrium of freezing soil was described by the Clausius-Clapeyron equation for phase change and the freezing characteristic function by the van Genuchten model with saturation. Darcy’s law was used for fluid flow in the porous medium. An elastoplastic hardening constitutive model, based on the Barcelona Basic model for unsaturated soils, was proposed, which is briefly discussed in the next section. The proposed fully coupled model was applied to the analysis of frost heave prediction and reasonably good agreements were obtained. CODE_BRIGHT was used for the simulation.

Another fully coupled THM model for frozen soil was proposed by [11]. The effective stress equation for frozen soil was presented as a function of thermal stress, ice swelling and pore pressure. The continuity equation was derived based on the conservation of mass and Darcy’s law. The energy conservation equation considered the deformation energy, adsorbed heat energy of soil skeleton and ice, pore pressure energy, adsorbed heat energy of pore water and phase-changing energy. No advanced mechanical constitutive model was included in the fully coupled model but rather simple thermo-elasticity was considered. The numerical implementation was performed by solving the governing equations in a finite element program. A simple thermo-elastic consolidation and numerical modeling of a pile foundation were studied by the proposed model.

A relatively recent attempt in fully coupled THM modeling of frozen soil is made by [14]. The theoretical formulation comprises description of the thermal, hydraulic and stress-strain fields. The thermal field uses the modified Fourier equation with both conduction and convection and empirical equation for the overall thermal conductivity. The hydraulic field uses the mixed type Richard’s equation with modification for ice term to described fluid motion. Analogy of the soil water characteristic curve from unsaturated soils was applied to describe the freezing process. In the stress-strain field, the total strain was defined as the sum of elastic, thermal, phase change of water, change of matric potential and initial strains. COMSOL Multiphysics was used for the numerical simulation and problems where benchmark data are available were analyzed. The proposed model fails to consider the
dependence of viscosity on temperature and there was no proper mechanical constitutive model.

[12] also proposed a fully coupled THM model for a frozen medium. The continuity flow equation was derived from fluid mass balance where the kinematic flow was defined by Darcy’s law. To define the equilibrium equation, the total stress was decomposed into effective and hydrostatic components by using the Bishop parameter. The energy conservation equation was obtained by combining energy conservation equations for solid and fluid constituents. For the mechanical part of the model, the Mohr-Coulomb criterion was used to define the yield locus and plastic potential, with associated flow. The governing equations were solved by using the finite element method based on Galerkin’s formulation. Temperature distribution and deformation close to a heat source are investigated as an evaluation problem, in addition to simulation of a freeze-thaw experiment.

Fully coupled THM modeling is a relatively new approach for frozen soils compared to other areas of geomechanics, such as unsaturated soils. A few other researchers have also attempted the coupled modeling of frozen soils. [45] discussed the numerical simulation of water-heat coupled movements in a seasonally frozen soil. [43] presented the simulation of a fully coupled THM system in freezing and thawing rock. [46] proposed a model for coupled heat, moisture and stress-field of saturated soil during freezing. Perhaps fully coupled THM modeling is more extensively studied in the geomechanics of high temperature regime geomaterials. The experience and findings from these studies are believed to provide an important background for modeling in low temperature regimes.

4 MECHANICAL CONSTITUTIVE MODELS

A review of mechanical constitutive models for frozen soil is presented in this section. The reviewed models are categorized in to different groups based on the complexity of the model and the plasticity framework used. The models reviewed range from simple thermoelastic models, advanced elastoplastic and viscoplastic models to damage coupled models. These models are discussed separately in the following sections.

4.1 Elastoplastic Models

Elastoplastic modeling of frozen soils has been a subject for some researchers including [10], [22] and [23]. [10] adopt an effective stress definition based on Bishop’s parameter where the weighted pore water and ice pressures are subtracted from the total stresses. The proposed elastoplastic constitutive model was developed based on the Barcelona Basic Model (BBM) for unfrozen-unsaturated soils, [47].
[22] proposed an elastoplastic frozen soil model based on triaxial compression test results under varying confining pressures and at selected temperatures, implying an isothermal model. They proposed a strength criterion based on a combination of an extended Lade-Duncan strength function in the $\pi$-plane and in the $p$-$q$ plane. Another isothermal elastoplastic model developed based on triaxial compressive test results at a discrete temperature is proposed by [23]. The presented model focuses mainly on frozen soils subjected to high confining pressures and it was observed that the strength of frozen soil increases to a peak value with increasing confining pressure, but with a further increase in confining pressure, the strength decreases due to pressure melting and crushing phenomena.

4.2 Viscoplastic/Creep Models

Viscoplastic or creep modeling of frozen soils has been studied by some researchers as found from the literature study. Some of the authors that studied this include [24], [25], [26], [27], [28], [29], [30] and [31].

[24] studied creep of artificially frozen soil by conducting a series of experiments at
discrete temperatures and by proposing an isothermal viscoplastic constitutive model based on the results from the experiments. [25] proposed a combined creep and strength model by a single (unified) constitutive equation. The model was developed for the entire primary, secondary and tertiary creep stages and the long-term strength of frozen soil under multiaxial stress at both constant stresses and constant strain rates. [26] presented a brief review of some frozen soil creep models and proposed a simple hypoplastic constitutive model. [27] discussed the creep behavior of frozen soils starting with uniaxial state of stress and proposed expressions for the effect of temperature on creep rate and strength.

4.3 Damage Coupled Models

A number of researches claim that failure in frozen soils can be associated with damage. Some of the studies that deal with damage coupled modeling of frozen soils include [32], [33], [34], [35], [36] and [37].

[36] proposed a statistical damage constitutive model for warm frozen clay based on the Mohr-Coulomb criterion. They argue that there is a need for stochastic damage modeling of frozen soils due to the microdefects randomly distributed in frozen soils leading to randomness of the mechanical properties. [33] also studied a damage statistical constitutive model and stochastic simulation for warm ice-rich frozen silt. [35] developed a constitutive model of frozen soil with damage and studied a numerical simulation for the coupled problem. One of the earliest studies which tried to couple creep of frozen soil with damage mechanics was presented by [34]. The modeling was based on experimental study and microscopic observation.

4.4 Other Models

The most common frozen soil models in the well-known plasticity frameworks of elasticity, elastoplasticity and viscoplasticity have been discussed in the previous sections, including damage coupled models. A few other models that consider different approaches are available in the literature. A model based on the hypoplasticity framework is proposed by [26], where a brief review of existing creep models is also presented. [38] proposed a nonlinear fracture model for frozen soil and presented the corresponding numerical simulation. [39] used the method of temperature-time analogy to determine the long-term strength of frozen soil in triaxial compression. A photo-viscoelastic creep modeling approach for frozen soils was proposed by [40]. [41] discussed the constitutive modeling of saturated frozen silt in torsion. The triaxial creep modeling of frozen soil under dynamic loading was studied by [42].

5 NUMERICAL ASPECTS OF THM COUPLED MODELING

Coupled numerical modeling of thermal, hydraulic and mechanical properties together is observed to exhibit a highly nonlinear behavior. The governing and constitutive equations involved in formulating a fully coupled THM model are usually complex and large in number. Solving such equations using the finite difference or the finite element method has the potential to create several numerical difficulties.
Studying the numerical difficulties and proposing optimized modeling approaches and solution methods for coupled THM models has been the subject of a number of researchers. [15] proposed a parallel finite element scheme for THM coupled problems in porous media. A parallel approach was also presented by [16]. [17] discussed the numerical difficulties and computational procedures for THM coupled problems of saturated porous media. [18] proposed an object-oriented finite element analysis of THM coupled problems in porous media. [19] and [20] used an algebraic multigrid method for solving coupled THM coupled problems. [21] discussed the application of standard and staggered Newton schemes in THM coupled problems.

6 SUMMARY AND CONCLUSIONS

A state-of-the-art review of coupled THM models for frozen soils has been presented. The applications and components of a coupled THM model were discussed and the governing equations in general form were presented. A review of the related literature was performed in two parts. Firstly, studies which presented coupled THM models for frozen soils were reviewed and discussed. And secondly, separate mechanical constitutive models for frozen soil varying from simple thermoelastic models to advanced elastoplastic, viscoplastic or creep and damage coupled models were reviewed. The numerical aspects of THM coupled modeling and mechanical constitutive models outside the well-known plasticity framework were briefly mentioned.

The following observations have been made from the state-of-the-art review of THM coupled modeling of frozen soils:

- The governing equations of a coupled THM model strongly depend on the assumptions made for the multiphase porous medium such as phase composition, saturated or unsaturated, mechanisms of heat transfer etc.
- THM coupled modeling in geomechanics has been widely studied in the high temperature regime but is gaining popularity for the low temperature regime i.e. frozen ground engineering.
- All the mechanical models reviewed here deal with isothermal conditions i.e. they are based on a discrete temperature, or experiments at a constant temperature. Non-isothermal constitutive modeling remains a challenge.
- Mechanical constitutive modeling of frozen soil has been studied under the various branches of the plasticity framework i.e. elasticity, elastoplasticity, viscoplasticity or creep and damage. Damage coupled modeling was frequently used in recent studies.
- Numerical solution of the coupled governing and constitutive equations shows a strongly nonlinear behavior and usually results in numerical difficulties during implementation.

In general, there remain several challenges both in the theoretical formulation and the numerical implementation of THM coupled models for frozen soils. A further study aims to look at some of the challenges and contribute to a progress in the research topic.
ACKNOWLEDGEMENT

The author wishes to acknowledge the support from the Research Council of Norway through the Centre for Research-based Innovation SAMCoT and the support from all SAMCoT partners, and particularly Knut V. Høyland for reviewing the paper and Thomas Benz for discussions.

REFERENCES


32. Lai, Y.M., L. Jin, and X.X. Chang, *Yield criterion and elasto-plastic damage constitutive
INFLUENCE OF CRACKS ON THE SOIL-ATMOSPHERE INTERACTION: NUMERICAL COUPLED MODEL OF THERMO- ATMOSPHERE- POROUS MEDIA

D.TABBAL *†, I. SHAHROUR *, A. AL-QADAD *, F. HAGE CHEHADÉ † AND M. SADEK *

* Laboratory of Civil Engineering and Geo-Environment (LGCgE) University Lille1 of sciences and Technology Polytech-Lille – Av. Paul Langevin – 59655 Villeneuve d’Ascq cedex, France e-mail: diala_tabbal@hotmail.com, www.lgcge.fr

† Ecole Doctorale des Sciences et de Technologie (EDST) Lebanese University, Beirut, Lebanon e-mail: edst@ul.edu.lb , www.edst.ul.edu.lb

Key words: Clay, cracks, soil-atmosphere interaction, evaporation, finite element model

Abstract. Soil shrinks as it desiccates, and the magnitude of shrinkage can be large for clayey soils. The drying of soil leads to cracks formation, causing high suctions to develop within. Cracks expose the deep soil and more evaporation can be expected in dry periods. To illustrate the effect of cracking, a numerical model of soil-atmosphere interaction has been developed taking into account the thermo-fluid coupling of an unsaturated clay soil. The model is used to simulate the evolution of evaporation during the drying process. The main results show a significant influence of the presence of cracks on the evaporation. This study also offers a simple method for taking into account the presence of cracks in the soil-atmosphere exchange.

1 INTRODUCTION

Desiccation cracking is a common phenomenon in clay materials. Since superficial soils are directly exposed to seasonal changes in climate, desiccation cracks develop at the surface of soils as they dry and shrink. Cracks induce changes in soil surface patterns and can greatly modify their hydraulic properties. Moreover, shrinkage cracks create in soil zones of weakness affecting the stability of buildings and structures that are constructed on clayey soils.

Many attempts have been made to follow the initiation of cracks and to describe the cracking process. Most of these studies are based on field and laboratory experiments [1-9].

In view of modeling the cracking process, there is very little literature on the impact of cracks on evaporation from soil surface and on suction development [10-12]. In this study, a numerical model is developed to investigate the effect of cracks on evaporation evolution from a soil surface exposed to dry conditions. Then a simplified approach able to reproduce the impact of cracks on the soil behaviour is presented and discussed.
2 NUMERICAL MODEL

The theory is based on the principles of Darcy’s law and Fick’s law to describe the flow of liquid water and water vapor in the saturated - unsaturated soil below the surface. Penman-Wilson (1948) [13] method for evaporation is used to predict evaporation from the soil surface, as follows:

\[ E_{\text{vap}} = \frac{\Delta R_a / L_v + \eta E_a}{\Delta + \frac{1}{\varphi_s}} \]
\[ E_a = 0.35(1 + 0.15U_a)\left(\frac{1}{\varphi_a} - \frac{1}{\varphi_s}\right)P_a \]

\( \Delta \) is the slope of the saturation vapor pressure versus temperature curve at the mean temperature of the air, \( R_a \) is the net radiation at the soil surface, \( L_v \) is the latent heat to vapor, \( \eta \) is the psychrometric constant, \( U_a \) is the wind speed, \( P_a \) is the vapor pressure in the air above the evaporation surface; \( \varphi_a \) and \( \varphi_s \) designate the relative humidity in the air and at the soil surface, respectively.

Permeability is calculated using (Fredlund et al. 1994; Leong and Rahardjo 1997) formulation [14,15] and Wilson’s equation (1990) [16] was used to estimate the temperature at the surface of soil. For water retention curve, the Van Genuchten equation (1980) [17] is used and Van de Griend and O’Neill (1986) [18] equation is used to compute the thermal conductivity.

2.1 Model description

A conceptual model is developed that quantitatively describes the soil-atmosphere interaction for a clayey soil undergoing desiccation. Figure 1 shows finite element mesh and the domain boundary used in numerical analysis. The model was 1 m deep, modeling the entire depth of soil and 20 cm wide, half of the physical length because of the symmetry. Mesh element sizes were reduced near the soil surface to increase model accuracy. To create a domain 1 m deep by 20 cm wide, 21 rows and 20 columns, 800 elements were built. The numerical model was used to simulate the results of 30 days evaporation for a clayey soil undergoing desiccation.

2.2 Boundary & Initial conditions

2.2.1 Initial Conditions

The same initial conditions were used in all simulations:
- Soil initially saturated given by: \( \Psi = 0 \) for \( t = t_0 \) and
- Surface temperature set at 20°C for \( t = t_0 \)

2.2.2 Boundary Conditions

Flow was allowed only from the upper surface to simulate the evaporation and thermal flux. Outward flux was not allowed at the right, left and lower boundaries (Figure 1). For the sake of simplicity and to represent the drying environment, weather parameters were set constant for all the period of simulations with a temperature set at 20°C, relative humidity at 60%, solar radiation at 800 W/m² and air velocity set equal to 60 Km/h.
2.3 Soil properties

The model requires several input parameters specific to the soil undergoing desiccation. Hydraulic conductivity and water retention characteristics are needed to predict the evolution of the suction profile with time, relative humidity of the air and at the surface are needed to evaluate evaporation from soil surface.

3 RESULTS OF THE NUMERICAL SIMULATIONS

Results of the evaporation from the interface soil-atmosphere for intact and cracked cases are shown in the following sections.

3.1 Case 1: Intact soil
Figure 2 presents the cumulated evaporation flux for the 30 days. Evaporation shows a value of 400 mm/day at the end of the first day then increased almost linearly with time and attains a value of 9500 mm/day after 30 days. Because soil is exposed to drying conditions, loss of water starts since the first day of simulations and continues; which, in turn, promote the evaporation process and increases the evaporation rate along the 30 days.

3.2 Case 2: Cracked soil

The purpose of this section is to show the impact of soil cracks on evaporation from soil surface. To that purpose, a crack of 5 cm depth and 1 cm width was incorporated into the model described above. The geometry of the cracked model and the boundary conditions are shown in Figure 3. We suppose that a crack occurs every 40 cm at soil surface. And, since only half of the model is analyzed, the width of the model was then equal to 20 cm with a depth of 1 m. The rightmost column represented the crack, with a new vertical front of 5 cm. To increase model accuracy, mesh element sizes were reduced near the crack and soil surface. Row widths away from the crack increased from 1.0 to 10 cm (Figure 3).

In the case of cracked soil, evaporation occurs from both surface and crack front. Then in the boundary conditions, the crack front is subjected too to atmosphere conditions. Simulations were conducted in the same manner of the intact soil using the same soil properties, same initial conditions and same weathers parameters.
4 RESULTS COMPARISON: CRACKED/INTACT SOIL SURFACE

To show the influence of crack on the evaporation, the results from intact and cracked models are combined and compared in Figure 4. It can be seen that the addition of the crack increased the evaporation from the first day and for the whole days of the simulations. Crack was found to increase evaporation by 17% and 10% after 15 days and 30 days respectively. The presence of crack exposes the deep soil and more evaporation can be expected in dry periods since the evaporation surface is increased. Cracks provide an excess of water loss mainly due to an increase in the hydraulic conductivity resulting in a higher evaporation rate.

**Figure 3:** Geometry and boundary conditions of FEM used in the simulations of cracked soil

**Figure 4:** Evaporation pattern from intact and cracked surface
5 A simplified approach for cracked model

In this approach, a new finite element constitutive model, where we propose a new tool for modelling cracks in a simplified way is developed and discussed. It was found from the previous study in this paper that, the surface of evaporation is the key parameter that controls the changes in the evaporation evolution and the cracking process. Hence, we suggest that, for a cracked soil, the water evaporates from a fictive length \( L_{eff} \) with the same boundary conditions as the intact soil surface.

As the length \( L_{eff} \) is larger than \( L \) intact soil and in order to reproduce the effect of cracks, \( L_{eff} \) was estimated to be equal to \( L_{intact} \) soil multiplied by a parameter \( \lambda \).

Numerical modelling to demonstrate the validity of the new methodology described herein was carried out. The results of the homogenized method subjected to desiccation conditions were compared to the previous model. The same soil model of 1m x 20 cm with a crack of (1cm x 5cm) is considered which corresponds to \( \lambda =1.25 \) in the proposed model.

5.1 Validation of the proposed model

The results of non-cumulative evaporation using both models are shown in figure 5. The graph show that the results from both approaches are very similar and that for all days of simulations.

The proposed approach was shown as an effective tool for modeling the effect of cracks on soil performance especially when the geometry of the sample is difficult to reproduce.

![Figure 5: Evaporation flow pattern from both approaches](image)

6 CONCLUSION

Desiccation cracking is a common phenomenon in clay materials, which can significantly affect the performance of soil. In this study, the effect of soil cracks on evaporation was investigated. Two cases were studied: intact case and cracked case. Two finite element models for intact and cracked surface were created and evaporation was simulated using a multiphase numeric simulator. The model quantified the contribution of soil cracks on
evaporation rates for a simulation of 30 days. Results were then compared with those of an intact soil. They show the important contribution of cracks in evaporation flux. In cracked soil, further loss of moisture from soil occurs as direct evaporation from the fronts of crack. Actually, the simulations that don’t consider the presence of crack could severely underestimated evaporation flow.

A new approach to reproduce the impact of cracks on the soil behaviour in a simplified method was presented. Results from both models were shown and discussed; Remarkable agreement can be seen between the two models. The proposed model was shown to be able to reproduce the cracked soil behavior very well.

In further study, cracks should be taken in consideration when modelling soil-atmosphere interaction in dry periods.

7 REFERENCES


INFLUENCE OF FLUID-MECHANICAL COUPLING IN GAS GENERATION IN UNDERSATURATED PETROLEUM RESERVOIRS

CARLOS E. R. LAUTENSCHLÄGER*, GUILHERME L. RIGHETTO*, NELSON INOUE* AND SERGIO A. B. DA FONTOURA*

*ATHENA – Computational Geomechanics Group
GTEP – Group of Technology in Petroleum Engineering
PUC-Rio – Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil

e-mail: fontoura@puc-rio.br, web page: http://www.gtep.civ.puc-rio.br/

Key words: Reservoir Geomechanics, Blackoil Reservoir, Fluid-mechanical Coupling, Gas Release, Geomechanical Effect.

Abstract. Among the several mechanisms of producing an oil reservoir, the gas expansion mechanism is an important primary recovery process. During the depletion of the reservoir, the pore pressure may reach values below the bubble pressure of the oil, allowing the gas release. From the geomechanical point of view, the change in pore volume, due to production, changes the dynamics of gas generation, since it is dependent upon the change in pore pressure. Studies considering the fluid-mechanical coupling show the relationship between variations of fluid pressure and porous structure of the reservoir. This work aims to study the influence of the fluid-mechanical partial coupling (one and two-way) in the process of gas release during recovery of hydrocarbon. It was used the partial coupling methodology developed by ATHENA/GTEP – PUC-Rio. The model called “A” has only one producing well, while the model called “B” has four injection wells, besides the producer. Initially, the oil present in the reservoir is in undersaturated condition. In model “A” was observed that the pressure drop of fluid is more accentuated, until it reaches the bubble pressure, when considering the two-way coupling. Consequently, the gas release initiation occurs earlier than one-way coupling scheme. After starting gas generation, the rates of pressure change in both partial coupling scenarios tend to equalize. In terms of compaction and subsidence, it was observed most significant displacements values in two-way coupling, highlighting the rigorous consideration of the geomechanical effects in the applied methodology. In model “B”, it was observed that the consideration of the two-way coupling resulted in a recovery scenario without generation of gas, unlike the results shown by the one-way coupling in which gas was generated during 40% of total simulation time. In geomechanical terms it was observed, as presented previously, that the values of vertical displacement were greater in the two-way coupling. The methodology used in this paper proved to be capable of simulating coupled process in a blackoil reservoir, as could be observed by the results. Furthermore, the use of one-way partial coupling scheme, which is widely used in the oil industry, showed results quite different in terms of gas liberation, when are compared with the two-way partial coupling scheme, which was developed in a more rigorous way.
1 INTRODUCTION

A petroleum reservoir may contain fluids in various configurations of phases, and depending on the state of pressure and temperature that found, there may be a predominance of one over the other phases. The properties of the fluids contained in hydrocarbon reserves, as well as their relationship with the types of existing reservoirs, consists of a theme widely presented and discussed in classical and modern literature, see [1, 2, 3].

The combination of fluid properties can be made from theoretical and experimental evaluations, using thermodynamic laws and PVT test results, leading to predictions about the behavior of the reservoir during its production life. It is denoted, however, that currently interests are not solely directed to hydraulic point of view of oil extraction, but also to effects of this extraction on the geomechanical behavior of reservoir. Such effects have great practical relevance, as have been reported in recent publications on this subject [4, 5, 6, 7].

In the context presented, it is observed the importance of the jointly evaluation of geomechanical effect in the reservoir and the changes in produced fluid phase, since both are a result of the pressure gradient resulting from the reservoir development. Using concepts from fluid-mechanical coupling, associated to production mechanisms involving gas release, it becomes possible to establish a panorama of geomechanical action on the dynamics of the phases present in the system.

2 GAS GENERATION IN BLACKOIL RESERVOIRS: AN OVERVIEW

Among the various mechanisms of producing an oil reservoir, the solution gas drive is an important primary production process. According to [8], this mechanism is favourable to the production of oil reserves with dissolved gas, since the pressure drop in the reservoir, induced during production, can generate gas release. From the geomechanical point of view, the liberation of dissolved gas due to pressure relief caused by production is not independent of geomechanical effects, since the consideration of these effects causes substantial variations in the pressure state of the reservoir [5, 6].

Figure 1, adapted from [2], shows two situations in which an oil reservoir may be found: (a) shows a condition of undersaturated oil (fluid pressure higher than bubble pressure) and (b) shows a condition in which the fluid pressure falls below the bubble pressure.

![Figure 1: Solution gas drive reservoir (a) above the bubble point pressure; (b) below bubble point pressure](image-url)

These production scenarios consider that the fluid is trapped by a sealant layer, and that the water influx in production is not relevant, in such a way that the boundaries do not permit flow in either direction. It should be emphasized that, because of the condition of flow isolation, the reservoir will undergo a gradual reduction of pressure during production, resulting in the...
expansion of the fluids present in the formation and the compaction of mineral skeleton of reservoir rock [8]. These phenomena contribute to the expulsion of fluid from the reservoir, allowing the production. Nevertheless, the pressure reduction may even result in the phase change of the dissolved gas, which upon become free, configures an important agent in this mechanism of production. It is therefore evident the interrelationship between the geomechanical effects and behavior of fluid in production as well as the need to perform a separate evaluation for possible situations, with and without free gas together with the oil.

In the undersaturated condition, the interaction between the compressibilities values of the components influences on the production capacity of the reservoir, mainly because there is no predominance of one phase over the other, in terms of compressibility. In this case the influence of variation of the pore volume due to geomechanical effects affects the system response front of the change of pressures resulting from the production.

In the situation where the oil pressure is below the bubble point pressure, the solution gas present in the oil is released and a gas saturation in the free form is developed in the reservoir. The compressibility of the gas, according to [2], can be two orders of magnitude higher than the value of the water compressibility, and about thirty times the compressibility of the porous structure. Thus, it is observed that the geomechanical effect in the reservoir may be differentiated in situations with and without free gas into the fluid, due to the difference between the magnitudes of phases compressibilities.

Assuming the compressibility of the system as being governed primarily by gas compressibility, it is assumed that the variation of pore volume resulting from the change in the stress state of rock due to production, does not significantly interfere on the gas pressure as occurs on pressure of the fluid in the liquid state. In terms of fluid-mechanical coupled analysis, this consideration is equivalent to saying that, upon fluid pressure reaching the bubble point, the phenomenon of coupling changes its nature, since gas is more compressible and the production mechanism does not refer solely to reservoir compaction.

3 PANORAMA OF THE FLUID-MECHANICAL COUPLING SCHEME

Studies conducted by the Computational Geomechanics Group – ATHENA/ GTEP – PUC-Rio showed substantial influence of geomechanical effects on the history of fluid pressures during the development of a reservoir, and also on other aspects such as subsidence and compaction. The simulated cases and reported in the literature [5, 6] considered no free gas, falling on the condition of undersaturated oil. In such situations, the use of iterative techniques in the fluid-mechanical coupling, i.e. with exchanging information between the flow and stress simulators, gave results of better quality than explicit coupling, when compared to full coupling, since the fluid and the pore structure compressibilities influence the production process. The use of partial coupling in the simulations with complete saturation of fluids in porous media has led to significant changes in rates of pressure change over time. However, studies involving coupled simulations with the presence of free gas in the porous medium should be performed taking into account that the gas has a high compressibility value regarding to other components, and the observed behavior must be distinct from complete saturation liquid.

The coupling methodology developed by the ATHENA/ GTEP – PUC-Rio consists of a one and two-way partial coupling scheme, where the flow variables (pore pressure and saturation of the phases) and the stress variables (displacement field, stress and strain state) are calculated separately and sequentially, by a conventional reservoir simulator and a stress
analysis program, respectively. The coupling parameters are exchanged at each time step until reaching the convergence. The quality of this methodology was ensured by the rigorous development of a coupling parameter which approximates the geomechanical response to the fully coupled behavior. In this work, it will be shown just the equations of the flow problem and the stress analysis problem. For more details about the development of the formulation see [6].

The flow equation can be obtained by combining the mass conservation equation and the Darcy’s law. The law of mass conservation is a material-balance equation written for a component in a control volume. In hydrocarbon reservoirs, a porous medium can contain one, two and three fluid phases. The governing flow equation for the conventional reservoir simulation and the governing equation used in the fully coupled scheme are given by Eq. (1) and Eq. (2), respectively. In the conventional reservoir simulation, the porosity is related to pore pressure through the rock compressibility using a linear relation, and in the fully coupled scheme, the porosity equation is composed of four components that contribute to the fluid accumulation term.

\[
\left( c, \phi^0 \right) \frac{\partial p}{\partial t} - \frac{k}{\mu} \nabla^2 p = 0
\]

\[
\left[ c, \phi^0 \right] \frac{\partial p}{\partial t} - \frac{k}{\mu} \nabla^2 p = -\alpha \frac{\partial \varepsilon_v}{\partial t}
\]

where \( \phi \) is the porosity, \( p \) is the pore pressure, \( t \) is the time, \( k \) is the permeability, \( \mu \) is the viscosity, \( c_f \) is the fluid compressibility, \( c_s \) is the solid matrix compressibility, \( c_r \) is the rock compressibility, \( \alpha \) is a Biot’s coefficient and \( \varepsilon_v \) is the bulk volumetric strain.

The formulation of the geomechanical problem takes into account the equilibrium equations, stress-strain-displacement equations, rock-flow interaction and the boundary conditions. The governing equation of the geomechanical problem may be written as indicated in Eq. (3).

\[
G \nabla^2 u + \frac{G}{1-2\nu} \nabla \nabla \cdot u = \alpha \nabla p
\]

where \( G \) is the shear modulus, \( u \) is the nodal displacement and \( \nu \) is the Poisson ratio.

The methodology used herein for the coupling between flow and stress problem was described in [5, 6]. The coupling is achieved through a convenient approximation between of the flow equation of the conventional reservoir simulation and the flow equation of the fully coupled scheme. In this methodology, the effect of solids compressibility is removed from the fully coupled scheme and the effect of volumetric strain of the porous medium is added to conventional reservoir simulation.

The parameters responsible for the coupling, which honor the fully coupling equation, are the porosity \( \phi \) and the pseudo-compressibility \( c_p \), presented in Eq. (4) and (5) respectively. These parameters are updated during each iteration through the coupled analysis.

\[
\phi = \phi^0 + \alpha (\varepsilon_v - \varepsilon^* \phi^0) + \frac{1}{Q} (p - p^*)
\]

\[
c_p = \frac{\varepsilon^*_{v}^{n+1} - \varepsilon^*_{v}^{n}}{\phi^0 (p^*_{v}^{n+1} - p^*_{v}^{n})}
\]

Furthermore, the partial coupling between the stress analysis program and the conventional reservoir simulator is reached using a staggered procedure, implemented in a C++ code [6].
The fact that the gas is more compressible than other medium components as well as being in process of gradual release of its dissolved state, it provides more significant production rates than those caused by the reduction of the pore volume, given the smallest variations of pressure and consequent changes in stress state. Therefore, the use of an iterative coupling scheme in this particular situation probably does not lead to substantial differences in the rate of change of pressure, since the variation of the pore volume would not be the dominant mechanism in the recovery of hydrocarbons. The use of a fluid-mechanical coupling scheme, to perform the analysis of the transition between undersaturated and saturated oil configurations, consists of a relevant study, since denoted the influence of geomechanical effects on the pressure variation in the cases where the fluid has not reached the bubble point pressure. Such assessments will be conducted via blackoil models built in this study, as presented in the following sections.

4 COUPLED SIMULATIONS IN BLACKOIL SCENARIOS

4.1 Scenarios

The simulations were performed using two different numerical models, based on model presented by [5, 6, 7]. The first model, called A, consists in a simplified reservoir with one producer well placed in its center. The second model, called B, presents the same configuration aforementioned, however, it was added four injector wells. Figure 2 shows these two configurations.

![Figure 2. Blackoil reservoirs, (a) model A and (b) model B](image)

It should be noted that the analysis time for the models A and B was 1200 and 5000 days respectively. The simulation time for the model B was longer than model A because the interest was observe the influence of the injection in long-term. Table 1 presents the geomechanical and fluid properties adopted for both models. The PVT data were adopted according to a real reservoir available.
Table 1. Rock and fluid properties for both models

<table>
<thead>
<tr>
<th>Properties</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>API (°)</td>
<td>27.5</td>
</tr>
<tr>
<td>Initial rock compressibility</td>
<td>$3.33 \times 10^{-4} \text{ psi}^{-1}$</td>
</tr>
<tr>
<td>Horizontal permeability</td>
<td>$9.86 \times 10^{-14} \text{ m}^2$</td>
</tr>
<tr>
<td>Vertical permeability</td>
<td>$9.86 \times 10^{-15} \text{ m}^2$</td>
</tr>
<tr>
<td>Initial porosity</td>
<td>0,25</td>
</tr>
<tr>
<td>Young’s modulus (reservoir)</td>
<td>0.689 GPa</td>
</tr>
<tr>
<td>Young’s modulus (surrounding rock)</td>
<td>6.89 GPa</td>
</tr>
<tr>
<td>Poisson’s ratio</td>
<td>0.25</td>
</tr>
<tr>
<td>Production rate – model A</td>
<td>50000 BBL/day</td>
</tr>
<tr>
<td>Production rate – model B</td>
<td>6500 BBL/day</td>
</tr>
<tr>
<td>Injection rate – model B</td>
<td>9000 BBL/day</td>
</tr>
</tbody>
</table>

In the following items will be discussed the results (pore pressure, gas release, compaction and subsidence) obtained from the numerical models aforementioned.

4.2 Results

This section is divided in two distinct parts. In the first one, it will be presented the results obtained considering the model with one producer well (model A). In another part, it will be discussed the results of the model with injectors and producer wells (model B).

- **Model A**

  Figure 3 shows the comparison between the conventional reservoir simulation (IMEX) and the simulation partially coupled ABAQUS-IMEX in two-way. It was evaluated the pore pressure distribution over time and the produced gas rate in reservoir conditions. It can be mentioned that the bubble pressure for oil is 5000 psi.

  It can be seen that the pore pressure decay rate was higher considering ABAQUS-IMEX two-way coupling than the conventional reservoir simulation (IMEX). In this case, gas generation occurs 160 days before the reservoir simulation, highlighting the importance of geomechanical effects. Besides that, it was observed that after gas release within reservoir, the pore pressure decay rate is quite similar in both methodologies compared, noting that fluid-mechanical behavior is also affected by this process. The pore pressure decay rate curves are presented in Figure 4.
It was evaluated also, the average pore pressure within reservoir considering one and two-way fluid-mechanical coupling schemes. It should be noted that one-way coupling simulation presents the same results when compared with conventional reservoir simulation, once the geomechanical effect does not affect the flow simulation, as discussed in [5, 6]. Figure 5 presents the average pore pressure behavior considering one and two-way coupling. It was clear that the correct information interchange between simulators (two-way coupling ABAQUS-IMEX) resulted in a higher pore pressure drop over production time. This fact makes sense physically, by virtue of the increased stiffness of system due consideration of surrounding rocks.

Figure 6 shows a set of outlook of pore pressure distribution considering one and two-way
coupling methodologies. It can be seen that in the iterative simulation, the variation of pore pressure was broadest than the simulation performed by one-way coupling (see Figure 6). Figure 6 is composed by a set of images, which are: top view, cross section through the producer well and tridimensional.

![Figure 5. Average pore pressure over production time for one and two-way coupling schemes – model A](image1)

![Figure 6. Outlook of pore pressure variation after 1200 days of production considering one and two-way coupling schemes – model A](image2)

It was evaluated also, the compaction of reservoir and the subsidence of seafloor, as can be seen in Figure 7. The measurement elements chosen were placed in the producer well position in relation to reservoir and overburden. Again, the behavior observed is quite similar those analyzed previously, i.e., both results present important differences between methodologies, in which two-way coupling shows values of compaction and subsidence higher than the one-way scheme. Thus, these results suggest that one-way coupled analyses could provide geomechanical forecasts against the security.
Therefore, in a general way, the analyses performed considering the influence of a rigorous fluid-mechanical coupling in a blackoil production scenario proved to be fundamental when the aim is to determine geomechanical behavior. As aforementioned, one-way coupling can forecast results against the security, once the pore pressure, compaction and subsidence obtained were less than the iterative coupling. Furthermore, considering gas release, two-way analyses provided a larger gas volume when compared with one-way analyses, highlighting its importance for field operations.

- **Model B**

As presented for the previously model, Figure 8 shows the comparison between the conventional reservoir simulation (IMEX) and the simulation partially coupled ABAQUS-IMEX in two-way. It was evaluated the pore pressure distribution and the produced gas rate in reservoir conditions after 5000 days for injection and production. The flow rates (producer and injectors wells) are presented in the Table 1.

It can be observed that pore pressure behavior, for two methodologies, present some similarities, i.e., both curves present decrease in the beginning of the analyses and start to increase when the influence of injector wells reaches the producer well. Nevertheless, there is a quite difference in the gas generation behavior. The consideration of iterative fluid-mechanical coupling led to an increase in pore pressure, due injection, were not fast enough to generate gas in the reservoir, in contrast to conventional reservoir simulation, in which was observed gas release during around 2000 days. In other words, the effect of injection, as production presented previously, is affected strongly by consideration of the geomechanical effect, i.e., in the production scenario, the pore pressure decrease faster than one-way coupling and in the scenario when injection prevails, the increase of pore pressure also is observed, by the system, faster than the one-way coupled analyses.
The average pore pressure behavior within reservoir was evaluated considering one and two-way fluid-mechanical coupling schemes (see Figure 9). The pore pressure increases in a higher rate when the two-way scheme is compared to one-way coupling. It was expected once the injection rate is higher than production rate and the geomechanical effect has a fundamental importance in the behavior, as discussed in the former analyses.

As already performed to model A, Figure 10 shows a set outlook of pore pressure distribution considering one and two-way coupling methodologies. Again, it can be seen that in the iterative simulation, the variation of pore pressure was broadest than the simulation performed by one-way coupling. Besides that, the pore pressure gradients around the wells are higher in the two-way scheme than one-way scheme. Figure 10 is composed by a set of
images, which are: top view, cross sections through the producer/injector wells and tridimensional.

![Pore Pressure Variation](image1)

**Figure 10.** Outlook of pore pressure variation after 5000 days of production and injection considering one and two-way coupling scheme.

Since the effects of injection are prevalent in a global way, it was observed that curves decrease to negative values which indicate expansion in the reservoir and overburden (see Figure 11). It can be noted also, that in the beginning of the analyses, there is a small value of compaction. This fact is related to a time necessary to injection effect reach the measurement point, once this point was adopted in the location of producer well. As observed in model A, both results present important differences between methodologies, in which two-way coupling shows values of compaction (reservoir expansion) and subsidence (seafloor expansion) higher than the one-way scheme.

![Compaction and Subsidence](image2)

**Figure 11.** Compaction of reservoir and subsidence of seafloor for on and two-way coupling schemes – model B.

As concluded for a production scenario, the consideration of a rigorous fluid-mechanical coupling in reservoir simulation proved to be primordial when the aim is to predict...
geomechanical behavior. As already discussed, one-way coupling can forecast results against the security, once the pore pressure, compaction and subsidence obtained were less than the iterative coupling. Moreover, considering gas generation within reservoir, two-way analyses did not predict any gas release during 5000 days of field development, unlike the analyses presented by one-way coupling scheme, which predicted around 2000 days of gas release. Thus, the use of one-way coupling should be done carefully, once its results can have a quite significant impact in field activities.

6 CONCLUSIONS

After analyses presented in this work, some conclusions can be done in relation to the influence of fluid-mechanical coupling in gas generation in undersaturated petroleum reservoirs. It should be highlighted that the use of the strong coupling methodology, as the one developed by ATHENA/GTEP-PUC-Rio, has fundamental influence in a forecast of reservoir geomechanical behavior.

For a production scenario, it can be concluded that when the porous medium is filled by fluid (composed by a mixture of gas and oil), the influence of coupling methodology is not fundamental, once the production mechanism is ruled by the gas expansion within the reservoir. On the other hand, for scenarios that combines injection and production, the iterative coupling methodology presents more accurate results when compared with one-way coupling scheme. Furthermore, the choice of coupling methodology must be careful, mainly when the purpose is to predict the time and volume of gas released when the pore pressure drop below the bubble point pressure.

REFERENCES

LATTICE BOLTZMANN BASED MULTICOMPONENT REACTIVE TRANSPORT MODEL COUPLED WITH GEOCHEMICAL SOLVER FOR PORE SCALE SIMULATIONS

RAVI A. PATEL†,‡,*, JANEZ PERKO‡, DIEDERIK JACQUES‡, GEERT DE SCHUTTER‡, GUANG YE‡,#, AND KLAAS VAN BREUGEL#

† Belgium nuclear research center (SCK•CEN)
Mol, Belgium.
Email: rpatel@sckcen.be, Web page: http://www.sckcen.be

‡ Magnel lab for concrete research
Ghent University
Ghent, Belgium.
Email: rapatel@ugent.be, Web page: http://www.ugent.be/ea/labomagnel/en

# Microlab
Delft University of Technology
Delft, Netherlands.
Email: K.vanBreugel@tudelft.nl, Web page: http://www.microlab.citg.tudelft.nl/

Key words: Lattice Boltzmann Method, reactive transport, pore scale modelling, geochemistry, PHREEQC coupling

Abstract. A Lattice Boltzmann (LB) based reactive transport model intended to capture reactions and solid phase changes occurring at the pore scale is presented. The proposed approach uses LB method to compute multi component mass transport. The LB multi-component transport model is then coupled with the well-established geochemical reaction code PHREEQC which solves for thermodynamic equilibrium in mixed aqueous-solid phase system with homogenous and heterogeneous reactions. This coupling enables us to update solid phases volumes based on dissolution or precipitation using static update rules which, on pore scale, affects the change of potentially pore network geometry. Unlike conventional approach, heterogeneous reactions are conceptualized as volumetric reactions by introducing additional source term in the fluid node next to solid node, and not as flux boundaries. To demonstrate the validity of this approach several examples are presented in this paper.

1 INTRODUCTION

Multicomponent reactive transport modelling is an important tool for analysing different applications involving coupled physical-chemical processes such as chemical degradation of cementitious materials, transport and sorption of radionuclides, environmental contaminant transport, bioremediation and geologic sequestration of carbon dioxide. Advancements in geochemical reaction modelling and reactive transport modelling have led to development of
several geochemical reaction modelling codes [1-4] and efficient reactive transport codes to simulate coupled continuum scale physical-chemical processes [5-7]. Continuum scale models do not account explicitly for heterogeneities present at pore scale; rather these effects are often lumped up in macroscopic parameters. However, reactions such as precipitation, dissolution or surface reactions occurs at mineral surfaces which in turn changes properties of porous media, affecting flow and transport at both pore and continuum scale [8]. Pore scale modelling can be used to resolve these heterogeneities at lower scale to understand the mechanism governing up scaling of parameters to continuum scale. A number of pore scale reactive transport approaches have been presented in recent years based on conventional computational fluid dynamics approaches [8], pore network models [9], smooth particle hydrodynamics [10], hybrid approaches [11], and Lattice Boltzmann (LB) methods [12-15]. LB methods are easy to implement, efficient and due to its inherent local computation easily parallelizable and scalable for computationally intensive applications. However application of LB methods has been restricted to reaction systems with simple kinetics and geochemistry [14, 15]. In this paper we present an approach to couple LB method with the well-established geochemical modelling tool PHREEQC [1] which allows LB based methods to be applied to variety of geochemical systems with different set of complexities. Additional mass given to or taken out of the aqueous phase due to dissolution and precipitation processes is modelled as a source term in this study, as opposed to the more commonly used boundary flux formulation which simplifies the coupling with external geochemical codes.

First, the standard LB method with Bhatnagar-Gross-Krook (BGK) collision operator [16] for multi-component solute transport is briefly described. For more detailed information, reader is referred to specialized textbooks [17-21]. Next, the principles of the coupling procedure with PHREEQC and rules for updating pore geometry are outlined. Finally, benchmark results and example cases are discussed.

2 LATTICE BOLTZMANN METHOD FOR MULTI COMPONENT MASS TRANSPORT

Multicomponent mass transport at pore scale mainly occurs through pore water and is governed by advective-diffusive transport equation given by equation (1), if the electro kinetic effects due to presence of charged species is not considered and if mass transport is considered as passive scalar i.e. feedback of change of concentration to change in density is neglected.

\[
\frac{\partial C_j}{\partial t} = -\nabla \cdot \vec{j}_j + R_j
\]

\[
\vec{j}_j = (-D_j \nabla C_j) + \vec{u} C_j
\]

where, \( C_j \) is the Concentration of \( j^{th} \) species \([N \ L^{-3}]\), \( \vec{j}_j \) is the flux for \( j^{th} \) species \([N \ L^{-2} \ T^{-1}]\), \( D_j \) is the isotropic diffusion tensor of \( j^{th} \) species in water \([L^2 \ T^{-1}]\), \( \vec{u} \) is the velocity vector field for fluid flow \([L \ T^{-1}]\) and \( R_j \) represents reaction term for \( j^{th} \) species \([N \ L^{-3} \ T^{-1}]\).

Further simplification is made by reducing the number of species to be transported to be equal to number of primary species [22] and by having same diffusion coefficient for all species [8], thus reducing the computational burden.

In LB methods, a discrete velocity Boltzmann equation is solved instead of solving...
equation (1) directly which is further discretised in space and time domain (equation (2)). The collision term of Boltzmann equation in equation (2), is simplified using BGK assumption [16] and the resulting LB-BGK method [23] is represented as

\[
g_i^j(\vec{r} + \vec{e}_i \Delta t, t + \Delta t) = g_i^j(\vec{r}, t) + \Omega_{BGK,j}^i(\vec{r}, t)
\]

\[
\Omega_{BGK,j}^i(\vec{r}, t) = \frac{\Delta t}{\tau} \left[ g_i^{eq,j}(\vec{r}, t) - g_i^j(\vec{r}, t) \right]
\]

where, \(\vec{r}\) is position vector, \(g_i^j\) represents the particle’s distribution function along lattice \(i^{th}\) direction for \(j^{th}\) species [N L^{-3}], \(\vec{e}_i\) is the velocity vector in \(i^{th}\) direction which depends on type of lattice [L T^{-1}], \(\Delta t\) is the time step [T], \(\tau\) is relaxation time [T], and \(g_i^{eq,j}\) is the particle’s equilibrium distribution function for \(j^{th}\) species [N L^{-3}] given by equation (3).

\[
g_i^{eq,j}(\vec{r}, t) = w_i C_i \left[ 1 + \frac{\vec{e}_i \cdot \vec{u}}{e_s^2} \right]
\]

where, \(w_i\) are the weights for particle’s distribution function along \(i^{th}\) direction and \(e_s\) is the speed of sound on lattice [L T^{-1}].

Orthogonal lattices are isotropic enough to recover equation (1) [18] and one of the most commonly used lattices are \(D1Q3\) (equation (4)) for 1-D with 3 lattice directions and \(D2Q5\) (equation (5)) for 2-D with 5 lattice directions as shown in Figure 1.

**D1Q3**,\n
\[
\begin{align*}
\vec{e}_i & = 0 & i = 0 \\
\vec{e}_i & = \left( \cos \left( \frac{(i-1)\pi}{2} \right) \right) e & i = 1,3 
\end{align*}
\]

**D2Q5**,\n
\[
\begin{align*}
\vec{e}_i & = 0 & i = 0 \\
\vec{e}_i & = \left( \cos \left( \frac{(i-1)\pi}{2} \right), \sin \left( \frac{(i-1)\pi}{2} \right) \right) e & i = 1,2,3,4 
\end{align*}
\]

where \(e = \Delta x / \Delta t\) [L T^{-1}] and \(\Delta x\) is the distance between two lattice nodes [L]. For these lattices, weights and speed of sound for equation (3) are given by \(w_i = \frac{4}{6}, \frac{1}{6}\) and \(\frac{6}{6}\) for \(i = 0-3\) in case of \(D1Q3\) and \(w_i = \frac{4}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\) and \(\frac{6}{6}\) for \(i = 0-4\) for \(D2Q5\). Speed of sound on lattice is \(e_s = e / \sqrt{3}\) for both cases.

It can be shown that equation (2) together with the particle’s equilibrium distribution function given by equation (3), and lattices described by equation (4) or equation (5) it is possible to recover equation (1) using multiscale Chapman-Enskog expansion [21, 24]. Moreover, from multiscale Chapman-Enskog expansion, the relation between diffusion coefficient and relaxation time for \(D1Q3\) and \(D2Q5\) lattice is given as

\[
D = e_s^2 \left( \tau - \frac{\Delta t}{2} \right)
\]
The macroscopic quantities such as concentration \( (C_j) \) and flux \( (j'_p) \) in terms of particle’s distribution functions is given as

\[
C_j = \sum_i g'_i \cdot j'_i = \sum_i g'_i \vec{u} - (1 - \frac{\Delta t}{2\tau}) \left[ \sum_i g'_i \vec{u} - \sum_i g'_i \vec{v}_i \right] \tag{7}
\]

Reaction term of equation (1) is obtained in LB model by addition of an additional collision term [25-27] in equation (2) as follows

\[
g'_i(\vec{r} + \vec{v}_i \Delta t, t + \Delta t) = g'_i(\vec{r}, t) + \Omega_{BGK,j}(\vec{r}, t) + \Omega_{\text{reaction},j}(\vec{r}, t) \tag{8}
\]

\[
\Omega_{\text{reaction},j}(\vec{r}, t) = \Delta t \cdot w_i \cdot R_j
\]

3 COUPLING LATTICE BOLTZMANN MASS TRANSPORT SOLVER WITH PHREEQC

Previously, LB method has been used to solve reactive mass transport equations and mimic dissolution and precipitation reactions. A short review on developments of LB for reactive transport modelling has been made by Kang et al. [28]. However, application of LB methods has been restricted to predefined simple chemical systems [14, 15] and recently further developed for incorporating exchange reactions [29]. Furthermore, existing approaches of applying heterogeneous reaction in LB method does not allow use of external geochemical reaction codes directly thus restricting application of developed solver to predefined chemical systems. In this section we show methodology to couple a LB mass transport solver with a generic geochemical solver viz., PHREEQC, to overcome this barrier.

PHREEQC is a widely used geochemical reaction modelling code with capabilities to seek thermodynamic equilibrium for solution speciation and redox reactions, mineral dissolution and precipitation reactions, ion exchange reactions, surface complexation based on diffused double layer model, gas phase exchange, (non)ideal solid solutions or mixed-equilibrium reaction networks with user-defined rate equations depending on solution speciation, temperature or moisture content [1].

An RD time splitting approach [30] for equation (8) is used to couple reactions with LB which is equivalent to Sequential Non-Iterative Approach used for continuum scale finite
element/finite difference numerical methods [22], e.g. [31] when coupling with PHREEQC. The LB advective-diffusive transport solver written in MATLAB in this study calls the recently released PHREEQC COM version of PHREEQC-3 [32, 1] to obtain the reaction collision term of equation (8). Similar to the approach of Wissmeier & Barry [33], a LB model is set up by assigning initial and boundary conditions through PHREEQC speciation calculations at the beginning of simulation. After setting up the model, a typical calculation step consists of

a.) Executing transport step as in equation (2)

b.) Transferring concentrations at the end of transport time step \( t + \Delta t \) to PHREEQC

c.) Executing reaction calculation using PHREEQC and obtain new concentrations at the end of reaction time step \( t + \Delta t' \)

d.) Correcting the particle’s distribution function using equation (9) for time step \( t + \Delta t \)

\[
g^{j}_t(\vec{r}, t + \Delta t) = g^{j}_{transport}(\vec{r}, t + \Delta t') + \Omega^{reaction,j}(\vec{r}, t + \Delta t'')
\]

where, \( C^{j}_{phreeqc} \) is the concentration for \( j^{th} \) species obtained from PHREEQC after executing reaction step and \( C^{j}_{transport} \) is the concentration for \( j^{th} \) species obtained at the end of transport step.

Unlike, previous approaches where heterogeneous reactions were implemented using flux boundary, we implemented heterogeneous reactions as an additional reaction term in the fluid node next to solid node which can be expressed using equation (1), thus representing heterogeneous reaction as a volumetric reaction instead of flux boundary [8]. Hence, a fluid node next to solid consists of both homogenous and heterogenous reactions and change in concentration due these reactions is computed using PHREEQC. This approach thus ensures complete decoupling of LB calculations and reaction calculations allowing use of existing geochemical reaction modelling codes to execute reaction step.

To update pore geometry, static update rules in essence similar to the one proposed by Kang et al. [28] are implemented. Information of the amount of moles of dissolved or precipitated solid phase at each time step is obtained from PHREEQC based on which the volume occupied by a single mineral can be calculated as

\[
V_m(t + \Delta t) = V_m(t) + \bar{V}_m[N_m(t + \Delta t) - N_m(t)]
\]

where, \( V_m \) is the volume occupied by a solid mineral phase \( m \) [L^3], \( \bar{V}_m \) is the molar volume for a solid mineral phase \( m \) [Na L^3], and \( N_m \) represents number of moles of solid mineral phase present at a node [N]. Hence, to allow the presence of multiple mineral phases at a given node, the total amount of volume occupied by the solid phase, \( V_{total} \), is the sum of the volumes occupied by all mineral phases at that node given by

\[
V_{total}(t + \Delta t) = \sum_m V_m(t + \Delta t)
\]

When the ratio of \( V_{total} \) to the effective volume (effective volume can be defined as the
maximum volume that can be occupied by solid mineral phase at a given node, $V_{\text{effective}}$ reaches a threshold value the pore geometry is updated. For dissolution track of solid volume is kept at solid node and when $\varnothing = V_{\text{total}}/V_{\text{effective}} \leq 0.5$, the corresponding solid node is converted into a fluid node. For precipitation track of solid volume is kept at fluid node and when $\varnothing = V_{\text{total}}/V_{\text{effective}} \geq 0.5$ the corresponding fluid node is converted into a solid node.

4 BENCHMARK CASES AND EXAMPLES

In this section, the following benchmarks and example are presented: (i) cation exchange example for verification of the coupling between the LB and a generic geochemical solver with an alternative code, (ii) verification of solid boundary update of LB code with analytical solution by simulating diffusion controlled dissolution and (iii) a hypothetical example of portlandite dissolution with geometry update to demonstrate the applicability of develop code.

4.1 Example of cation exchange

The developed code is applied to model advective-diffusive transport in the presence of a cation exchanger to demonstrate coupling of the LB method with PHREEQC. This benchmark aims at verifying the formulation defined in section 3 for a porous media at the continuum scale, i.e. solving equation (1). The chemical composition of effluent from a column containing a cation exchanger is simulated and the result of LB method is compared with COMSOL Multiphysics (this example is based on example 11 described in [1]). The model setup consists of an 8 cm long column containing initially 1mM of NaNO$_3$ solution and 0.2 mM of KNO$_3$ solution in equilibrium with 1.1 mM of cation exchanger. The column is then injected with a 0.6 mM of CaCl$_2$ solution. At each time step, the exchanger is in equilibrium with Na, K and Ca. The water flow velocity in domain is assumed to be $2.77 \times 10^{-6}$ m/s and the diffusion coefficient is $5.54 \times 10^{-9}$ m$^2$/s. At inlet Cauchy boundary equal to the product of fixed boundary concentration and velocity is applied using equation (7) and the outlet is an open boundary. This example was found to be sensitive to the type of boundary condition applied, due to the presence of exchange reaction and serves as a good benchmark example. It should be noted that the formulation of flux boundary condition in terms of particle’s distribution function given by Verhaeghe et al. [34] represents only the diffusive part of flux [30] whereas, equation (7) represents both diffusive and advective components of flux and thus offers more generic formulation for application of Cauchy boundary condition. Figure 2 shows a comparison between the LB results and COMSOL Multiphysics and indicates an excellent agreement.
4.2 Diffusion controlled dissolution

A Stefan problem for diffusion controlled dissolution [35] is used to validate the geometry update during dissolution. Comparison of LB model is made with analytical given by Aaron et al. [35]. The model setup consists of a 20 mm long domain with species concentration of 0.1 mol/m³ with last 3 mm domain containing solid with initial mass of 1 mol/m³. Hence, the fluid-solid interface, \( x_0 \) is initially located at 17 mm from the inlet. An equilibrium concentration of 0.4 mol/m³ is maintained at the fluid-solid interface. The diffusion coefficient is taken as 1 x 10⁻⁹ m²/s. Figure 3(a) shows a comparison of the LB results with analytical solution for the following two dissolution rules: (i) when a solid node contains less than 50% solid (\( \emptyset < 0.5 \)), it is converted to fluid node and (ii) when a solid node contains no solid (\( \emptyset = 0 \)), it is converted to fluid node. It can be seen that the rate of movement of boundary is slower for \( \emptyset = 0 \) and \( \emptyset \leq 0.5 \) better represents the movement of fluid solid interface. In Figure (3), the movement of interface is in discrete steps as update of boundary is carried out in a static way and as shown in Figure 3(b) that for \( \emptyset \leq 0.5 \), with higher number of nodes more continuous movement of fluid solid interface. However, the movement of fluid solid interface is fairly independent of number of nodes and good agreement with analytical solution is observed for all discretization.
4.3 Dissolution of portlandite

Portlandite is an important mineral phase in hardened cement paste and dissolution of portlandite due to calcium leaching has adverse effects on transport and mechanical properties of cement paste [36]. In this example, applicability of the developed code for handling geochemical reactions along with geometry update to mimic the dissolution of portlandite is demonstrated. The model setup consist of 8 cm long domain with the first 2 cm of domain consisting of water in equilibrium with portlandite (Ca ≈ 20 mM, pH ≈ 12.5) and the last 6 cm representing solid portlandite with initial mass of 4 mM. An aggressive water solution with pH 3 (obtained by adjusting Cl ion concentration to maintain charge balance) is present at the inlet boundary. At the fluid-solid interface, portlandite keeps the water in equilibrium by dissolving excess portlandite at a given time step. Diffusion coefficient for all species is assumed to be $1 \times 10^{-9}$ m$^2$/s. Fixed concentration boundary is applied at the inlet and no flux boundary is imposed at outlet boundary on complete dissolution of solid.

Results of simulation are shown in figure 4. Figure 4(a) shows that when water with low pH is introduced from inlet boundary Ca ions gets flushed out from the water in the domain, causing decrease in Ca ion concentration in water. Once, the cumulative concentration of Ca in water drops to around 1.08 mM, dissolution of portlandite begins causing increase of Ca ion concentration in water till all portlandite is dissolved from the system. After complete dissolution of portlandite, a gradual drop in the Ca concentration in the water is observed due to the outward diffusion of Ca. Figure 4(b) shows the movement of interface with $\sqrt{t}$ which is linear as would be expected for diffusion controlled dissolution.
Figure 4: Diffusion controlled dissolution
5. CONCLUSIONS

In this paper we outlined the development of a LB–BGK model for pore scale reactive transport problems involving complex chemical systems by coupling LB method with PHREEQC. The LB mass transport step and reaction step were separated using a RD operator splitting approach and PHREEQC was used to solve both homogenous and heterogeneous reactions. The developed code was tested with a series of benchmarks to test the coupling of LB with PHREEQC and accuracy of geometry update. Finally, the ability of the code to handle geochemical reactions along with geometry update to mimic dissolution of portlandite was demonstrated. All the cases were restricted to 1-D but the examples and benchmarks show the ability of the developed approach to handle complex geochemical reactions and effective updating of pore geometry caused by dissolution and precipitation reactions.

In future, the developed code will be applied to 2D and 3D problems and further coupled with LB –BGK Navier stokes solver to handle coupled flow and reactive transport problems. Finally, the coupled model will be applied to study the evolution of cement microstructure due to cement degradation mechanisms such as carbonation and leaching.

6. ACKNOWLEDGEMENTS

First author would like to thank Suresh Seetharam and Joan Govaerts of Belgian Nuclear research center (SCK•CEN) for discussions, help on coupling PHREEQC and developing benchmark example with COMSOL Multiphysics. This work is financially supported by SCK•CEN and their support is greatly acknowledged.

7. REFERENCES


ON NUMERICAL MODELING OF COUPLE HEAT, AIR AND MOISTURE TRANSFER THROUGH MULTILAYERED WALLS

M. MALIKI*, N. LAREDJ*, H. NAJI†, ††, K. BENDANI*, H. MISSOUM* AND K. MENAD*

* Construction, Transport and Protection of the Environment laboratory (LCTPE)
Abdelhamid Ibn Badis University
Belhacel Road, 27000 Mostaganem, Algeria
e-mail: mus27000@yahoo.fr, http://www.lctpe-univ.mosta.dz

† Université Lille Nord de France
F-59000 Lille, France
e-mail: Hassan.Naji@polytech-lille.fr, http://www.univ-lille-nord-de-france.fr

†† Laboratoire Génie Civi & géo-Environnement (LGCgE- EA 4515)
UArtois/FSA Béthune, F-62400 Béthune, France
e-mail: hassane.naji@univ-artois.fr, http://www.lgcge.fr

Key words: Coupled transfer, heat, air, moisture, multi-layer wall, porous media, simulation.

Abstract. This paper reports on numerical modeling of heat, air, and moisture transfer through multilayered walls. Building materials are often subjected to temporal climatic variations, which can induce a transfer of heat and moisture through the walls of the building and the foundation soil. These materials are generally considered as porous media. The coupled heat, air and moisture transfer in building materials is of paramount importance in the construction area. In this way, a mathematical model has been elaborated and validated using a benchmark example. Here, we aim to determine the energy losses. The capillary pressure is considered as potential moisture which represents both the transport of vapor and liquid phases of the water. Basing on basic functions of partial differential equations, one can convert certain measurable properties of porous media as coefficients depending on the temperature and the capillary pressure. The results obtained compare favorably with other available in the literature.

1 INTRODUCTION

Heat, air and moisture transfer through porous media is explored in many engineering areas such as pollutants infiltration, granular materials drying, heat exchangers and thermal insulation among others. Building envelopes are constantly subjected to random climatic loads on the outer surface and relatively stable conditions on its inner face. These loads generate a transport of heat, air and moisture through the structure. The flow's direction of these entities depends on the gradient of the potential director of each entity. It should be noted that all porous media, in which the combined heat, air and moisture transport occurs simultaneously and strongly coupled, does have to be taken into account. In addition to these
temporal variations of external load due, thermal characteristics and those related to storage moisture's layers of the wall should be added. This combination makes the transport of heat and moisture transient and relatively complex in the building envelope. Despite this complexity, we can numerically simulate the dynamic processes drying-moistening of the envelope component. In this context, hygrothermal models have been used to evaluate the performance of a wall exposed to the weather in different geographical sites [1, 2, 3].

To carry out this study, a model has been developed to describe the transfer of heat, air and moisture in building envelopes. The model developed here belongs to the class of models that have recently emerged to handle the coupled transfer mentioned above. These are implemented using commercial softwares to solve problems inherent to buildings physics [4, 5].

It is advisable to note that these models are more flexible for future extensions (e.g. for 2D and/or 3D simulations), because allowing the addition of new features, as well as an easy integration with other existing models. One advantage of having a transient heat, air and moisture (HAM) model as a whole building hygrothermal model is that it enables to capture the potential moisture release from the building enclosure to the indoor space. In fact, moisture sources from construction and from wet soil through foundation walls and floor slab could dominate all internal moisture sources. Similarly, the importance of quantification of the moisture release from foundation slabs when calculating indoor humidity levels was recently emphasized.

The use of a transient HAM model when conducting whole building performance analysis yields also to a better estimation of energy demand for heating or cooling of a building. This is possible due to the fact that a transient HAM model takes into account the effect of moisture in the heat transfer through building enclosures. Usually energy simulation models ignore the moisture effect when conducting the thermal analysis [6], and use constant thermal storage and transport (thermal conductivity and heat capacity, respectively) property values despite the fact that these properties can be strongly dependent on moisture content.

The aim of this work is towards the numerical simulation of heat, air and moisture transfer in building systems. The numerical model of combined heat and moisture transfer based on basic functions of partial differential equations is tested in this work leaning on a benchmark case.

The remaining of this paper is organized as follows: in the next section, description of the physical problem and the mathematical model herein used are supplied. Section 3 is devoted to boundary conditions. Subsequently, in section 4, we present the model validation and we discuss the obtained results. Finally, section 5 outlines conclusions drawn from the present simulations.

2 MATHEMATICAL MODELING

The mathematical model for heat and moisture transfer, to be implemented for building materials, is described in this section.

It is worth noting that, in the coupled transfer, the moisture transport in building materials appears under two different phases: liquid and vapor. The phase vapor is divided into diffusion and convection parts. Indeed, the diffusive flow of vapor is engendered by vapor
pressure gradient and the corresponding conductivity represents the permeability of the vapor. As for, the convective vapor flow (the vapor flow), it is advected by the moving air [3].

The modeling of the transfer in vapor phase by the gradient of capillary pressure as conductive potential, and the permeability of the liquid as the conductivity of moisture transfer became the most appropriate approach and the most used one in this kind of modeling [7]. For the liquid flow, moisture content gradient has been used as the driving potential in some hygrothermal tools, and moisture diffusivity was used as the moisture transfer conductivity.

According to the principle of the preservation of the combined transport of heat and humidity of a representative elementary volume (REV), which is defined as being large enough when compared to pore dimensions but small enough compared to the size of the sample, governing equations of the coupled transfer in building materials can be formulated. Below, the balance equations for moisture and heat transport are described.

2.1 Moisture transport

The moisture content being the mainspring of transfer, its distribution can be expressed by the following equation:

\[
\frac{\partial w}{\partial t} = \nabla \left( \delta_p \nabla p_v - K_l \nabla P_c \right) - \nu \cdot \nabla \rho_v + F_m
\]  

(1)

where \( w \) (kg/m\(^3\)) is the moisture content, \( t \) (s) is the time, \( \delta_p \) (kg/m.s.Pa) is the water vapor permeability, \( p_v \) (Pa) is the partial water vapor pressure, \( K_l \) (s) is the liquid water permeability, \( P_c \) (Pa) is the capillary pressure, \( \nu \) (m/s) is the air velocity, \( \rho_v \) (kg/m\(^3\)) is the water vapor density and \( F_m \) is the moisture source term.

2.2 Heat transport

Here, we conjecture that the main mechanisms which govern the transfer of heat are the thermal conduction and the convection due to air movement and latent heat. This is due to the presence of low temperature gradients.

The the energy conservation equation can be written as

\[
\left( c_p \rho + c_p w \right) \frac{\partial T}{\partial t} = \nabla \left( \lambda \nabla T \right) + L_v \nabla \left( \delta_p \nabla p_v \right) - \nu L_v \cdot \nabla \rho_v - \nu \rho_a c_{p,a} \nabla T + F_h
\]  

(2)

where \( c_p \) (J/kg.K) is the dry specific heat of material, \( \rho \) (kg/m\(^3\)) is the dry density of the material, \( c_p,l \) (J/kg.K) is the specific heat of liquid water, \( T \) (K) is the temperature, \( \lambda \) (W/m.K) is the thermal conductivity, \( L_v \) (J/kg) is the enthalpy of evaporation, \( \rho_a \) (kg/m\(^3\)) is the dry air density, \( c_{p,a} \) (J/kg.K) is the specific heat of dry air and \( F_h \) (W/m\(^3\)) is the heat source term.
2.3 Conservation equations and modeling

In this approach, we transformed variables depending on moisture (see Eqs. 1-2) into a single variable called capillary pressure, denoted $P_c$.

The relation between the partial water vapor pressure and the relative humidity can be expressed as:

$$p_v = \phi P_{sat}$$  

where $\phi$ is relative humidity, and $P_{sat}$ (Pa) is the saturated water vapor pressure.

The relation between relative humidity and capillary pressure is given by Kelvin’s law:

$$\phi = \exp \left( \frac{-P_c}{\rho_i R_v T} \right)$$

where $\rho_i$ (kg/m$^3$) is the water density and $R_v$ (J/kg.K) is the gas constant for water vapour.

It ensues from this that the conservation equations of the combined heat and moisture transfer can be rewritten in term of coefficients, and by considering temperature as independent variable for heat transfer and capillary pressure as independent variable for moisture transfer, as shown in Eq. (5) and (6).

$$C_T \frac{\partial T}{\partial t} = \nabla \left( C_{11} \nabla T + C_{12} \nabla P_c \right) + \nu \left( D_{11} \nabla T + D_{12} \nabla P_c \right) + F_h$$

$$\delta \frac{\partial P_c}{\partial t} = \nabla \left( C_{21} \nabla T + C_{22} \nabla P_c \right) + \nu \left( D_{21} \nabla T + D_{22} \nabla P_c \right) + F_m$$

with $C_T = c_p \rho + c_{pi} \rho_i$; $\Omega = \partial w / \partial P_c$, and $\Omega$ is the moisture storage capacity, defined as the slope of water retention curve.

It is worth noting that Eqs. (5) and (6) can be written in the following matrix form:

$$\frac{\partial T}{\partial t} = \nabla \left( C \nabla T \right) + \beta \nabla \left( P_c \right) + \begin{bmatrix} F_h \\ F_m \end{bmatrix}$$

where damping ($d_a$), diffusion ($C$) and convection ($\beta$) matrices are respectively defined by:

$$d_a = \begin{bmatrix} C_T & 0 \\ 0 & \Omega \end{bmatrix}$$
where \( \frac{P'_s}{\partial T} \) is the derivative of saturation vapor pressure.

\[
\beta = \begin{bmatrix}
D_{11} & D_{12} \\
D_{21} & D_{22}
\end{bmatrix} = \begin{bmatrix}
-\left( \rho_\phi \frac{\partial \rho_v}{\partial T} + L_v \frac{\partial \rho_v}{\partial T} \right) & -\frac{L_v \phi}{\rho_v R_v T} \frac{\partial \rho_v}{\partial \phi} \\
\frac{\partial \rho_v}{\partial T} & -\left( \frac{\phi}{\rho_v R_v T} \frac{\partial \rho_v}{\partial \phi} \right)
\end{bmatrix}
\]

where \( F_h \) and \( F_m \) represent respectively, heat and moisture source.

To go further, the current model consists of converting, via MatLab, the measurable physical properties of the material such as \( K, \phi, \delta_p \) and \( \lambda \) which depend on moisture content \( w \) into partial differential equations (PDE) \( C_{11}, C_{12}, D_{11}, D_{12}, \Omega \) and \( C_T \) which are dependent on \( P_c \) and \( T \) [8]. This is schematically depicted in Fig. 1.

**Figure 1:** Conversion of measurable material properties

### 3 BOUNDARY CONDITIONS (BCS)

External boundary conditions of building envelopes can be grouped in three categories [9]: moisture saturation, constant heat and moisture flow, and heat/moisture flow through surface resistance film fixed on external surface. Eqs. (11) and (12) describe the exterior boundary conditions. For the internal surface of the wall, the temperature and pressure are maintained constant.

#### 3.1 Moisture BCs

The moisture flux through the outside surface of the envelope, \( g_{w,e} \) (kg/m².s), is based on the following relationship:
\[ g_{n,e} = \beta_{p,e} (p_{v,e} - p_{surf,e}) \]  

where \( \beta_{p,e} \) (kg/m².s.Pa) is the vapor transfer coefficient of the exterior surface, \( p_{v,e} \) (Pa) is the water vapor pressure of the outdoor air, and \( p_{surf,e} \) (Pa) is the water vapor pressure on the exterior surface.

For the external side of the wall, the moisture flux is obtained according to the following relationship:

\[ g_{n,i} = \beta_{p,i} (p_{v,i} - p_{surf,i}) \]  

where \( \beta_{p,i} \) (kg/m².s.Pa) is the vapor transfer coefficient of the interior surface, \( p_{v,i} \) (Pa) is the water vapor pressure of the indoor air and \( p_{surf,i} \) (Pa) is the water vapor pressure of the interior surface.

### 3.2 Heat BC

Recall that the heat flux through external surface, \( q_{n,e} \) (W/m²), includes conductive, convective and latent heat effects only. It can be expressed as:

\[ q_{n,e} = \alpha_e (T^{eq} - T_{surf,e}) + L_v \beta_{p,e} (p_{v,e} - p_{surf,e}) \]  

where \( \alpha_e \) (W/m².K) is the heat transfer coefficient of the exterior surface, \( T^{eq} \) (K) is the equivalent exterior temperature and \( T_{surf,e} \) (K) is the temperature of the exterior surface.

Heat transfer through internal surface of the building envelope, \( q_{n,i} \) (W/m²), is given by:

\[ q_{n,i} = \alpha_i (T_i - T_{surf,i}) + L_v \beta_{p,i} (p_{v,i} - p_{surf,i}) \]  

where \( \alpha_i \) (W/m².K) is the heat transfer coefficient of the interior surface, \( T_i \) (K) is the temperature of the indoor air, and \( T_{surf,i} \) (K) is the temperature of the interior surface.

### 4 VALIDATION OF THE HEAT AND MOISTURE MODEL

As was stated above, the working of the aforementioned model equations has been checked considering a benchmark case. The latter arises from a series of five benchmark cases from a work outcome of the UE-initiated project for standardization of heat, air and moisture calculation methods (HAMSTAD WP2) [10,11].

The benchmark considered here deals with interstitial condensation occurring at the contact surface between two materials. The construction, from lowest x-coordinate (external side) to the highest, is built up as follows; vapor tight seal, 100 mm load bearing material and 50 mm thermal insulation as shown in fig. 2. The materials have different thermal and moisture properties: the load bearing material is capillary active, while the insulation is hygroscopic but capillary non-active (infinite resistance to liquid flow), and thermal conductivities differ by a factor 50 (at dry conditions). The structure is perfectly airtight.
4.1 Initial conditions

The following initial conditions were adopted:

- For load bearing material: \( w = 145 \text{ kg/m}^3, T = 10 \text{ C} \)
- For insulation: \( w = 0.065 \text{ kg/m}^3, T = 10 \text{ C} \)

4.2 Boundary conditions

The boundary conditions corresponding to the considered problem are as follows:

- For heat and moisture, a data file supplies the hourly values for a period lasting over one year. For intermediate values of time, they are obtained by interpolation.
- The outside equivalent temperatures encompass both the temperatures of the ambient air and that of the radiation.
- No difference in pressure is considered.

The surface transfer coefficients are given by:

\[
\alpha_{v,e} = 25 \text{ W/m}^2K, \alpha_{e,i} = 7 \text{ W/m}^2K, \beta_{p,e} = 0 \text{ s/m}, \beta_{p,i} = 2 \times 10^{-8} \text{ s/m} 
\]

Note that these conditions allow a very good case for checking the heat and moisture transfer model.

The numerical simulation was carried out using the multi-physics software COMSOL V3.5 [12], and the required results are the following ones:

- Capillary pressure \( P_c \) in space and time for load bearing element A, and insulation B (see Fig. 2).
- Temperature \( T(x,t) \).
- Total moisture weight \( M \) (kg/m²) in each layer.
- Heat flux \( q \) (W/m²) crossing the structure from interior to the wall.

The model simulates, for over a period of one year, the distributions of the capillary pressure \( P_c \) and the temperature \( T \).

The model simulates for over a period of one year, the distributions of the capillary pressure \( P_c \) and the temperature \( T \). In Fig. 3, the temperature distribution in load bearing element and insulation is depicted.
4.3 Results and discussion

- Moisture transport

Using properties of the material, the total moisture content can be easily computed from the distribution of capillary pressure \( P_c \). In Fig. 4, the total moisture content in load bearing during the first year is presented.

It can be seen from this figure that the predicted profile generally corroborates over the considered period with a maximum relative difference of 1.65%. This allows us to state that the current model is able to reproduce available results in literature.
• **Heat transport**

Figure 5 displays the heat flow crossing the structure from the inside during the first 500 hours obtained with the current model in comparison with HAMSTAD test. From this figure, we observe that our prediction is very close to that of the test case. Moreover, some peaks appear suggesting a possible numerical instability. This is going to be elucidated soon.

![Figure 5: Heat flux from interior to the wall](image)

### 5 CONCLUSION

A mathematical model considering the coupled heat, air and moisture transport through porous media has been elaborated and validated using a benchmark case. It has been designed to handle the coupling between processes of transfer of heat and moisture in the building envelopes. A hygrothermal simulation of a porous medium has been carried out to see the performance of the heat and moisture model. The commercial code COMSOL was used to solve the governing equations of HAM transport, which provides the flexibility for building science researchers to operate and modify the presented model. The numerical results show good agreement with available data.

Finally, the proposed model can be readily extended to predict the hygrothermal behavior of building envelopes, and the relevant work will be reported in the future.

### 6 REFERENCES


COUPLED ELECTRO-ELASTICITY AT FINITE STRAINS

CARLO SANSOUR∗, SEBASTIAN SKATULLA† AND MOHAMED HIJIAJ∗†

∗Senior Mary Curie Fellow, INSA de Rennes, France, on leave from Department of Civil Engineering, The University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom e-mail: carlo.sansour@nottingham.ac.uk, web page: http://www.nottingham.ac.uk/

†CERECAM, Department of Civil Engineering, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa e-mail: sebastian.skatulla@uct.ac.za - Web page: http://www.uct.ac.za

∗† Department of Civil Engineering, INSA de Rennes, France e-mail: mohammed.hjiaj@insa-rennes.fr - Web page: http://www.insa-rennes.fr

Key words: electromechanical coupling, electro polymers, multiplicative decomposition, meshfree methods

Abstract. In this paper we propose a multiplicative split of the deformation gradient into a part related to the elastic behaviour of the material and further one which describes the deformation induced by the electric field. Already available and well tested functions of elastic free energy functions can be immediately deployed without any modifications provided the argument of the function is considered to be the deformation tensor which is defined by the elastic part of the deformation gradient only. An appropriate constitutive relation is formulated for the electrically induced part of the deformation gradient. The approach is elegant, straightforward and above all, provides clear physical insight. A numerical example of highly non-linear coupled deformations demonstrate the potential and strength of the theory.

1 INTRODUCTION

Electroactive polymers (EAP) have been discovered to be very useful, because, in contrast to piezoelectric ceramics, they are less critical with regard to deformability and formability. The field of application as actuators, sensors and energy harvesting devices shows a broad versatility ranging from bio- and micro-manipulation, biomimetic robotics, prosthetics, and smart structures [1, 2].

A model was developed in [3] for an application to electric-sensitive hydrogels which is called the refined multi-effect-coupling electric-stimulus (rMECe) model, wherein the
fixed charge density and finite deformation were considered. An overview about Ionic polymer-metal composites (IPMCs) was given by [4]. An experimental study on dielectric elastomers undertaken by [5] demonstrated that the homogeneous deformation of a layer of a dielectric elastomer subjected to a voltage could be unstable, giving way to an inhomogeneous deformation, such that two regions coexisted within the layer. In [6] an electrical measurement technique was presented for obtaining information on the transient strain in the actuator and analyze the behavior of the actuator in safe and failure operation regimes. Similarly, another electromechanical loading experiments were conducted that captured the large deformation dynamic behavior of axisymmetric dielectric elastomer membranes subjected to dynamic electrical loading and dynamic mechanical loading experiments [7]. A model for a biaxially pre-strained circular actuator was proposed to characterize the electromechanical behavior of a dielectric elastomers [8]. In this approach, the deformation of the actuator for a given activation voltage depended on the three-dimensional mechanical behavior of the film.

The general non-linear electroelasticity equations were developed by [9] and [10]. Constitutive relations of non-linear electroelastic material were formulated by [11] as well as [12]. [13] formulated the variational approach for realistic modelling of EAP, and provided the finite element implementation details. An extension of the theories of electroactive polymers into the realm of generalised continua was recently achieved in [14]. Such continua naturally provide scale effects.

By the very fact that large elastic deformations can be sustained, electroactive polymers lends themselves to frameworks based on the formulation of a free energy function. One of the pressing issues of many of such formulations relates to the adequate consideration of coupling terms, where the deformation tensor, usually considered to be the right Cauchy-Green tensor $C$, and the electric field $e$ become intertwined. Existing formulations, however, lack transparency. The coupling terms are chosen to be very simple and no real guidance is provided as to the appropriate choice of such terms. To overcome these difficulties the paper embarks on a new formulation based on the idea of splitting the deformation gradient $F$ in a multiplicative fashion into an elastic (mechanically induced) part and a part induced by the electric field: $F = F_{mech}F_{elec}$. The advantages of such a decomposition become immediately apparent: 1) Existing well established and tested formulations of free energy functions for elastic polymers can be directly adopted provided the function is formulated in terms of deformation tensors only based on $F_{mech}$. 2) The coupling is directly given by an appropriate constitutive law for $F_{elec}$ in terms of the electric field. Such formulation becomes transparent, straightforward and can be extended to arbitrary non-linearities. 3) Physical insight regarding the nature of the constitutive law is immediately provided. We note that such a decomposition has been extensively considered in the theory of plastic deformations; see e.g. [15, 16, 17]. The nature of the constitutive law is, however, completely different since plasticity is a dissipative process where the inelastic part of the deformation gradient is defined by means of an integration process.
The paper is organized as follows: In Sec. 2 the basics of electrostatics is presented followed by introducing the multiplicative electromechanical coupling theory. Finally, a numerical example illustrates the capabilities of the formulation.

2 ELECTROMECHANICAL COUPLING

A summary of the standard equations of electromechanical coupling is given in this section. For further details the reader is also referred to [18, 19].

Let $\mathbf{F}$ be the deformation gradient. As a deformation measure we make use of the right Cauchy-Green deformation tensor $\mathbf{C}$ defined by

$$\mathbf{C} = \mathbf{F}^T \mathbf{F}. \tag{1}$$

In the current configuration the electric field is given by

$$\mathbf{e}_t = -\nabla \phi, \tag{2}$$

where $\phi$ denotes the electric potential. For dielectric material the electric displacement is expressed as

$$\mathbf{d}_t = \epsilon_0 \mathbf{e}_t + \mathbf{p}_t, \tag{3}$$

where $\epsilon_0$ denotes the \textit{vacuum electric permittivity} and $\mathbf{p}_t$ the dielectric polarization or simply the polarization. The polarization represents a derived quantity, as it depends on the electric field and on the material and its state which includes density, temperature and strain. The latter essentially means that not only does a dielectric material deform when subjected to an externally induced electric field, but also in turn does the deformation influence the electric field. This is clear, as the polarization generates its own electric field which contributes to the total electric field $\mathbf{e}_t$. In case of linear dielectrics, i.e. the material exhibits a linear response to an applied electric field, the electric displacement (Eq. 3) can be expressed as

$$\mathbf{d}_t = \epsilon_0 (1 + \chi_e) \mathbf{e}_t = \epsilon \mathbf{e}_t, \tag{4}$$

where the constant $\chi_e$ denotes the \textit{electric susceptibility} and $\epsilon$ the material’s permittivity. \textit{In vacuo} the electric displacement is reduced to $\mathbf{d}_t = \epsilon_0 \mathbf{e}_t$.

In electrodynamics, electric and magnetic fields are governed by \textit{Maxwell’s} equations. Here, however, we want to focus on electrostatics and do not consider motion and changes in time. Then, if magnetic fields, free electric currents and charges are not taken into account, only \textit{Gauss’s} law for electricity and \textit{Faraday’s} law of induction need to be considered:

$$\text{div} \ \mathbf{d}_t = 0 \quad \text{in } \mathcal{B}_t \quad \text{and} \quad \text{curl} \ \mathbf{e}_t = 0 \quad \text{in } \mathcal{B}_t, \tag{5a,b}$$

830
respectively, where the div and curl denote the divergence and curl operator in the current configuration respectively. The governing equations are supplemented with corresponding Neumann and Dirichlet boundary conditions which read as follows:

$$\mathbf{d}_t \cdot \mathbf{n}_t = -q_s^t \text{ on } \partial B_{t,N}^q,$$

(6)

and

$$\phi = h_\phi \text{ on } \partial B_{t,D}^\phi,$$

(7)

respectively. $q_s^t$ denotes the electric surface charge density in the deformed configuration, $\mathbf{n}$ the normal vector on $\partial B$ with $\mathbf{n}_t$ as its equivalent in the deformed configuration and $\partial B_{t,D}^\phi \subset \partial B$ and $\partial B_{t,N}^q = \partial B \setminus \partial B_{t,D}^\phi$.

For later use we also want to provide all needed equations and definitions in the Lagrangian description. For further details the reader is referred to [14]. The electrical field in the reference configuration $\mathbf{e}$ is related to the corresponding quantity in the current configuration $\mathbf{e}_t$ by

$$\mathbf{e} = \mathbf{F}^T \mathbf{e}_t$$

(8)

and the electric displacement by

$$\mathbf{d} = J \mathbf{F}^{-1} \mathbf{d}_t,$$

(9)

where $J = \det \mathbf{F}$ denotes the Jacobian with det being the determinant. In accordance with Eq. (3) the polarization in the undeformed configuration may be related to its equivalent in the deformed configuration via

$$\mathbf{p} = J \mathbf{F}^{-1} \mathbf{p}_t,$$

(10)

assuming that $\mathbf{p}_t$ is given by a constitutive equation. Note that these equations represent “pull back” operations. With Eqs. (3), (8) and (9) the electric displacement in the reference configuration takes

$$\mathbf{d} = \epsilon_0 J \mathbf{C}^{-1} \mathbf{e} + \mathbf{p}.$$

(11)

The electric field $\mathbf{e}$ with respect to the undeformed configuration is defined via Eqs. (2) and (8) by

$$\mathbf{e} = -\text{Grad} \phi.$$

(12)

Finally, from the governing equations of the electric field and the electric displacement in the Eulerian description, Eqs. (5a,b), and by making use of Eqs. (8) and (9) and
resorting to standard algebraic manipulations, their Lagrangian counterparts are obtained as follows:

$$\text{Curl } \mathbf{e} = 0 \quad \text{in} \quad \mathcal{B} \quad \text{and} \quad \text{Div } \mathbf{d} = 0 \quad \text{in} \quad \mathcal{B},$$

(13a,b)

where the Div and Curl denote the divergence and curl operator in the reference configuration respectively.

The electric Neumann and Dirichlet boundary conditions (Eq. 6) and (Eq. 7) are expressed in the undeformed configuration as

$$\mathbf{d} \cdot \mathbf{n} = -q^s \quad \text{on} \quad \partial \mathcal{B}_N^q,$$

(14)

and

$$\phi = h_\phi \quad \text{on} \quad \partial \mathcal{B}_D^\phi,$$

(15)

respectively, where $q^s$ denotes the electric surface charge density in the reference configuration.

To complete the theory, the equilibrium equation in its Lagrangian form is written as

$$\text{Div} (\mathbf{FS}) + \mathbf{b} = 0 \quad \text{on} \quad \mathcal{B},$$

(16)

and the corresponding Neumann boundary condition is given by

$$\mathbf{FSn} = \mathbf{t}^{(n)} \quad \text{on} \quad \partial \mathcal{B}_N^\sigma,$$

(17)

where $\mathbf{S}$ is the symmetric second Piola-Kirchhoff stress tensor, $\mathbf{b}$ denotes the mechanical body force and $\mathbf{t}^{(n)}$ the external traction. Note that generally the external traction consists of a mechanical and an electric contribution. The electric part is due to the fact that associated with the electric field outside of body $\mathcal{B}$ we find a Maxwell stress field which causes an electric traction force acting on the charges lying on $\partial \mathcal{B}$. For simplicity however, we do not consider an electric field outside of $\mathcal{B}$, here and in the following. Consequently, the traction forces reduce to purely mechanical ones.

3 MULTIPlicative Decomposition of the Deformation GrADIENT

3.1 A modified framework and a multiplicative decomposition

As a starting point we consider purely hyperelastic deformations together with deformations induced by the electric field suggesting the existing of a corresponding coupled free energy function $\Psi_{\text{coupled}} (\mathbf{C}, \mathbf{e})$ which depends on the deformation tensor $\mathbf{C}$ and the electric field $\mathbf{e}$. The free energy is considered per unit volume. Note that this contribution relates solely to the deformation of the body. Further contributions must be considered to account for the electric field in a non-deforming medium which will be considered at
a later stage. Hence, in a completely uncoupled formulation the dependency on \( e \) will vanish and \( \Psi_{\text{coupled}} \) will depend on \( C \) alone. The question now arises of how to define \( \Psi_{\text{coupled}}(C, e) \). The material constants involved are to be determined by experiments. Very often, however, the nature of these coupling terms is not clear from the outset and one is motivated to resort to very simple forms of them. While it is true that one can resort to representation theorems of tensor-valued functions, such representations are by far not adequate as they result in general in a large number of material constants and entails complicated issues regarding convexity and existence of solutions.

Instead of the above approach, in describing the polarization of the dielectric material and to achieve more clarity about a coupled formulation we resort now to an alternative approach and apply here the idea of multiplicatively splitting the deformation gradient into a mechanical (elastic) part and a part induced by the electric field. That is, we consider the multiplicative decomposition of the deformation gradient

\[
F = F_{\text{mech}} F_{\text{elec}}. \tag{18}
\]

With the help of the above split in Eq. (18) a purely mechanical (elastic) right Cauchy-Green deformation tensor \( C_{\text{mech}} \) can be defined as

\[
C_{\text{mech}} = F_{\text{mech}}^T F_{\text{mech}} = F_{\text{elec}}^{-T} C F_{\text{elec}}^{-1}. \tag{19}
\]

Now, \( \Psi_{\text{coupled}}(C, e) \) can be defined as purely elastic energy given as \( \Psi_{\text{mech}}(C_{\text{mech}}) \) which is to depend solely on \( C_{\text{mech}} \). The dependency on \( e \), and so the coupling, is implicit through the dependency of \( F_{\text{elec}} \) on \( e \).

To complete the formulation of the free energy we consider the function

\[
\Psi(C, e) = \Psi_{\text{mech}}(C_{\text{mech}}) - \frac{1}{2} c_3 J C^{-1} : e \otimes e, \tag{20}
\]

where the extra term accompanied by the electro-mechanical material constant \( c_3 \) accounts for an energy contribution of the nearly rigid material. It refers to the polarization of linear dielectrics as in Eq. (4) but transferred to the reference configuration via Eq. (10). Consequently, it can be assumed \( c_3 = \epsilon_0 \chi_e \).

Now, from well established fundamental thermodynamical considerations (see e.g. [19]) the relation hold:

\[
p = - \frac{\partial \Psi}{\partial e}. \tag{21}
\]

Making use of the following relation

\[
\frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} : \frac{\partial C_{\text{mech}}}{\partial e} = -4 \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} \cdot \frac{4}{C_{\text{elec}}} \right) e, \tag{22}
\]
the polarization in the un-deformed configuration is given by
\[ p = -\frac{\partial \Psi}{\partial e} = 4 \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} : C_{\text{elec}} \right) e + c_3 J C^{-1} e, \tag{23} \]
where the double contraction operation is here defined as follows
\[ \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} : C_{\text{elec}} \right)_{yz} = \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} \right)_{kn} C_{(elec)knyz}. \tag{24} \]
Substituting Eq. (23) into Eq. (11), the electric displacement or electric charge potential is expressed as
\[ d = \epsilon_0 J C^{-1} e + 4 \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} : C_{\text{elec}} \right) e + c_3 J C^{-1} e \]
\[ = \epsilon J C^{-1} e + 4 \left( C_{\text{mech}} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} : C_{\text{elec}} \right) e. \tag{25} \]
Proceeding similarly as in [13, 14] an augmented free energy function per unit volume is formulated as
\[ \Psi^A (C, e) := \Psi_{\text{mech}}(C_{\text{mech}}) - \frac{1}{2} \epsilon J C^{-1} : e \otimes e. \tag{26} \]
which additionally includes the purely electric energy contribution we find in vacuo with \( \epsilon = \epsilon_0 + c_3 \). Then, the electric charge potential (Eq. (25)) is given by
\[ d = -\frac{\partial \Psi^A}{\partial e}. \tag{27} \]
Finally, the stress tensor is given by
\[ S = 2 \frac{\partial \Psi^A}{\partial C}. \tag{28} \]
With the help of
\[ \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} : \frac{\partial C_{\text{mech}}}{\partial C} = F_{\text{elec}}^{-1} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T}, \tag{29} \]
we arrive at the expression
\[ S = F^{-1} P = 2 F_{\text{elec}}^{-1} \frac{\partial \Psi_{\text{mech}}}{\partial C_{\text{mech}}} F_{\text{elec}}^{-T} + \epsilon J \left[ C^{-1} (e \otimes e) C^{-1} - \frac{1}{2} \left( C^{-1} : e \otimes e \right) C^{-1} \right]. \tag{30} \]
Note that \( \Psi_{\text{mech}} \) must be formulated in a way such that \( S \) vanishes at the reference configuration. The theory is complete once expressions have been found for \( \Psi_{\text{mech}}(C_{\text{mech}}) \) and a constitutive law is formulated for \( F_{\text{elec}} \) as a function of \( e \). In this paper \( \Psi_{\text{mech}}(C_{\text{mech}}) \) is assumed to be the one proposed in [20, 14] which accounts for the incompressible material behaviour of EAP. It features a non-linear statistically based hyperelastic material law originally developed to describe nearly incompressible behaviour of rubber material. As to \( F_{\text{elec}} \) the formulation is addressed in the next subsection.
3.2 A constitutive law for $F_{elec}$

In what follows we present briefly a constitutive law for $F_{elec}$ for full details the reader is referred to [21]. Since polymers exhibit almost no volume change during deformations, it is meaningful to assume for incompressibility to hold for $F_{elec}$. That is $\det F_{elec} = 1$ and so $F_{elec} \in SL^+(3, \mathbb{R})$, where $SL^+(3, \mathbb{R})$ is the special linear group defined over the real numbers of matrices with determinants equal one. To achieve this requirement we assume the following form for $F_{elec}$:

$$F_{elec} := \exp D, \quad \text{tr} D = 0,$$

(31)

where $\exp$ denotes the exponential map defined by

$$\exp D = 1 + D + \frac{1}{2!}D^2 + \frac{1}{3!}D^3 + \ldots,$$

(32)

$D$ is a second order tensor given via a separate electro-mechanical constitutive law, and tr denotes the trace operation of a second order tensor. Accordingly, $D$ must be trace free. The above assumption makes use of the fact that $\det(\exp D) = \exp (\text{tr} D)$. Obviously the choice $\text{tr} D = 0$ results in $\det(\exp D) = 1$. The advantage of the assumption lies in the very fact that it is much easier to fulfil the linear condition of $\text{tr} D = 0$ than the highly non-linear condition of $\det F_{elec} = 1$.

We assume now that $D$ depends only on the electric field vector in the undeformed configuration. An immediate relation is $D = \mathbb{H} e$, with $\mathbb{H}$ being a third order tensor. The relation is linear and must satisfy objectivity requirements which essentially mean that the formulation is to be defined in the local co-ordinates and the results then transformed to the global co-ordinate system. Such requirements can be more easily dealt with if we consider a quadratic form such as:

$$D = 4 C_{elec} : e \otimes e,$$

(33)

where $4 C_{elec}$ denotes a fourth order constitutive tensor and the $: \otimes$ operator represents the double contraction of a fourth order tensor with a second order tensor defined as $D_{ij} = C_{(elec)ijkl} e_k e_l$. The component matrix $C_{(elec)ijkl}$ is considered to be symmetric with respect to indices $k,l$. In addition, due to the near incompressibility of the material, the resulting deformation in compression and expansion needs to accommodate the volume preservation. The condition results in

$$\text{tr} D = 0 \quad \Rightarrow D_{11} + D_{22} + D_{33} = 0.$$

(34)

Now, as the electro-mechanical constitutive law (Eq. 33) is assumed to account only for contraction and expansion due to the electric field but no shear, consequently, the
component matrix of the electro-mechanical constitutive tensor is sparsely set as shown below:

\[ C_{(elec)\ 1111} = C_{(elec)\ 2222} = C_{(elec)\ 3333} = c_1 \]

as well as

\[ C_{(elec)\ 1122} = C_{(elec)\ 1133} = C_{(elec)\ 2211} = C_{(elec)\ 2233} = C_{(elec)\ 3311} = C_{(elec)\ 3322} = c_2 . \]

The additional electro-mechanical material constants \( c_1 \) and \( c_2 \) are required to be determined by experiments. The remaining entries of the component matrix have zero values. Accordingly, with Eq. (34) we find that the two material constants are related by

\[ c_2 = -0.5 c_1 , \]

which reduces the material constants to only one.

We close this subsection by noting that it is straightforward to extend the dependency of \( F_{elec} \) on \( e \) to account for higher order terms than quadratic, should any experimental evidence suggests that, which reflects further advantages of the present formulation.

4 NUMERICAL EXAMPLES

In the following an examples is presented to demonstrate the applicability of the above theory. For further details about the theoretical framework and the numerical implementation the reader is referred to [21]. As mentioned before, the mechanical part of the constitutive model resorts to a non-linear hyperelastic material law [20, 14]. It has been originally developed to describe nearly incompressible behaviour of rubber material making use of three constants, the shear modulus \( C_R \), the bulk modulus \( \kappa \) and parameter \( N \) which addresses the the limited extensibility of the macromolecular network structure of the material. The electro-mechanical coupling of the dielectric material subjected to an electric field is assumed as to respond with contraction in direction of the electric field vector and corresponding transversal expansion. Consequently, the material constant \( c_1 \) in Eq. (33) need to have a negative value and \( c_2 \) a positive one to reflect the presumed coupling of electric field and mechanical deformation. Note that the choice of those two parameters as well as the electric susceptibility constant \( \chi_e \) is purely academic due to lack of experimental data.

The example displayed in Fig. 1 is a spherical shell with 6 holes subjected to variable electric potential boundary conditions uniformly applied to rim of both the top and the bottom hole with opposite sign each. Four holes are arranged around the \( z \) coordinate axis by rotating lines connected to the origin of the coordinate system and enclosing an angle of \( \alpha_{side} = 45^\circ \) with the positive as well as negative \( x \) and \( y \) coordinate directions, respectively. The hole at the top and the bottom of the sphere are cut out by rotating lines enclosing an angle \( \alpha_{pole} = 30^\circ \) with the positive and negative \( z \) coordinate axis, respectively. Due
to the symmetric configuration of this problem only one eighth of the domain is modelled applying the appropriate symmetry conditions. The problem domain is discretized by 384 particles distributed in 3 equally spaced horizontal layers. Corresponding to the linearly increasing electric potential boundary conditions with opposite sign at the top and bottom hole an we find an electric field which leads to compression of the sphere as depicted in Fig. 3. It can be clearly seen that the at first the compression in z coordinate direction results in expansion in x and y coordinate direction as illustrated in Fig. 2. At the final stage of the simulation the sphere is also contracting in x and y direction.

5 CONCLUSION

In this paper, the electro-mechanically coupled material behaviour is implicitly incorporated via a multiplicative decomposition of the deformation gradient into an elastic
and an electrically-induced part. The electric deformation gradient is formulated via a separate constitutive law which provides the flexibility to address different kinds of electromechanical coupling. In fact, in the light of the mentioned decomposition, the formulation of the electromechanical coupling becomes almost straightforward. The strength of the present framework is demonstrated by various numerical examples of large electromechanically coupled deformations based on a meshfree method.

REFERENCES


PARAMETRIC IDENTIFICATION OF MATHEMATICAL MODELS OF COUPLED CONDUCTIVE-RADIATIVE HEAT TRANSFER

ALEKSEY V. NENAROKOMOV*, OLEG M. ALIFANOV* AND DMITRY M. TITOV*

*Dept. of Space System Engineering
Moscow Aviation Institute (Technical University)
4 Volokolamskoe Hgw.
Moscow, 125993, Russia
e-mail: aleksey.nenarokomov@mai.ru

Key words: thermal conductivity, heat capacity, emissivity, thermal conductivity, radiation

Abstract. In many practical situations it is impossible to measure directly such characteristics of analyzed materials as thermal and radiation properties. The only way, which can often be used to overcome these difficulties, is indirect measurements. This type of measurements is usually formulate as the solution of inverse heat transfer problems. Such problems are ill-posed in mathematical sense and their main feature shows itself in the solution instabilities. That is why special regularizing methods are needed to solve them. The experimental methods of identification of the mathematical models of heat transfer based on solving of the inverse problems are one of the modern effective solving manners.

The goal of this paper is to estimate thermal and radiation properties of advanced materials using the approach based on inverse methods (as example: thermal conductivity $\lambda(T)$, heat capacity $C(T)$ and emissivity $\varepsilon(T)$). New metrology under development is the combination of accurate enough measurements of thermal quantities, which can be experimentally observable under real conditions and accurate data processing, which are based on the solutions of inverse heat transfer problems. In this paper, the development of methods for estimating thermal and radiation characteristics is carried out for thermally stable high temperature materials. Such problems are of great practical importance in the study of properties of materials used as non-destructive surface coating in objects of space engineering, power engineering etc.

Also the corresponded optimal experiment design problem is considered. The algorithm is based on the theory of Fisher information matrix.

1 INTRODUCTION

In the modern engineering systems we deal with structures operating in the conditions of intensive, often extreme thermal effects. The general tendency in the development of technology is connected with the increase of the number of responsible, thermally loaded engineering objects. For such systems the support of thermal conditions is one of the most important aspects of design, determining the main design solutions. The modern approaches
to the design of structures assume broad application of mathematical and physical simulation methods. But mathematical simulation is impossible if there is no true information available on the characteristics (properties) of objects analyzed. In the majority of cases in practice the direct measurement of materials’ thermophysical properties, especially of complex composition, is impossible. There is only one way which permits to overcome these complexities - the indirect measurement. Mathematically, such an approach is usually formulated as a solution of the inverse problem: through direct measurements of system’s state (temperature, component concentration, etc.) define the properties of a system analyzed, for example, the thermophysical properties. Violation of cause-and-effect relations in the statement of these problems results in their correctness in mathematical sense (i.e., the absence of existence and/or uniqueness and/or stability of the solution). Hence to solve such problems we develop special methods usually called regularized.

In estimating properties of modern structural, thermal-protective and thermal-insulating materials - as temperature-dependent - the most effective are methods based on solution of the coefficient inverse heat conduction problems. The most promising direction in further development of research methods for non-destructive composite materials using the procedure of inverse problems is the simultaneous determination of a combination of material’s thermophysical and radiation properties (thermal conductivity $k(T)$, heat capacity $C(T)$ and integral emissivity $\varepsilon(T)$). Such problems are of great practical importance in the study of properties of composite materials used as non-destructive surface coating in objects of space technology, power engineering etc. [1], [2], [3], [4], [5], [6], [7]. The experimental equipment and the method developed could be applied for determination of material’s three characteristics; the availability of two specimens of the material allows us to provide uniqueness of the solution. The mathematical model of heat transfer in specimen is

$$C(T) \frac{\partial T}{\partial \tau} = \frac{\partial}{\partial x}\left( k(T) \frac{\partial T}{\partial x} \right), \quad x \in (X_0, X_1), \quad \tau \in (\tau_{\text{min}}, \tau_{\text{max}}]$$  \hspace{1cm} (1)

$$T(x, \tau_{\text{min}}) = T_0(x), \quad x \in [X_0, X_1]$$  \hspace{1cm} (2)

$$- \beta k(T) \frac{\partial T(X_0, \tau)}{\partial x} + \alpha_1 T(X_0, \tau) = q_1(\tau), \quad \tau \in (\tau_{\text{min}}, \tau_{\text{max}}]$$  \hspace{1cm} (3)

$$- k(T) \frac{\partial T(X_1, \tau)}{\partial x} = \varepsilon_{\text{eff}}(T) \sigma \left( T_h^4(\tau) - T^4(X_1, \tau) \right), \quad \tau \in (\tau_{\text{min}}, \tau_{\text{max}}]$$  \hspace{1cm} (4)

Where

$$\varepsilon_{\text{eff}}(T) = \frac{\varepsilon(T) \varepsilon_h(T_h)}{\varepsilon(T) + \varepsilon_h(T_h) - \varepsilon(T) \varepsilon_h(T_h)}$$

In model (1)-(4) the quantities $C(T)$, $k(T)$ and $\varepsilon(T)$ are unknown. If emissivity of the heater material ($\varepsilon_h$) is known a-priori, and its temperature is measured, the heat flux from the heater can be calculated as irradiation with known (measured by thermocouple) temperature.
of the heater $T_h(\tau)$ and a-priori known (theoretically) emissivity of heater $\varepsilon_h(T)$. In presented paper a case is considered, when $\varepsilon_h(T)$ is not known a-priori and should be estimated. At this paper the emissivity of the heater is considered as additional (forth) estimating functions. Therefore the accuracy of the inverse problems, considered bellow, will not depend to the a-priory information about the radiation properties of the heater’s material.

The results of temperature measurements inside the specimen are assigned as necessary additional information to solve an inverse problem

$$T^{exp}(x_m, \tau) = f_m(\tau), \quad m = \overline{1, M}$$

With the presented statement of inverse problem, the data gained in one experiment are not sufficient for simultaneous recovery of three thermal and radiative characteristics (thermal conductivity, heat capacity and emissivity), because data by values of the heat flux applied to a specimen are also needed.

The execution of the single experiment is not enough to provide the conditions of uniqueness of the inverse problem solving by simultaneous estimating of thermal conduction, heat capacity and emissivity of the testing material. To solve this problem the data of several $N$ (in partial three) similar experiments with equal material specimen and different heating regimes were processed simultaneously.

The experimental equipment and the method described below could be applied for estimating of material’s three characteristics; the availability of a few specimens of the material allows us to provide uniqueness of the solution. This paper is concerned with modification of the approach, presented at [8].

2. INVERSE PROBLEM ALGORITHM

In the inverse problem Eqn.(1)- Eqn.(5) it is necessary first of all to indicate as a temperature range $[T_{\min}, T_{\max}]$ of the unknown functions, which is general for all experiments, and for which the inverse problem analysis has a unique solution. For $T_{\min}$ the minimum value of initial temperature is used. Of much greater importance is a correct sampling of value $T_{\max}$. Proceeding from the necessity to provide uniqueness of solution, it seems possible to sample, for $T_{\max}$, a minimum among maximum temperature values gained on the thermocouple positioned on the heated surface at every testing specimen. The same should be done for the heater temperature range $[T_{h,\min}, T_{h,\max}]$.

Suppose then that the unknown characteristics are given in their parametric form. With this purpose introduce in the interval $[T_{\min}, T_{\max}]$, three uniform difference grids with the number of nodes $N_i$, $i = 1,2,3$.

$$\omega_i = \left\{ T_k = T_{\min} + (k - 1)\Delta T, \quad k = \overline{1, N_i - 1} \right\}, \quad i = \overline{1, 3} \tag{6}$$

The function
\[ B^{(j-1)}(t) = B^{(j-1)}(T_k, T_{k+1}, ..., T_{k+j}, \tau) = \sum_{i=1}^{k+1} \frac{(T_i - T)^{j-1}}{\omega_k(T_i)} \]  

(6a)

where \( \omega_k = \prod_{i=1}^{j} (T - T_{k+i}) \) and \( \max \left\{ (T_i - T)^{j-1} \right\} \) is called B-spline of the \((j-1)\) degree relatively with nodes \( T_k, T_{k+1}, ..., T_{k+j} \).

When solving practical problems, B-splines are used with so-called "natural" boundary conditions

\[ u'(T_{\text{min}}) = u''(T_{\text{max}}) = 0 \]  

(6b)

where \( u \) is desired function.

Then, in case of cubic B-splines \((j-1=3)\), the unknown function is presented as

\[ u_i(\tau) = \sum_{k=1}^{N_i} u_k \varphi_{i,k}(\tau) \]  

(6c)

\[ \varphi_1(T) = 2B_1(\bar{T} + \Delta T) + B_0(\bar{T}) \]  

(6d)

\[ \varphi_2(T) = -B_0(\bar{T} + \Delta T) + B_0(\bar{T} - \Delta T) \]  

(6e)

\[ \varphi_k(T) = B_{k-1}(\bar{T}), \quad k = 3, ..., N_i - 2 \]  

(6f)

\[ \varphi_{N_i-1}(T) = B_0(\bar{T} - (N_i - 2)\Delta T) - B_0(\bar{T} - N_i\Delta T) \]  

(6g)

\[ \varphi_{N_i}(T) = B_0(\bar{T} - (N_i - 1)\Delta T) + 2B_0(\bar{T} - N_i\Delta T) \]  

(6h)

where

\[ B_k(T) = B_0(\bar{T} - k\Delta T) \]  

(6i)

\[ \bar{T} = T - T_{\text{min}} ; \]  

\[ B_0(T) = \left( \frac{1}{6}T + 2\Delta T \right)^3 - 4\left( T + \Delta T \right)^3 + 6\left( T - \Delta T \right)^3 + \left( T - 2\Delta T \right)^3 \]  

(6j)

The function \( B_0(T) \) has the property

\[ B_0(T) = \begin{cases} >0, & \text{if} \quad -2\Delta T < T < 2\Delta T \\ =0, & \text{if} \quad |T| \geq 2\Delta T. \end{cases} \]  

(6k)

This property makes the computational algorithm simpler.

Approximating the unknown functions on grids Eqn. (6) using the cubic B-splines
\[ C(T) = \sum_{k=1}^{N_1} C_k \varphi_k(T) \]
\[ k(T) = \sum_{k=1}^{N_2} k_k \varphi_k(T) \]
\[ \varepsilon(T) = \sum_{k=1}^{N_3} \varepsilon_k \varphi_k'(T) \] (7)

where \( C_k, k = 1, \ldots, N_1, k_k, k = 1, \ldots, N_2, \varepsilon_k, k = 1, \ldots, N_3 \) - parameters.

Let’s introduce in the interval \([T_{\text{min}}, T_{\text{max}}]\) uniform difference grids with the number of nodes \( N_4 \)

\[ \omega_4 = \left\{ T_k = T_{\text{min}} + (k - 1)\Delta T_k, \quad k = 1, N_4 - 1 \right\} \] (6l)

and approximate the unknown function \( \varepsilon_h(T) \) on grids (6a) using the cubic B-splines

\[ \varepsilon_h(T) = \sum_{k=1}^{N_4} \varepsilon_{hk} \varphi_k(T) \] (7a)

where \( \varepsilon_{hk}, k = 1, \ldots, N_4 \) - parameters.

As a result of approximation, the inverse problem is reduced to the search of a vector of unknown parameters \( \bar{p} = \{ p_k \}, \quad k = 1, \ldots, N_p \), with dimensions \( N_p = N_1 + N_2 + N_3 + N_4 \). Writing down a least-square discrepancy of the calculated and experimental temperature values in points of thermal sensors positioning, than the residual functional will depend to four functions

\[ J(\bar{p}) = J(C(T), k(T), \varepsilon(T), \varepsilon_h(T)) = \sum_{n=1}^{N} \sum_{m=1}^{M} \int_{r_{\text{min}}}^{r_{\text{max}}} \left( T^n(x^n_m, \tau) - f^n_m(\tau) \right)^2 d\tau \] (8)

where \( T^n(x, \tau) \) is determined from a solution of the boundary-value problem Eqn. (1)- Eqn. (4) for \( n \)-th experiment using the approximations of Eqn. (7). It is assumed here that the conditions of uniqueness of the inverse problem solving are satisfied. Below to simplify the notation of equations index \( n \) will be excluded.

So, proceeding from the principle of iterative regularization [8], [9], [10], the unknown vector can be determined through minimization of functional Eqn. (8) by gradient methods of the first order prior to a fulfilment of the condition

\[ J(\bar{p}) \leq \delta_f \] (9)

where \( \delta_f = \sum_{m=1}^{M} \int_{r_{\text{min}}}^{r_{\text{max}}} \sigma_m(\tau) d\tau \) - integral error of temperature measurements \( f_m(\tau), \quad m = 1, M \), and \( \sigma_m \) - measurement variance.
To construct an iterative algorithm of the inverse problem solving a conjugate gradient method was used.

The greatest difficulties in realizing the gradient methods are connected with calculation of the minimized functional gradient. In the approach being developed the methods of calculus of variations are used.

\[
J_{c_i}' = - \int_{\tau_{\text{max}}}^{\tau_{\text{max}}} \psi(x, \tau) \cdot \frac{\partial T}{\partial \tau} \, dx \, d\tau
\]

\[J_{c_i}' = - \int_{\tau_{\text{max}}}^{\tau_{\text{max}}} \psi(x, \tau) \left( \frac{\partial^2 T}{\partial \tau^2} \cdot \phi_k(T) + \left( \frac{\partial T}{\partial \tau} \right)^2 \cdot \frac{\partial \phi_k}{\partial T} \right) \, dx \, d\tau - \]

\[\int_{\tau_{\text{max}}}^{\tau_{\text{max}}} \psi(x_0, \tau) \frac{\partial T}{\partial \tau} \left( X_0, \tau \right) \phi_k(T) \, dx \, d\tau + \]

\[\int_{\tau_{\text{max}}}^{\tau_{\text{max}}} \psi(x_1, \tau) \frac{\partial T}{\partial \tau} \left( X_1, \tau \right) \phi_k(T) \, dx \, d\tau,
\]

\[k = 1, N_1, N_2, N_3, N_4\]

where \(\psi(x, \tau)\) - solution of a boundary-value problem adjoint to a linearized form of the initial problem Eqn. (1)- Eqn. (4).

\[-c(T) \frac{\partial \psi_m}{\partial \tau} = k \frac{\partial^2 \psi_m}{\partial \tau^2}, \]

\[x \in (x_{m-1}, x_m), x_0 = X_0, x_{M+1} = X_1, m = 1, M+1, \tau \in (\tau_{\text{min}}, \tau_{\text{max}})\]

\[\psi_m(x, \tau_{\text{max}}) = 0, x \in [x_{m-1}, x_m], m = 1, M+1\]

\[-k \frac{\partial \psi_{M+1}}{\partial \tau}(\varepsilon + \varepsilon_h - \varepsilon_h) - \frac{d\varepsilon}{dT} \frac{\partial T}{\partial \tau} \left( 1 - \varepsilon_h \right) k \frac{\partial T}{\partial \tau} \psi_{M+1} \left( X_1, \tau \right) + \]

\[\]
3. OPTIMAL EXPERIMENT DESIGN

As has been said above, the developed method of determining the $C(T)$, $k(T)$ and $\varepsilon(T)$ of a material necessitates the solution of an ill-posed inverse problem. The accuracy of estimating the desired properties is determined largely by the experimental merits, and one is connected with the problem of optimal experiment design. When the processing of the experimental data is the solution of ill-posed problems, optimal experiment design essentially entails the sampling of merits that will ensure the best conditioning of the computational algorithm [9]. The available a priori information about the experimental data are used to formulate a certain scalar criterion of optimality $\Phi(\xi)$, which depends on the experiment design $\xi$ and characterizes the conditioning of the algorithm for solving the inverse problem. It is reasonable to assume that $\Phi(\xi)$ has a lower bound on the set of possible designs $\Sigma$ and that optimal design $\xi^*$ exists such that

$$
\xi^* = \arg\inf_{\xi \in \Sigma} \Phi(\xi)
$$

In order to formulate the experiment design problem, it is necessary to select the design criterion $\Phi(\xi)$, to identify the experimental merits comprised in its actual design, i.e., the merits that significantly influence the criterion $\Phi(\xi)$, and to formulate the domain of possible designs $\Sigma$.

On the basis of the simulation presented at [9] the following set of merits is used in the present paper for the experiment design

$$
\xi = \{M, x_m, m = 1, M\}
$$

Where $M$ is the number of thermocouple and $x_m, m = 1, M$ is the coordinate of the thermocouple installation. In order to solve the problem (11), it is necessary to determine the domain of possible designs $\Sigma$. The following considerations must be taken into account in forming this set:

1) $x_m \in [X_p, X_f]$;
2) $M \in [1; 3]$;

Following [9] the determination of the vector $\bar{p} = \{p\}^N_j$ can be reduced to the solution of the system of linear equations.
\[ A \overline{F} = d \]  

where (without number of experiments)

\[
A = \{ a_{kn} \}, \quad k = 1, N_p, \quad n = 1, N_p, 
\]

\[
a_{km} = \sum_{m=1}^{m} \int \vartheta_k(x_m, \tau) \vartheta_n(x_m, \tau) d\tau
\]

where \( \vartheta_n(x_m, \tau) \) and \( \vartheta_k(x_m, \tau) \) are the corresponded sensitivity functions, and \( A \) is the Fisher information matrix of the considered system [9]. Following [9], the determinant of \( A \) is adopted as the optimality criterion of the experimental design:

\[
\Phi(\xi) = - \det A(\xi)
\]

Problem (11) then acquires the form

\[
\xi^* = \arg \inf_{\xi \in \Sigma} (- \det A(\xi))
\]

The optimal design problem (23) is solved by the numerical projective gradient method of optimization. The iterative process is formulated as follows in this case:

1) an initial approximation of the experiment design \( \xi^0 \) is specified,
2) the value of the gradient of the functional \( \Phi(\xi) \), the descent step \( \alpha^s \), and the experimental design in the next iteration are computed, where

\[
\xi^{s+1} = \xi^s + \alpha^s \left( \Phi(\xi) \right)', \quad s = 0, 1, 2, \ldots, \xi \in \Sigma
\]

3) the iterative process is stopped when the optimality criterion has the same value in two successive iterations, i.e., when the following condition is satisfied

\[
\left| \Phi^{s+1}(\xi) - \Phi^s(\xi) \right| \leq \varepsilon^*
\]

where \( \varepsilon^* > 0 \) is the a priori specified relative error of exit from the iterative process.

The size of the step \( \alpha^s \) is selected on the basis of the condition

\[
\min_{\alpha^s \in \mathbb{R}, \xi \in \Sigma} \Phi\left( \xi^s + \alpha^s \left( \Phi(\xi) \right)' \right)
\]

by one of the conventional techniques of one-dimensional optimization.

4. EXPERIMENTAL VERIFICATION

An example to how apply the approach suggested is presented below. Given are the results of data processing of specimen experimental investigations with modern composite materials. The investigations were carried out on a set of pairs of specially manufactured specimens (the first in the pair for simultaneous estimating material’s heat capacity per and thermal conductivity and the second for determining boundary conditions).
The models of test material are the square plates of 50x50x15 mm (Fig. 2) with four thermocouples installed in the specimen. An installation of thermosensors in specimens was chosen from a solution of the problem of optimal experiment design. The co-ordinates of thermocouple positioning (according optimal experiment design) in the first set of specimens, for estimating the material’s thermal characteristics, had the following values: $x_0 = 0\ mm$ (for a boundary condition of the first kind sensor readings on the internal surface were used), $x_1 = 7.5\ mm$, $x_2 = 11.8\ mm$, $x_3 = 15\ mm$ (positioned on the exposed surface). The second set of specimens, for defining the emissivity, has the thermocouples at points $x_0 = 0\ mm$, $x_1 = 12.65\ mm$, $x_2 = 13.65\ mm$, $x_3 = 15\ mm$.

The number of approximation parameters $N_1$, $N_2$, $N_3$ and $N_4$ for every characteristic was assumed to be 5. During specimen heating a theoretically preset time dependence of surface temperature (Fig. 3) was provided. The measurement errors were estimated as 5 %. A comparison of experimentally measured and calculated (with the help of thermal characteristics obtained from a solution of the inverse problem) temperature values in points of thermocouple positioning is shown on Fig. 3 (only for one specimen). The results are in agreement, which shows the robustness of the inverse problem algorithm.

The results proper of the inverse problem solving - the composite material thermal characteristics and emissivity are given on Fig. 4 (the results for two sets by two experiments in vacuum and air). The accuracy of these results of the inverse heat conduction problem was verified using different (quite distinct from each other) initial approximations for an iterative process. The results show reasonable agreement.
CONCLUSIONS

The paper seeks to describe the algorithm developed to process the data of unsteady-state thermal experiments. The algorithm is suggested for determining these unknown on the surface of a slab as a solution of the nonlinear inverse heat conduction problem in an extreme formulation.

The following main factors have an influence on the accuracy of the inverse heat conduction problem (in sequence of significance): the errors in coordinates of thermosensor positions; the errors in values of different characteristics; the errors in estimating the residual...
level. It was shown that in the cases considered the accuracy of the inverse problems solution is compatible with the errors of the simulated "experimental measurements". Next step in the development of the proposed approach is to consider an estimating interface conductance between periodically contacting surface of specimen and heater foil using the approach similar [11].

ACKNOWLEDGMENTS

This work was done with the financial support of RBRF (grant #11-08-00952-a).

REFERENCES

SIMULATION OF THE THERMAL PROFILE OF A MUSHY METALLIC SAMPLE DURING TENSILE TESTS

F. Palama Bongo*, J. Favergeon†, M. Rachik†, D. Marceau+++  
*†Laboratoire Roberval UMR 7337  
Université de Technologie de Compiègne (UTC)  
Centre de Recherche de Royallieu, BP 20529 60205 Compiègne cedex France  
E-mail: Franck.Palama-Bongo@utc.fr

+++ Centre Universitaire de Recherche sur l’Aluminium (CURAL)  
Université du Québec à Chicoutimi  
555, Boulevard de l’Université, Chicoutimi (Québec) Canada G7H 2B1  
Email: Daniel_Marceau@uqac.ca - Web page: http://www.cural.qc.ca/

Key words: Joule effect Heating, Gleeble 3800, Numerical simulation, Inverse method.

Abstract. Strain measurement is a major challenge in tensile tests performed in a mushy state. While non-contact technique devices like the laser speckle extensometer remain the most reliable facility for this type of measurement, these devices are often not readily available. So the strain measurement is usually performed by determining the length of the ‘‘hot zone’’ of the sample. This is possible with the help of the thermal profile associated with the sample under heating. The purpose of our work is to develop a numerical model to predict the thermal profile of a A356 aluminum alloy sample at high temperature, taking into account the device geometry and characteristics.

We simulate the joule heating effect using the FE software Abaqus. Our model takes into account the grips of a Gleeble machine, the thermal contact conductance and electrical contact resistance at the grip-sample interfaces, as well as the convection heat transfer on the free surfaces of the system. These thermo-physical properties have been determined by fitting the experimental thermal profile obtained at 545°C. The model was then used to simulate the temperature profile on the sample at higher temperatures (when the sample is in the mushy state).

The thermal profile predicted by our model is in excellent agreement with the profile obtained experimentally.

1 INTRODUCTION

The Digital Image Correlation is the most reliable technique for measuring the deformation on a mushy sample under tensile test. It consists in using a camera that follows the motion of a speckle printed on the sample surface. Non-contact measuring devices include laser speckle
Extensometer used by S. Dziallach [1] and laser dilatometer used in the work of AB Phillion [2]. However, these measuring devices are expensive and not available in our case.

When non-contact measuring devices are not available, deformation on a sample under tensile test is determined as the ratio between two amounts: the displacement of the testing machine grips and the change in length of the specimen hot zone. In a system with induction heating, the extent of the hot zone is known without difficulties: it corresponds to the length of the induction coil [3, 4]. In contrast, in the case of heating by Joule effect, the hot zone is not known \textit{a priori}. In an experimental approach [4, 5], the hot zone length of a sample heated by Joule effect is obtained from the measured temperature profile along the specimen. This temperature profile has a parabolic shape along the test sample length, whose apex flattened which indicates the length of the hot zone.

However, the determination of the temperature profile on a mushy sample is very challenging. It is difficult to weld thermocouples on the semi-solid part of a joule effect-heated sample. In addition, holes made in the sample for thermocouples insertion induce current flow modification, and then, temperature increases around the region of the hole. For this reason, inserted thermocouples are not reliable for temperature measurement. A technique for controlling the temperature of the hot zone without having to set thermocouples was developed by Q. Han [6]. In his work, the author has established calibration curves giving the temperature of the hot zone of the specimen from the temperature of a point located outside it. These curves are drawn as long as it is possible to attach a thermocouple to the hot zone. When the temperature increases considerably and the hot zone becomes mushy, AB Phillion [2] suggests to extrapolate calibration curves to higher temperatures. This extrapolation is performed assuming a linear calibration curve. This assumption leads to an underestimation of the temperature of the hot zone.

Due to the experimental difficulties in measuring the temperature of the hot zone, numerical simulation is a useful alternative. This approach was proposed for the first time in the work of Changli Zhan et al [7] who study the thermal profile of ultra-resistant steels. This is the approach that we will deploy here to characterize the length of the hot zone of a A356 alloy sample heated by Joule effect via the Gleeble 3800 thermomechanical platform. The simulation is performed using the Abaqus software. First, we present the experimental setup consisting of the Gleeble machine and a cylindrical sample. Then the electrical transfer and heat exchange associated with the components of the experimental setup are presented. At the same time, the equations describing the electrical and thermal phenomena involved are given. Following this description, the experimental setup is modelled by assuming some geometric simplifications and some other assumptions are taken into account in order to reduce the number of simulation parameters. Finally, unknown parameters of the numerical model are identified by the inverse method.

2. Experimental setup

In this work, we used for heating tests the Gleeble 3800. This latter is a thermo-mechanical platform whose main characteristics are the high heating and cooling rates that may be reached during tests. Those characteristics allow to reproduce conditions encountered in
foundry industry. The figure 1 shows the experimental setup with each component of the Gleeble indicated by a number: on each end of the specimen are located a pair of jaws (4), two pairs of grips (5) and some other metallic components. Through these components, electric current flows as indicated by the diagram of figure 2. Continuous lines indicate a significant current flow while dashed lines show a negligible one. Heat exchange between components takes place by conduction while free surfaces of these components are submitted to convection and radiation.

3 Electrical transfer and heat exchange balance

The Joule effect heating modeling passes through the review of the transfer electrical balance and heat exchange between the specimen and the various components of the Gleeble. In Joule heating problems, the main unknowns are electrical potential and temperature at each point of the experimental setup. These parameters are determined by solving Maxwell's equations related to electric charge and energy conservation.
3.1. Maxwell’s equations

Electrical potential field in a conductor is governed by Maxwell's equations related to the conservation of electric charge. In steady state, direct current is given by the following equations:

\[
\begin{align*}
J &= -\sigma_{\text{elec}} \nabla \phi \\
\nabla \cdot J &= 0
\end{align*}
\] (1) (2)

Where \( \phi \) is electrical potential, \( \sigma_{\text{elec}} \) electrical conductivity, and \( J \) is the electrical current density vector.

The key of electrical problem then consists in solving the equation given by:

\[
\nabla \cdot (\sigma_{\text{elec}} \nabla \phi) = 0
\] (3)

The boundary conditions encountered in electrical problems are of two types: prescribed values of electric potential \( \phi_{\text{imp}} \) (4a) and current density \( J_{\text{imp}} \) (4b). Those boundary conditions are applied on two distinct surfaces of the studied domain. In the case of two surfaces in contact, an effective electrical contact resistance \( h_{\text{elec}} \) is applied at interface (4c).

\[
\begin{align*}
\phi_{\text{imp}} &= 0 \\
-J \cdot n &= J_{\text{imp}} \\
-J \cdot n &= h_{\text{elec}}(\phi - \phi_{\text{contact}})
\end{align*}
\] (4a) (4b) (4c)

In the above equations, \( n \) is the outward unit vector either normal to the electrical charge surface (4a) or normal to one of the surfaces of two bodies in contact (4c). \( \phi_{\text{contact}} \) denotes the local electrical potential at the limit of the body.

3.2. Energy conservation equation

In thermoelectric problems, total energy includes energy created by Joule effect \( P^\text{elec} \) and heat associated with the power of deformation due to thermal expansion. The latter being neglected, the equation of energy conservation is written as follows:

\[
\rho \frac{dH}{dt} - \nabla \cdot (\lambda \nabla T) = P^\text{elec}
\] (5)

Where \( \rho \) denotes the density, \( \lambda \) the thermal conductivity, \( T \) the temperature. The specific enthalpy \( H \) is defined as follows:
\[ H = \int_{T_{\text{ref}}}^{T} c_p(\tau) \, d\tau + f_i L \]  

(6)

With \( T_{\text{ref}} \) an arbitrary reference temperature, \( c_p \) the specific heat, \( f_i \) the mass fraction of liquid and \( L \) the specific latent heat of fusion.

The heat generated by electrical resistance is given by Joule's law:

\[ p_{\text{elec}}^\text{elec} = \sigma_{\text{elec}}^{-1} \mathbf{J} \cdot \mathbf{\nabla} \phi = \sigma_{\text{elec}} \mathbf{\nabla} \phi \cdot \mathbf{\nabla} \phi \]  

(7)

As in the case of electrical problem, boundary conditions in a heat flow problem are applied either at the border or at the interface of two bodies in contact. Regarding the boundary conditions at the border, it may be prescribed by a heat flux \( q_{\text{imp}} \) (8a) or a heat transfer with the surrounding environment through an equivalent exchange coefficient \( h_{\text{th,eff}} \) (8c). At interfaces, heat transfer from one surface to another is provided via a thermal contact conductance \( h_c \). Heat balance at the interfaces is given by equation (8b), with the interface energy denoted by \( p_{\text{interface}}^\text{elec} \).

\[ -\lambda \nabla T \cdot \mathbf{n} = q_{\text{imp}} \quad \text{(8a)} \]
\[ -\lambda \nabla T \cdot \mathbf{n} = h_c (T - T_{\text{contact}}) + \frac{b}{b + b_{\text{contact}}} p_{\text{interface}}^\text{elec} \quad \text{(8b)} \]
\[ -\lambda \nabla T \cdot \mathbf{n} = h_{\text{th,eff}} (T - T_{\text{env}}) \quad \text{(8c)} \]

Where \( b = \sqrt{\lambda \rho c_p} \) denotes the thermal effusivity and the subscript “contact” refers to amounts at interfaces and the subscript “env” refers to amounts associated with the surrounding environment. The interface energy is given by \( p_{\text{interface}}^\text{elec} = h_{\text{elec}} (\phi - \phi_{\text{contact}})^2 \).

The review of thermo-electrical energy and equations governing the Joule heating shows several thermo-physical parameters involved in the problem. At interfaces, the parameters to be managed include electrical contact resistance, and thermal effusivity. On the free surfaces, the main parameters are the emissivity and convective heat transfer coefficient. In view of the relatively large number of interfaces identified in the experimental setup, and thermo-physical parameters involved in the problem, some simplifications are introduced in the numerical model, for obvious reasons of reduction of computation time.

4. Numerical model
The numerical model developed in this work includes only the specimen and the Gleeble grips. This simplification leads to postulate an equivalent heat transfer coefficient on the grips surfaces in order to take into account the components of the Gleeble omitted in the geometric model. In addition, the numerical model is only half of the experimental setup thanks to the horizontal symmetry plane (figure 3). Simplifying assumptions are also made to reduce the number of unknown thermo-physical parameters. First, we assume a perfect contact at grips-sample interfaces. That leads to no heat generation ($p_{\text{interface}} = 0$) and a quasi-continuity of the electrical potential $\bar{\varnothing}$. Also the electrical contact resistance $h_{\text{elec}}$ is fixed to a low value. Secondly, the level of temperatures involved in the case of a A356 sample being relatively low, radiation is negligible compared to convection and conduction. This assumption leads to impose a zero value for emissivity.

Figure 3: Gleeble Setup Modelisation on Abaqus/CAE 6.11-3: grips and sample (Section along an horizontal plane).

The thermo-physical parameters to be determined in this study are convection coefficient on the free surfaces of grips and sample, the thermal contact conductance at grips-sample interfaces and the equivalent heat transfer coefficient (defined on grips surfaces in contact with the rest of the experimental setup not shown in the model). The last two are obtained by inverse method based on calibration of the numerical model to the experimental data while the first parameter is derived from empirical relationships provided by the literature on heat transfer by convection.

Heat transfer through the free surfaces of the setup is natural convection type. Transfer coefficients are derived from the following empirical relations:

\[
h = 1.32 \left(\frac{T}{D}\right)^{1/4} \quad (9a)
\]

\[
h = 1.42 \left(\frac{T}{L}\right)^{1/4} \quad (9b)
\]
In the above equations, $T$ is the temperature in degree Celsius. Equation (9a) is valid for a horizontal tube of diameter $D$. It is employed to determine the coefficient of convection through the free surface of the sample. Relations (9b) and (9c) are valid respectively for a vertical plate and a upper surface of a horizontal plate of characteristic length $L$. They are used to determine the convection coefficient on the free surfaces of the grips.

The various electrical and thermo-physical parameters listed above, whether the parameters set by assumption (electrical contact resistance) or settings determined by calculation (convection coefficient) or parameters to be determined by the inverse analysis (thermal contact conductance and equivalent heat transfer coefficient) will be assumed to be temperature independent for the numerical simulation. It is the same for thermo-physical properties of sample and grips materials, namely the A356 alloy and copper. Thermo-physical properties of the former are obtained from the data base of Auburn University while the second properties are provided by the manufacturer of the Gleeble machine. These properties are summarized in the table 1.

<table>
<thead>
<tr>
<th>Material</th>
<th>Density (Kg/m³)</th>
<th>Thermal conductivity (W/m/K)</th>
<th>Electrical conductivity (Ω.m⁻¹)</th>
<th>Spécific Heat (J/Kg/K)</th>
<th>Latent heat (J/Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A356</td>
<td>2500</td>
<td>170</td>
<td>9803921</td>
<td>1315</td>
<td>437637</td>
</tr>
<tr>
<td>Copper</td>
<td>8900</td>
<td>350</td>
<td>20000000</td>
<td>450</td>
<td>176000</td>
</tr>
</tbody>
</table>

4.1. Loading and boundary conditions

During heating tests performed in this study, temperature was the controlled parameter defined in a given heat cycle. The current density that causes heating is adjusted via a PID controller so that heating is carried out at a prescribed rate. In numerical simulation, control is provided in the same way with the difference that the current density is controlled manually to achieve the desired temperature.

$J_{imp}$ current density was imposed on the left grips of the Gleeble. It was associated with a zero potential $\varnothing_{imp} = 0$ applied on the right grips. At grips-sample interfaces, an electrical conductance (the inverse of electrical contact resistance) of the order of $10^8 \Omega^{-1} m^2$ was applied for electrical current transfer. In addition to the boundary conditions related to the electrical problem, there are those related to thermal one. The free surfaces of the grips as well as those of the sample are submitted to heat exchange by convection. This heat transfer is ensured through the exchange coefficients derived from the relationships described in the previous paragraph (9a, 9b and 9c). At grips-sample interfaces heat transfer is ensured through the thermal contact conductance $h_{theff}$. 
4.2. Identification of parameters

To determine the two unknown parameters of this simulation (namely thermal contact conductance at grips/sample interfaces and the equivalent heat transfer coefficient), the numerical model was calibrated on the experimental data from heating tests achieved via the Gleeble machine. The identification is performed for a test temperature of 500°C. The approach used to determine these parameters consists in varying the value of the unknown parameters in the numerical model until the experimental temperature profile coincides with the thermal profile resulting from the numerical simulation. During heating, temperature of grips and sample are recorded. The locations of thermocouples welded on the sample and grips are shown in figure 4. Giving the temperature at the middle, TCc, and at the extremity of the sample, TC_E, the numerical value of the thermal contact conductance at the grips/sample interface is chosen such that the numerical couple (TCc, TC_E) is closed to the experimental values. Once the thermal contact conductance was determined, the equivalent coefficient of heat transfer is selected such that the numerical grips temperature is close to that recorded experimentally.

5. DISCUSSION

With the values of thermal contact conductance (3500 W. m⁻²K⁻¹) and equivalent heat transfer coefficient (1000 W. m⁻²K⁻¹) thus determined by inverse method based on the experimental results performed with 545°C temperature at the center of the sample, simulations are performed for set points of 200, 300, 400°C and 500°C at the center of the sample. Also, new heating tests were performed with the same set points at the center of the sample (the thermocouples are still located as shown in figure 4). Figure 5 shows the agreement between the experimental data and the numerical model, regarding thermal profile.
in the axial direction. This agreement is particularly good in the central part of the sample. That validates our numerical model.

Based on this numerical model, temperature of the hot zone can be predicted giving the measured temperature at any point of the sample located outside the central part. This avoids overloading the hot zone with inserted or welded thermocouples and thus improving the stress measuring during tensile test. Moreover, the extent of the hot zone can be determined more precisely and more easily without many heating tests. The knowledge of the hot zone length leads to a better calculation of strain.

Figure 5. Thermal profile along the sample axial direction: experimental data vs numerical model for various set points temperatures (● 500°C, ▲ 400°C, ■ 300°C, ■ 200°C).

6. CONCLUSIONS

In this work, we have developed a model intended to characterize the length of the hot zone formed on a sample heated by the Joule effect. The model determines the thermal profile in the axial direction of the sample set on the Gleeble machine. The modeling of the experimental setup included only the grips and the sample. Moreover, due to the existence of a plane of symmetry, we were able to make a geometric simplification that led to a half grips-sample assembling. To take into account the rest of the set up not represented in the model, we assumed an equivalent heat transfer coefficient at the grips/jaws interfaces.

For the simulation, some assumptions have been made, including perfect contact at grips/sample interfaces and negligible radiation losses on the free surfaces of the setup. This led to a low value for the first parameter while the second was set to zero. As far as the heat transfer along the free surfaces of the model is concerned, the convective coefficients were computed using empirical relations. The inverse method has been used to determine the equivalent heat transfer coefficient assumed at the grips/jaws interfaces as well as the thermal conductance at grips/sample interfaces. The procedure consisted of finding a couple of both
parameters which allowed us to fit the numerical thermal profile to the experimental one for a set point temperature of 500°C. To validate the model, simulations and heating tests were conducted for set points of 200°C, 300°C and 400°C. The good agreement between numerical and experimental thermal profile validated our model.

In this study, the thermo-physical properties are assumed to be temperature independent. The agreement between experimental data and numerical simulation suggests that the identified parameters do not vary much in the temperature interval considered. Indeed, for the temperature of the hot zone varying between 100°C and 500°C, the temperature at grips/sample interface did not vary much (between 25°C and 290°C) while the temperature of the grips did not exceed 40°C. For temperatures outside the set point interval considered, the variation of thermo-physical properties at grips/sample interfaces would probably need to be taken into account.

Non-contact techniques being often expensive or difficult to implement, numerical simulation has proved to be an alternative for characterization of the hot zone length. In the development of the numerical model, the identification of heat flow parameters was the main challenge. This study presented the inverse method as an alternative for the identification of parameters such as thermal contact conductance. This depends on so many factors such as the clearance between the contact surfaces as well as the average temperature and pressure at the interface that experimental techniques are not feasible.

ACKNOWLEDGEMENTS

This paper is written in the framework of the TOMOPIC project funded by the Picardie region. It includes the following partners: Université de Technologie de Compiègne, Montupet SA, CETIM and ESI Group SA. The technical support from the Centre Universitaire de Recherche sur l’Aluminium, Québec, is greatly acknowledged. The authors would like to thank the Professor X.-Grant Chen for his advises, D. Racine and E. Brideau for their technical help.

REFERENCES


THE DRBEM SOLUTION OF THE GENERALIZED MAGNETO-THERMO-VISCOELASTIC PROBLEMS IN 3D ANISOTROPIC FUNCTIONALLY GRADED SOLIDS

MOHAMED ABDELSABOUR FAHMY*

* Faculty of Computers and Informatics
Suez Canal University
Old Campus, El-Sheikh Zayed, 41522 Ismailia, Egypt
e-mail: mafahmy2001@yahoo.com, http://mafahmy.webs.com/

Key words: Magneto-Thermo-Viscoelastic Problems, Green-Naghdi Theory, Functionally Graded Solids, Anisotropic, Dual Reciprocity Boundary Element Method.

Abstract. Our current problem is an important due to its many applications in modern aeronautics, astronautics, soil dynamics, geophysics, plasma physics, nuclear reactors and high-energy particle accelerators. It is hard to find the analytical solution of a problem in a general case, therefore, an important number of engineering and mathematical papers devoted to the numerical solution have studied the special cases of current general problem. The basis of the boundary element method (BEM) is the transformation of the governing partial differential equations into a boundary integral equation. The presence of domain integrals in the BEM formulation implies domain discretization and this makes the BEM inefficient when compared with domain discretization techniques such as finite element method (FEM) or finite difference method (FDM). One of the most widely used techniques for converting the domain integral into a boundary one is the so-called dual reciprocity boundary element method (DRBEM), a numerical model based on the DRBEM taking into account the boundary and initial conditions is extended to solve the time-dependent generalized magneto-thermo-viscoelastic problems in anisotropic functionally graded solids. The unified formulation is tested through its application to the problem of a solid placed in a constant primary magnetic field. In the case of three-dimensional Cartesian coordinate system, a predictor-corrector implicit-implicit time integration algorithm was proposed and implemented for use with the DRBEM to obtain the temperature and displacement distributions with time in the context of the Green and Naghdi theory of type III. The results obtained are presented graphically in homogeneous and functionally graded materials. The examples that appear in the literature using the Meshless Local Petrov-Galerkin (MLPG) method are special cases of our general problem. Also, there are a lot of practical applications may be deduced as special cases from this general problem and may be implemented in commercial FEM software packages FlexPDE 6. In the considered special case, the results obtained with the DRBEM have been compared graphically with those obtained using the MLPG method and also the results obtained from the FlexPDE 6 are shown graphically in the same figures to confirm the validity of the proposed method. It can be seen from these figures that the DRBEM results are in excellent agreement with the results obtained by MLPG and FEM, thus confirming the accuracy of the DRBEM.
1 INTRODUCTION

Biot [1] introduced the theory of coupled thermoelasticity to overcome the first shortcoming in the theory of uncoupled thermoelasticity introduced by Duhamel [2] and Neuman [3] where it predicts two phenomena not compatible with physical observations. First, the equation of heat conduction of this theory does not contain any elastic terms. Second, the heat equation is of a parabolic type, predicting infinite speeds of propagation for heat waves. Later on, generalized theories of thermoelasticity were introduced in order to eliminate the shortcomings of the uncoupled thermoelasticity. Lord and Shulman (LS) [4] developed the theory of coupled thermoelasticity with one relaxation time by constructing a new law of heat conduction to replace the classical Fourier's law. This law contains the heat flux vector as well as its time derivative. It contains also new constant that acts as relaxation time. Since the heat equation of this theory is of the wave-type, it automatically ensures finite speeds of propagation for heat and elastic waves. Green and Lindsay (GL) [5] included a temperature rate among the constitutive variables to develop a temperature–rate-dependent thermoelasticity that does not violate the classical Fourier's law of heat conduction when the body under consideration has a center of symmetry; this theory also predicts a finite speed of heat propagation. This theory is known as the theory of thermoelasticity with two relaxation times. According to these theories, heat propagation should be viewed as a wave phenomenon rather than diffusion one. Relevant theoretical developments on the subject were made by Green and Naghdi (GN) [6, 7] they developed three models for generalized thermoelasticity of homogeneous isotropic materials which are labeled as model I, II and III. These theories of thermoelasticity LS, GL and GN theories are known as the generalized theories of thermoelasticity with finite thermal wave speed. In general, it is not easy to obtain analytical solutions of generalized magneto-thermo-visco-elastic problems in anisotropic functionally graded solids. Therefore, an important number of engineering and mathematical papers devoted to the numerical solution have studied the overall behavior of such materials [8-15].

The first step of the boundary element method (BEM) is the transformation of the physical problem at hand to an integral equation. Due to the pioneer work by Nardini and Brebbia [16], the boundary element method in conjunction with dual reciprocity method (DRM) and radial basis function (RBF), can now be used to obtain approximate solutions of general partial differential equation (PDE) systems. This strategy is often called as the dual reciprocity boundary element method (DRBEM). This method was initially developed in the context of two-dimensional (2D) elastodynamics and has been extended to deal with a variety of problems wherein the domain integral may account for linear-nonlinear static-dynamic effects. The DRBEM has been highly successful in a very wide range of engineering applications, including acoustics, aeroacoustics, aerodynamics, fluid dynamics, fracture analysis, geomechanics, elasticity and heat transfer. A more extensive historical review and applications of dual reciprocity boundary element method may be found in Refs. [17-20].

The main objective of this paper is to study the generalized magneto-thermo-viscoelastic problems in anisotropic solid of functionally graded material (FGM) in the context of the Green and Naghdi theory of type III. A predictor-corrector implicit-implicit time integration algorithm was developed and implemented for use with the dual reciprocity boundary element method (DRBEM) to obtain the solution for the temperature and displacement fields. The results obtained are presented graphically in homogeneous and functionally graded solids.
2 FORMULATION OF THE PROBLEM

Consider the coordinate system $Oxyz$ as shown in Fig. 1. We shall consider a functionally graded anisotropic solid placed in a primary magnetic field $H_0$ acting in the direction of the $z$-axis and occupies the region $R = \{(x, y, z): 0 < x < \gamma, 0 < y < \beta, 0 < z < \alpha\}$ with graded material properties in the thickness direction ($x$-axis).

The governing equations of the generalized magneto-thermo-viscoelastic problems in the context of the Green and Naghdi theory of type III are given by [21]:

$$\sigma_{ab} + \tau_{ab} = \rho (x + 1)^m \ddot{u}_a$$  \hspace{1cm} (1)

$$\sigma_{ab} = \mathfrak{N}(x + 1)^m \left[ C_{abfg} u_{fg} - \beta_{ab}(T - T_0 + \tau_1 T) \right], \mathfrak{N} = \left( 1 + v_0 \frac{\partial}{\partial T} \right)$$  \hspace{1cm} (2)

$$\tau_{ab} = \mu (x + 1)^m \left( \bar{h}_a H_b + \bar{h}_b H_a - \delta_{ba}(\bar{h}_f H_f) \right), \quad \bar{h}_a = \left( \mathfrak{V} \times (\mathfrak{u} \times \mathfrak{H}) \right)_a$$  \hspace{1cm} (3)

$$k_{ab}^* T_{ab} = -k_{ab} T_{ab} + \beta_{ab} T_0 \ddot{u}_{ab} + \rho c (x + 1)^m \ddot{T}, \quad (k_{ab} = k_{ba}), (k_{12})^2 - k_{11} k_{22} < 0$$  \hspace{1cm} (4)

where $\sigma_{ab}$ is the mechanical stress tensor, $\tau_{ab}$ is the Maxwell’s electromagnetic stress tensor, $u_a$ is the displacement, $T$ is the temperature, $C_{abfg}$ ($C_{abgf} = C_{fagb} = C_{bfag}$) and $\beta_{ab}$ ($\beta_{ab} = \beta_{ba}$) are respectively, the constant elastic moduli and stress-temperature coefficients of the anisotropic medium, $\mathfrak{N}$ is the viscoelastic material constant, $v_0$ is the viscoelastic relaxation time, $\mu$ is the magnetic permeability, $\bar{h}$ is the perturbed magnetic field, $k_{ab}$ are the thermal conductivity coefficients, $\rho$ is the density, $c$ is the specific heat capacity, $\tau$ is the time $\tau_1$ is the mechanical relaxation time, $m$ is a dimensionless constant and the traction vector can be written as

$$\tau_a = \sigma_{ab} n_b = \mathfrak{N}(x + 1)^m \left( C_{abfg} u_{fg} - \beta_{ab}(T - T_0 + \tau_1 T) \right) n_b$$  \hspace{1cm} (5)

3 NUMERICAL IMPLEMENTATION

The field equations (1) and (4) can now be written in operator form as follows [22]:

$$L_{gb} u_f = f_{gb}$$  \hspace{1cm} (6)

$$L_{ab} T = f_{ab}$$  \hspace{1cm} (7)

in which

$$L_{gb} = D_{abf} \frac{\partial}{\partial x_b}, \quad f_{gb} = \rho \ddot{u}_a - \left( D_a T + D_{af} + \Lambda D_{af} \right)$$  \hspace{1cm} (8)

$$L_{ab} = k_{ab}^* \frac{\partial}{\partial x_a} \frac{\partial}{\partial x_b}, \quad f_{ab} = -k_{ab} \ddot{T}_{ab} + \beta_{ab} T_0 \ddot{u}_{ab} + \rho c (x + 1)^m \ddot{T}$$  \hspace{1cm} (9)

where

$$D_{abf} = C_{abfg} e \mathfrak{e}, \mathfrak{e} = \frac{\partial}{\partial x_g}, D_{af} = \mu H_0^2 \left( \frac{\partial}{\partial x_a} + \delta_{a1} \Lambda \right) \frac{\partial}{\partial x_f}, \Lambda = \frac{m}{x + 1},$$

$$D_a = -\beta_{ab} \left( \frac{\partial}{\partial x_b} + \delta_{b1} \Lambda + \tau_1 \left( \frac{\partial}{\partial x_b} + \Lambda \right) \frac{\partial}{\partial \tau} \right)$$

864
Application of the weighted residual method (WRM) to Eq. (6), which, after integration by parts and use of the sifting property of the Dirac distribution, is written in the form of the following elastic integral representation formula

\[ u_d(\xi) = \int_C \left( u_{da}^* \xi_a - t_{da}^* u_a + u_{da}^* \beta_{ab} T n_b \right) dC - \int_R f_{gb} u_{da}^* dR \]  \quad (10)

By implementing the WRM and integration by parts using the sifting property, we obtain from (7) the thermal integral representation formula

\[ T(\xi) = \int_C (q^* T - q T^*) dC - \int_R f_{ab} T^* dR \]  \quad (11)

The thermoelastic representation formula (28) can be written in contracted notation as:

\[ U_D(\xi) = \int_C \left( U_{DA}^* T_A - \tilde{T}_{DA}^* U_A \right) dC - \int_R U_{DA} S_A^* dR \]  \quad (12)

To transform the domain integral in (12) to the boundary, we approximate the source vector \( S_A \) in the domain by a series of given tensor functions \( f_{AN}^q \) and unknown coefficients \( \alpha_{N}^q \)

\[ S_A \approx \sum_{q=1}^{E} f_{AN}^q \alpha_{N}^q \]  \quad (13)

Thus, the thermoelastic representation formula (12) can be written in the following form

\[ U_D(\xi) = \int_C \left( U_{DA}^* T_A - \tilde{T}_{DA}^* U_A \right) dC - \sum_{q=1}^{N} \int_R U_{DA} f_{AN}^q dR \alpha_{N}^q \]  \quad (14)

The dual representation formulae of elastic and thermal fields can be combined as follows

\[ U_{DN}^{q}(\xi) = \int_C \left( U_{DA}^* T_{AN}^q - T_{DA} U_{AN}^q \right) dC - \int_R U_{DA} f_{AN}^q dR \]  \quad (15)

With the substitution of (15) into (14), the dual reciprocity representation formula of coupled thermoelasticity can be expressed as follows

\[ U_D(\xi) = \int_C \left( U_{DA}^* T_A - \tilde{T}_{DA}^* U_A \right) dC + \sum_{q=1}^{N} \left( U_{DN}^{q}(\xi) + \int_C \left( T_{DA} U_{AN}^q - U_{DA} T_{AN}^q \right) dC \right) \alpha_{N}^q \]  \quad (16)

According to the steps described in Fahmy [23], the dual reciprocity boundary integral equation (16) can be written in the following system of equations

\[ \tilde{\zeta} \ddot{u} - \eta \ddot{\zeta} = (\zeta \ddot{U} - \eta \ddot{\psi}) \alpha \]  \quad (17)
The generalized displacements $U_F$ and velocities $\dot{U}_F$ are approximated by a series of tensor functions $f_{PD}^q$ and unknown coefficients $\gamma_D^q$ and $\tilde{\gamma}_D^q$ as follows [24]:

$$U_F \approx \sum_{q=1}^{N} f_{PD}^q(x) \gamma_D^q, \quad \dot{U}_F \approx \sum_{q=1}^{N} f_{PD}^q(x) \tilde{\gamma}_D^q \quad (18)$$

The same point collocation procedure described in Gaul, et al. [25] can be applied to (13) and (18). This leads to the following system of equations

$$\ddot{S} = J\alpha, \quad U = J'\gamma, \quad \dot{U} = J'\dot{\gamma} \quad (19)$$

Now, the coefficients $\alpha$ can be expressed in terms of nodal values of the unknown displacements $U$, velocities $\dot{U}$ and accelerations $\ddot{U}$ as follows:

$$\alpha = J^{-1} \left( [B^T J^{-1} - (D_{af} + \Lambda D_{alf})] U + \left( k_{ab} \frac{\partial}{\partial x_a} \frac{\partial}{\partial x_b} \delta_{AF} \right) \dot{U} + \left[ \tilde{A} - c \rho (x + 1)^m \delta_{AF} \right] \ddot{U} \right) \quad (20)$$

where $B^T = \begin{cases} -D_a & A = 1, 2, 3; F = 4 \\ 0 & \text{otherwise} \end{cases}$, $A = 1, 2, 3; F = 1, 2, 3$, $\ddot{\alpha} = \begin{cases} \rho \beta_{f} & A = 1, 2, 3; F = 4 \\ -T_{0} \beta_{f} & A = 4; F = 4 \end{cases}$, $\delta_{AF} = \begin{cases} -c \rho (x + 1)^m & A = 4; F = 4 \\ 0 & \text{otherwise} \end{cases}$

An implicit-implicit staggered algorithm was developed and implemented for use with the DRBEM for solving the governing equations which may now be written in a more convenient form after substitution of Eq. (20) into Eq. (17) as follows:

$$\tilde{M} \ddot{U} + \tilde{K} \dot{U} + \tilde{R} U = \tilde{Q} \quad (21)$$

$$\tilde{X} \ddot{T} + \tilde{A} T + \tilde{B} T = \tilde{Z} \ddot{U} \quad (22)$$

where $\tilde{M} = V \tilde{A}$, $\tilde{K} = V \left( k_{ab} \frac{\partial}{\partial x_a} \frac{\partial}{\partial x_b} \right) \delta_{AF}$, $\tilde{R} = \tilde{\xi} + VB^T J^{-1} - (D_{af} + \Lambda D_{alf})$, $\tilde{Q} = \eta T$, $\tilde{X} = -\rho c (x + 1)^m$, $\tilde{A} = k_{ab} \frac{\partial}{\partial x_a} \frac{\partial}{\partial x_b} \delta_{AF}$, $\tilde{B} = k_{ab} \delta_{AF}$, $\tilde{Z} = T_{0} \beta_{f}$, $V = (\eta \tilde{\beta} - \tilde{\xi} \ddot{U}) J^{-1}$.

where $V, \tilde{M}, \tilde{K}$ and $\tilde{R}$ represent the volume, mass, damping and stiffness matrices, respectively, $\ddot{U}, \dot{U}, U, T$ and $\tilde{Q}$ represent the acceleration, velocity, displacement, temperature and external force vectors, respectively, $\tilde{A}$ and $\tilde{B}$ are respectively the capacity and conductivity matrices, $\tilde{X}$ is a vector of new material constants and $\tilde{Z}$ is coupling matrix.

Hence the governing equations lead to the following coupled system of equations [26]:

$$\tilde{M} \ddot{U}_{n+1} + \tilde{K} \dot{U}_{n+1} + \tilde{R} U_{n+1} = \tilde{Q}^p_{n+1} \quad (23)$$

$$\tilde{X} \ddot{T}_{n+1} + \tilde{A} \dot{T}_{n+1} + \tilde{B} T_{n+1} = \tilde{Z} \ddot{U}_{n+1} \quad (24)$$

where $\tilde{Q}^p_{n+1} = \eta T_{n+1}^p$ and $T_{n+1}^p$ is the predicted temperature. For further details, see the recent works of Fahmy [27-29].
Integrating Eq. (21) with the use of trapezoidal rule and Eq. (23), we have

$$
\dot{U}_{n+1} = \dot{U}_n + \frac{\Delta \tau}{2} \left[ \dot{U}_n + \dot{\tilde{M}}^{-1} \left( \dot{\tilde{Q}}_{n+1} - \tilde{K} \dot{U}_{n+1} \right) \right]
$$

(25)

$$
U_{n+1} = U_n + \Delta \tau \dot{U}_n + \frac{\Delta \tau^2}{4} \left[ \dot{U}_n + \dot{\tilde{M}}^{-1} \left( \dot{\tilde{Q}}_{n+1} - \tilde{K} \dot{U}_{n+1} \right) \right]
$$

(26)

From Eq. (25) we have

$$
\dot{U}_{n+1} = \bar{Y}^{-1} \left[ \dot{U}_n + \frac{\Delta \tau}{2} \left[ \dot{U}_n + \dot{\tilde{M}}^{-1} \left( \dot{\tilde{Q}}_{n+1} - \tilde{K} \dot{U}_{n+1} \right) \right] \right]
$$

(27)

where $\bar{Y} = \left( I + \frac{\Delta \tau}{2} \dot{\tilde{M}}^{-1} \tilde{\Gamma} \right)$

Substituting from Eq. (27) into Eq. (26), we derive

$$
U_{n+1} = U_n + \Delta \tau \dot{U}_n + \frac{\Delta \tau^2}{4} \left[ \dot{U}_n + \dot{\tilde{M}}^{-1} \left( \dot{\tilde{Q}}_{n+1} - \tilde{K} \dot{U}_{n+1} \right) \right] - \tilde{K} \dot{U}_{n+1}
$$

(28)

From Eq. (25) we get

$$
\dot{U}_{n+1} = \hat{M}^{-1} \left[ \dot{\tilde{Q}}_{n+1} - \tilde{\Gamma} \bar{Y}^{-1} \left[ \dot{U}_n + \frac{\Delta \tau}{2} \left[ \dot{U}_n + \dot{\tilde{M}}^{-1} \left( \dot{\tilde{Q}}_{n+1} - \tilde{K} \dot{U}_{n+1} \right) \right] \right] \right] - \tilde{K} \dot{U}_{n+1}
$$

(29)

Integrating the heat equation (22) using the trapezoidal rule, and Eq. (24) we get

$$
\dot{T}_{n+1} = \dot{T}_n + \frac{\Delta \tau}{2} \left( \dot{T}_{n+1} + \dot{T}_n \right) = \dot{T}_n + \frac{\Delta \tau}{2} \left( \tilde{X}^{-1} \left[ \mathcal{Z} \dot{U}_{n+1} - \tilde{A} T_{n+1} - \mathcal{B} T_{n+1} \right] + \dot{T}_n \right)
$$

(30)

$$
T_{n+1} = T_n + \Delta \tau \dot{T}_n + \frac{\Delta \tau^2}{4} \left( \tilde{T}_n + \tilde{X}^{-1} \left[ \mathcal{Z} \dot{U}_{n+1} - \tilde{A} T_{n+1} - \mathcal{B} T_{n+1} \right] \right)
$$

(31)

From Eq. (30) we get

$$
\dot{T}_{n+1} = Y^{-1} \left[ \dot{T}_n + \frac{\Delta \tau}{2} \left( \tilde{X}^{-1} \left[ \mathcal{Z} \dot{U}_{n+1} - \tilde{B} T_{n+1} \right] + \dot{T}_n \right) \right]
$$

(32)

where $Y = \left( I + \frac{\Delta \tau}{2} \tilde{A} \tilde{X}^{-1} \right)$

Substituting from Eq. (32) into Eq. (31), we have

$$
T_{n+1} = T_n + \Delta \tau \dot{T}_n + \frac{\Delta \tau^2}{4} \left( \tilde{T}_n + \tilde{X}^{-1} \left[ \mathcal{Z} \dot{U}_{n+1} \right] \right) - \tilde{A} \left( Y^{-1} \left[ \dot{T}_n + \frac{\Delta \tau}{2} \left( \tilde{X}^{-1} \left[ \mathcal{Z} \dot{U}_{n+1} - \tilde{B} T_{n+1} \right] + \dot{T}_n \right) \right] - \mathcal{B} T_{n+1} \right)
$$

(33)

Substituting $T_{n+1}$ from Eq. (32) into Eq. (24) we obtain
Now, a displacement predicted staggered algorithm for the solution of (28) and (33) is:
(1) Predict the displacement field: \( \hat{u}_{n+1}^P = U_n \).
(2) Substituting for \( \hat{U}_{n+1}^P \) and \( \hat{U}_{n+1}^R \) from equations (25) and (23) respectively in Eq. (33) and solve the resulted equation for the temperature field.
(3) correct the displacement field using the computed temperature field for the Eq. (28).
(4) compute \( \hat{U}_{n+1}^P, \hat{U}_{n+1}^R, \hat{t}_{n+1} \) and \( \hat{t}_{n+1} \) from Eqs. (27), (29), (32) and (34) respectively.

4 NUMERICAL RESULTS AND DISCUSSION

For the purpose of numerical computations, the physical constants are as follows:

Elasticity tensor

\[
C_{abfg} = \begin{bmatrix}
17.77 & 3.78 & 3.76 & 0.24 & -0.28 & 0.03 \\
3.78 & 19.45 & 4.13 & 0 & 0 & 1.13 \\
3.76 & 4.13 & 21.79 & 0 & 0 & 0.38 \\
0 & 0 & 0 & 8.30 & 0.66 & 0 \\
0 & 0 & 0 & 0.66 & 7.62 & 0 \\
0.03 & 1.13 & 0.38 & 0 & 0 & 7.77
\end{bmatrix} \text{ GPa}
\]

Mechanical temperature coefficient and tensor of thermal conductivity are

\[
\beta_{ab} = \begin{bmatrix}
0.001 & 0.02 & 0 \\
0.02 & 0.006 & 0 \\
0 & 0 & 0.05
\end{bmatrix} \cdot 10^6 \text{ N/Km}^2, \quad k_{ab} = \begin{bmatrix}
1 & 0.1 & 0.2 \\
0.1 & 1.1 & 0.15 \\
0.2 & 0.15 & 0.9
\end{bmatrix} \text{ W/km}
\]

Mass density \( \rho = 2216 \text{ kg/m}^3 \) and heat capacity \( c = 0.1 \text{ J/(kg K)} \), \( H_0 = 1000000 \text{ Oersted} \), \( \mu = 0.5 \text{ Gauss/Oersted} \), \( \kappa = 2 \), \( h = 2 \), \( \Delta \tau = 0.0001 \).

The initial and boundary conditions considered in the calculations are:

at \( \tau = 0 \) \quad \sigma_{ab} = \hat{\sigma}_{ab} = \tau_{ab} = \hat{\tau}_{ab} = 0, \quad T = 0 \quad (35)

at \( x = 0 \) \quad \frac{\partial u_1}{\partial x} = \frac{\partial u_2}{\partial x} = 0, \quad \frac{\partial \tau}{\partial x} = 0, \quad \text{at} \ x = y \quad \frac{\partial u_1}{\partial x} = \frac{\partial u_2}{\partial x} = 0, \quad \frac{\partial T}{\partial x} = 0 \quad (36)

at \( y = 0 \) \quad \frac{\partial u_1}{\partial y} = \frac{\partial u_2}{\partial y} = 0, \quad \frac{\partial \tau}{\partial y} = 0, \quad \text{at} \ y = \beta \quad \frac{\partial u_1}{\partial y} = \frac{\partial u_2}{\partial y} = 0, \quad \frac{\partial T}{\partial y} = 0 \quad (37)

at \( z = 0 \) \quad \frac{\partial u_1}{\partial z} = \frac{\partial u_2}{\partial z} = 0, \quad \frac{\partial \tau}{\partial z} = 0, \quad \text{at} \ z = \alpha \quad \frac{\partial u_1}{\partial z} = \frac{\partial u_2}{\partial z} = 0, \quad \frac{\partial T}{\partial z} = 0 \quad (38)

The results obtained are presented graphically in homogeneous solid (HS) and functionally graded solid (FGS), where we assumed that \( m = 0.0 \) in the HS case and \( m = 0.5 \) in the FGS case.

Figures 2 and 3 show the variation of the temperature \( T \) with time \( \tau \). We can conclude from these figures that the maximum temperature \( T \) occurs at \( \tau = 3 \) for FGS case. But it occurs at \( \tau = 6 \) for HS case.
Figures 4 and 5 illustrate the variation of the displacement $u_1$ with time $\tau$. It can be seen that the maximum displacement $u_1$ occurs at $\tau = 3$ for the FGS case. But it occurs at $\tau = 6$ for the HS case.

Figures 6 and 7 show the variation of the displacement $u_2$ with time $\tau$. It can be seen from the figures that the maximum displacement occurs in the FGS case. It can be noted from these figures that the displacement $u_2$ appears in an intense oscillation about a zero value, where the wave fronts collide with each other for the FGS case and may accumulate for the HS case.

Figures 8 and 9 illustrate the variation of the displacement $u_3$ with time $\tau$. It can be seen that the maximum displacement $u_3$ occurs at $\tau = 27.5$ for the HS case. But it occurs at $\tau = 18$ for the FGS case.

The present proposed method should be applicable to any three-dimensional generalized magneto-thermo-viscoelastic problem. The examples that appear in the literature using the Meshless Local Petrov-Galerkin (MLPG) method are special cases of our general problem. Also, there are a lot of practical applications may be deduced as special cases from this general problem and may be implemented in commercial FEM software packages FlexPDE 6. In the considered special case of Hosseini et al. [30] who solved the special case from this study in the context of Green and Naghdi (GN) theory of type II, the results obtained with the DRBEM have been compared graphically, in figure 10, with those obtained using the MLPG method of Hosseini et al. [30] and also the results obtained from the FlexPDE 6 are shown graphically in the same figure to confirm the validity of the proposed method. It can be seen from this figure that the DRBEM results are in excellent agreement with the results obtained by MLPG and FEM, thus confirming the accuracy of the DRBEM.

5 CONCLUSION

In this paper, an algorithm to solve 3D generalized magneto-thermo-viscoelastic problems in anisotropic FGS was proposed and implemented for use with the DRBEM to obtain a solution for the displacement and temperature fields for both HS and FGS. Since there has been no previous solution is known for the current problem, the accuracy of the proposed algorithm was examined and confirmed by comparing the obtained results with those known previously from special cases of the present study. The DRBEM results are in excellent agreement with the special cases results obtained by MLPG method and FDM. So the proposed algorithm is very useful in solving 3D complex generalized magneto-thermo-viscoelastic problems.

REFERENCES

Mohamed Abdelsabour Fahmy


Fig. 5. Variation of the displacement $u_1$ with time $\tau$ (FGS).

Fig. 6. Variation of the displacement $u_2$ with time $\tau$ (HS).

Fig. 7. Variation of the displacement $u_3$ with time $\tau$ (FGS).

Fig. 8. Variation of the displacement $u_3$ with time $\tau$ (HS).

Fig. 9. Variation of the displacement $u_1$ with time $\tau$ (FGS).

Fig. 10. Variation of the temperature $T$ with time $\tau$ for three methods: DRBEM, MLPG, FEM.
1 INTRODUCTION

The competitiveness of renewable energy sources for heat generation and the efficient use of industrial process heat require the integration of reliable thermal energy storage systems. Aside from sensible and latent heat storage concepts thermochemical heat storage systems are an attractive option due to their high storage densities and low losses [8]. Their development is still at a laboratory stage and can be greatly supported by advanced theoretical and numerical models that facilitate process understanding and optimisation [11]. Here we present a model based on the Theory of Porous Media and the assumption of local thermal non-equilibrium for a solid reactor with a direct heat transfer gas carrying a gaseous reactant. Appropriate kinematic and balance relations are established to account for mass and heat transport as well as phase and mass transfer due to exo- and endothermic chemical reactions. They are supplemented by constitutive relations that are derived from and constrained by the Clausius-Duhem inequality in a thermodynamically consistent manner. The highly nonlinear governing equations are cast into the weak form and discretised in space following a standard Galerkin approach and in time using a generalized single-step algorithm. The model has been implemented into the scientific open source finite element code OpenGeoSys designed for strongly coupled THMC processes [7]. A calcium oxide / calcium hydroxide reaction system through which a nitrogen-steam gas
mixture passes is taken as a model system. The simulations show a fast and a slow reaction wave travelling through the reactor that are limited by system equilibrium temperature / heat transport and the degree of conversion of the solid material, respectively. It is shown how the system performance, e.g. in terms of outlet heat profiles, can be adapted to specific needs by varying process and material parameters.

2 METHODS

2.1 General balance equations

The model development was based on the well established Theory of Porous Media (TPM) [5, 3]. Without additional derivation the general local balance relations employed here will be listed in this section. The balance of mass for a phase $\alpha$ subject to mass exchange with other phases is

$$ \frac{d\alpha}{dt} \rho_\alpha + \rho_\alpha \text{div} \; v_\alpha = \hat{\rho}_\alpha \quad \text{with} \quad \sum_\alpha \hat{\rho}_\alpha = 0 \tag{1} $$

where $(\bullet)'_\alpha = \frac{d(\bullet)}{dt} = \frac{\partial (\bullet)}{\partial t} + \text{grad} \, (\bullet) \cdot v_\alpha$ denotes a material time derivative following the motion of the $\alpha^{th}$ constituent (velocity $v_\alpha$, apparent density $\rho_\alpha$). The symbol $(\bullet)$ denotes production terms due to internal interactions.

The quasistatic balance of linear momentum in the absence of body forces reads

$$ \text{div} \; \sigma_\alpha + \dot{s}_\alpha = \hat{\rho}_\alpha \, v_\alpha \quad \text{with} \quad \sum_\alpha \dot{s}_\alpha = \sum_\alpha (\hat{p}_\alpha + \hat{\rho}_\alpha \, v_\alpha) = 0 \tag{2} $$

with the partial Cauchy stress tensor $\sigma_\alpha$, the overall and direct momentum production terms $\dot{s}_\alpha$ and $\hat{\rho}_\alpha$, respectively.

Due to the presence of mass exchange the first law of thermodynamics was expressed in terms of enthalpy rather than inner energy. For that purpose, the Cauchy stress and rate of deformation tensors were split into deviatoric and hydrostatic parts [10]

$$ \sigma_\alpha = \sigma_\alpha^D - p_\alpha \, I \quad \text{with} \quad p_\alpha = -\frac{1}{3}(\sigma_\alpha : I) \tag{3} $$

$$ d_\alpha = d_\alpha^D + \frac{1}{3}(d_\alpha : I) \, I \tag{4} $$

The enthalpy balance (with $h_\alpha$ the specific enthalpy and $q_\alpha$ as the partial heat flux vector)

$$ \rho_\alpha (h_\alpha)'_\alpha - (p_\alpha)'_\alpha - \sigma_\alpha^D : d_\alpha^D - \rho_\alpha \, r_\alpha + \text{div} \; q_\alpha = \dot{u}_\alpha - \hat{\rho}_\alpha \left( h_\alpha + \frac{1}{2} v_\alpha \cdot v_\alpha \right) - v_\alpha \cdot \hat{p}_\alpha \tag{5} $$

is subject to a constraint on the inner energy production terms

$$ \sum_\alpha \dot{u}_\alpha = 0 \tag{6} $$
Constitutive relations are required to close the above system of equations. The Clausius-Duhem inequality was used in the form

$$\sum_{\alpha} \frac{1}{T_{\alpha}} \left[ -\rho_{\alpha}[(\psi_{s})'_{\alpha} + (T_{s})'_{\alpha} \eta_{\alpha}] - \hat{\rho}_{\alpha} \left( \psi_{\alpha} + \frac{p_{\alpha}}{\rho_{\alpha}} + \frac{1}{2} v_{\alpha} \cdot v_{\alpha} \right) + \sigma_{\alpha}^{D} : t_{\alpha}^{D} + p_{\alpha} \frac{(\phi_{s})'_{\alpha}}{\phi_{\alpha}} + p_{\alpha} \frac{(\rho_{s})'_{\alpha}}{\rho_{s}} - \hat{p}_{\alpha} \cdot v_{\alpha} - \frac{1}{T_{\alpha}} q_{\alpha} \cdot \text{grad} T_{\alpha} + \hat{u}_{\alpha} \right] \geq 0$$

(7)

to ensure the thermodynamic consistency of the relations used in this study. Here, $T_{\alpha}$ is the absolute temperature, $\psi_{\alpha}$ the specific Helmholtz free energy, $\eta_{\alpha}$ the specific entropy, $\phi_{\alpha}$ the volume fraction and $\rho_{\alpha R} = \rho_{\alpha}/\phi_{\alpha}$ the real density of phase $\alpha$. Specifically, a porous solid skeleton saturated by a gas was considered here. The saturation constraint

$$0 = \sum_{\alpha=S,G} (\phi_{\alpha})'_{\alpha} - \text{grad} \phi_{G} \cdot (v_{G} - v_{S})$$

(8)

was multiplied with a Lagrange multiplier and the inverse reference temperature $[4]$

$$\frac{1}{T_{R}} = \frac{1}{2} \frac{T_{S} + T_{G}}{T_{S} T_{G}}$$

(9)

and added to the entropy inequality. Additionally, the gas was considered a mixture of ideal gases (components $\zeta$) with $\phi_{G} = \phi_{\zeta}$ and $T_{G} = T_{\zeta}$ [6]. The final version of the entropy inequality thus reads

$$0 \leq \frac{1}{T_{S}} \left[ -\rho_{s}[(\psi_{s})'_{s} + (T_{s})'_{s} \eta_{s}] - \hat{\rho}_{s} \left( \psi_{s} + \frac{p_{s}}{\rho_{s}} + \frac{1}{2} v_{s} \cdot v_{s} \right) + \sigma_{s}^{D} : t_{s}^{D} + p_{s} \frac{(\phi_{s})'_{s}}{\phi_{s}} + p_{s} \frac{(\rho_{s})'_{s}}{\rho_{s}} - \hat{p}_{s} \cdot v_{s} - \frac{1}{T_{s}} q_{s} \cdot \text{grad} T_{s} + \hat{u}_{s} \right] + \frac{1}{T_{G}} \sum_{\zeta} \left[ -\rho_{\zeta}[(\psi_{\zeta})'_{\zeta} + (T_{\zeta})'_{\zeta} \eta_{\zeta}] - \hat{\rho}_{\zeta} \left( \psi_{\zeta} + \frac{p_{\zeta}}{\rho_{\zeta}} + \frac{1}{2} v_{\zeta} \cdot v_{\zeta} \right) + \sigma_{\zeta}^{D} : t_{\zeta}^{D} + p_{\zeta} \frac{(\phi_{\zeta})'_{G}}{\phi_{G}} + p_{\zeta} \frac{\phi_{\zeta}}{\phi_{G}} \text{grad} \phi_{G} d_{\zeta} + p_{\zeta} \frac{(\rho_{s})'_{\zeta}}{\rho_{s R}} - \hat{p}_{\zeta} \cdot v_{\zeta} - \frac{1}{T_{G}} q_{\zeta} \cdot \text{grad} T_{G} + \hat{u}_{\zeta} \right] + \frac{1}{T_{R}} \lambda [(\phi_{s})'_{s} + (\phi_{G})'_{G} - \text{grad} \phi_{G} \cdot w_{G}]$$

(10)

where the seepage velocity $w_{G} = v_{G} - v_{S}$ of the gas relative to the solid and the diffusion velocity $d_{\zeta} = v_{\zeta} - v_{G}$ of a gas constituent in the overall gas phase have been introduced. Further details as well as the evaluation of the above relation can be found in [9].

### 2.2 Governing equations

The general model above was specified under the following assumptions: Local thermal non-equilibrium albeit with small deviations from equilibrium; binary gas mixture of nitrogen and
steam; rigid solid skeleton; constant porosity; Fourier's law for heat conduction. The solid gas reaction
\[
\text{CaO (s) + H}_2\text{O (g) } \rightleftharpoons \text{Ca(OH)}_2 \text{ (s)}
\]
was taken as an example system.

The solid mass balance simply reads
\[
(1 - \phi_G) \frac{\partial \rho_{SR}}{\partial t} = \hat{\rho}_S
\]
(11)

The reaction kinetics used to determine \(\hat{\rho}_S\) were dependent on the degree of conversion from calcium oxide to calcium hydroxide as well as the equilibrium drop that was calculated from the Clausius-Clapeyron relation for this particular reaction:
\[
\hat{\rho}_S = - (1 - \phi_G) x_m V (\rho_{\text{Ca(OH)}_2 R} - \rho_{SR}) k^H_R \frac{T_S - T_{eq}}{T_{eq}}
\]
(12)
where \(k^H_R\) is a rate constant and \(T_{eq}\) the equilibrium temperature.

The gas mass balance is found as
\[
\phi_G \frac{\partial \rho_{GR}}{\partial t} + \text{div} (\phi_G \rho_{GR} v_G) = - \hat{\rho}_S
\]
(13)

Under certain linearity assumptions for the momentum production terms [9, 6] the linear momentum balance of the gas yields the extended Darcy-like expression for the filter velocity \(\tilde{w}_G\) (neglecting diffusion-reaction contributions)
\[
\tilde{w}_G = - k_S \mu_v \left[ p \phi_G \left(1 - \frac{T_S}{T_G} \right) \text{grad} \phi_G + \text{grad} p \right]
\]
(14)
with the viscosity \(\mu_v\) and the intrinsic permeability tensor \(k_S\).

The heat transport equation for the gas reads
\[
\phi_G \rho_{GR} c_{pG} \frac{\partial T_G}{\partial t} + \phi_G \rho_{GR} c_{pG} \text{grad} T_G \cdot v_G = h_{SG} (T_S - T_G) + \text{div} [\phi_G \lambda_{GR} \text{grad} T_G] +
+ \phi_G \frac{\partial p}{\partial t} + \phi_G \text{grad} p \cdot v_G
\]
(15)
with the specific heat capacity \(c_{pG}\), heat conductivity \(\lambda_{GR}\) and the volumetric interphase heat transfer coefficient \(h_{SG}\). Similarly, the heat transport equation for the solid is
\[
\rho_s c_{ps} \frac{\partial T_s}{\partial t} = \text{div} (\lambda_s \text{grad} T_s) + h_{SG} (T_G - T_s) - \tilde{p}_G \cdot w_G + \hat{\rho}_S [\Delta h + c_{pV} (T_G - T_s)]
\]
(16)
with \(\tilde{p}_G = - \frac{\phi_G^2 \mu_v}{k_S} w_G + \frac{T_S}{T_G} p \text{grad} \phi_G = \text{grad} (\phi_G p)\)
with the specific heat capacity of the solid \(c_{pS}\) and vapour \(c_{pV}\), heat conductivity \(\lambda_{SR}\), reaction enthalpy per unit mass \(\Delta h\) and seepage velocity \(w_G\). Gas properties such as \(c_{pG}\), \(\mu_V\), \(\rho_{GR}\) etc. were modelled to depend on gas composition [2].

Finally, the mass balance for the reactive vapour component was written in terms of the vapour mass fraction \(x_{mV}\) as

\[
\phi_G \rho_{GR} \frac{\partial x_{mV}}{\partial t} - \text{div} (\rho_{GR} \mathbf{D} \text{grad} x_{mV}) + \phi_G \rho_{GR} \text{grad} x_{mV} \cdot \mathbf{v}_G = -(1 - x_{mV}) \hat{\rho}_S \tag{17}
\]

Although the analysis based on Eq. (10) yielded the gradient of the mole fractions \(x_n\) as the driving force for diffusion [9, 6], the simplification used in the component mass balance is based on the assumption of a constant molar mass of the gas mixture:

\[
\phi_G \mathbf{d}_\zeta = -x_n^{-1} \mathbf{D} \text{grad} x_n \xrightarrow{M_G=\text{const.}} \phi_G \rho_{GR} \mathbf{d}_\zeta = -\rho_{GR} \mathbf{D} \text{grad} x_m \zeta \tag{18}
\]

This assumption is made in the present model only in the context of diffusion as it simplifies the treatment and diffusion is of minor importance compared to other mass transport effects in the applications under study. Anywhere else the mixture molar mass is calculated depending on the component mole fractions.

### 2.3 Finite element implementation

The primary variables \(p, T_G, T_S\) and \(x_{mV}\) were selected. Equations (13), (15), (16) and (17) were transformed into their weak forms and discretised following a standard Galerkin approach. Fluid velocities were substituted using Eq. (14). First order shape functions were employed. The reaction rate \(\hat{\rho}_S\) (see Eqs. (11) and (12)) was determined with a local ODE solver. The non-symmetric system of equations

\[
\begin{pmatrix}
M_{pp} & M_{pT}^G & 0 & M_{px} \\
M_{pT} & M_{TT}^G & 0 & 0 \\
0 & 0 & M_{TT}^S & 0 \\
0 & 0 & 0 & M_{cc}
\end{pmatrix}
\begin{pmatrix}
\dot{p} \\
\dot{T}_G \\
\dot{T}_S \\
\dot{x}_{mV}
\end{pmatrix}
+ \begin{pmatrix}
L_{pp} + \hat{A}_{pp} \\
A_{TT}^G + A_{TT}^S + L_{TT}^G \\
0 \\
0
\end{pmatrix}
\begin{pmatrix}
C_{TT}^G \\
C_{TT}^S \\
C_{TT} + L_{TT}^S \\
C_{cc} + A_{cc} + L_{cc}
\end{pmatrix}
= \begin{pmatrix}
f_p \\
f_T^G \\
f_T^S \\
f_c
\end{pmatrix}
\tag{19}
\]

was discretised in time with a generalised time stepping scheme using the implicitness parameter \(\theta\)

\[
y = \frac{y^{m+1} - y^m}{\Delta t} \quad \text{and} \quad \dot{y} = \theta y^{m+1} + (1 - \theta) y^m \tag{20}
\]

Detailed formulations of the element matrices and right hand side vectors used above as well as fundamental model verification can be found in [9]. Picard iterations were used to resolve further nonlinearities and the strongly coupled system of equations was solved monolithically for reasons of robustness.
2.4 Performed simulations

A one dimensional reactor model was set up. Dimensions and boundary conditions for the discharge process are listed in table 1 and material parameters can be found in table 2. Initial conditions were: $p\left(G, x, t=0\right) = 2 \text{ bar}$, $T\left(G, x, t=0\right) = T\left(S, x, t=0\right) = 300 \degree C$, $x_{mV}(x, t=0) = 0$. According to [11, 1] the effective volumetric heat transfer coefficient was determined based on a surface heat transfer coefficient $h$, the porosity $\phi_G$, the particle diameter $d_p$ of the solid grains and the solid thermal conductivity $\lambda_S$:

$$h_{SG} = \frac{6(1 - \phi_G)}{d_p} \left( \frac{1}{h} + \frac{d_p}{10\lambda_S} \right)^{-1} \quad (21)$$

The following variations were run to illustrate the effect of porosity, permeability and particle diameter:

1. $\phi_G = 0.8$, $k_S = 5.0 \cdot 10^{-12} \text{ m}^2$, $d_p = 50 \mu\text{m}$
2. $\phi_G = 0.9$, $k_S = 8.0 \cdot 10^{-12} \text{ m}^2$, $d_p = 50 \mu\text{m}$
3. $\phi_G = 0.7$, $k_S = 1.0 \cdot 10^{-12} \text{ m}^2$, $d_p = 50 \mu\text{m}$
4. $\phi_G = 0.8$, $k_S = 8.0 \cdot 10^{-12} \text{ m}^2$, $d_p = 500 \mu\text{m}$

A more detailed analysis of influential parameters and specifics of the reaction kinetics will be presented elsewhere. The numerically determined energy density values were compared to theoretical values calculated with the relation

$$e_{\text{theor}} = (1 - \phi_G) \frac{\Delta H}{M_{H_2O}} \left( \rho_{Ca(OH)2} R - \rho_{CaO} R \right) \quad (22)$$

where $\Delta H = 112 \text{ kJ/mol}$ is the heat of reaction and $M_{H_2O}$ the molar mass of water.

### Table 1: Geometry and boundary conditions for the discharge of the DLR reactor.

<table>
<thead>
<tr>
<th>Geometry</th>
<th>Boundary</th>
<th>Eq. (13)</th>
<th>Eq. (15)</th>
<th>Eq. (16)</th>
<th>Eq. (17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l = 0.08 \text{ m}$</td>
<td>$x = 0$ : $\rho_G w_n = 0.212 \text{ g/s}$</td>
<td>$T_G = 300 \degree C$</td>
<td>$q_n = 0$</td>
<td>$x_{mV} = 0.36$</td>
<td></td>
</tr>
<tr>
<td>$d = 0.055 \text{ m}$</td>
<td>$x = l$ : $p = 2 \text{ bar}$</td>
<td>$q_n = 0$</td>
<td>$q_n = 0$</td>
<td>$d_n = 0$</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Material parameters of the reference reactor.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{Ca(OH)2} R$</td>
<td>2200 kg m$^{-3}$</td>
</tr>
<tr>
<td>$\rho_{CaO} R$</td>
<td>1656 kg m$^{-3}$</td>
</tr>
<tr>
<td>$c_p S$</td>
<td>1200 J kg$^{-1}$ K$^{-1}$</td>
</tr>
<tr>
<td>$\lambda_S$</td>
<td>0.4 W m$^{-1}$K$^{-1}$</td>
</tr>
<tr>
<td>$h$</td>
<td>80 W m$^{-2}$K$^{-1}$</td>
</tr>
<tr>
<td>$D$</td>
<td>9.65·10$^{-5}$ m$^2$s$^{-1}$</td>
</tr>
</tbody>
</table>
3 RESULTS

3.1 Reference case

At the onset of the process a fast reaction wave travels through the reactor (Fig. 1a), partially converts the solid material (Fig. 1b), extracts most of the steam from the gas (Fig. 1d) and heats up the reactor to equilibrium temperature (Fig. 1c). The subsequent reaction wave is slower due to heat transport limitations and converts the remainder of calcium oxide into calcium hydroxide (Fig. 1). The developing reaction kinetics lead to a somewhat dispersed reaction front with a width of about 25% of the reactor length.

Figure 1: Results for the reference reactor (simulation 1) plotted over the length of the reactor at various time points during the discharge.
3.2 Influence of porosity and permeability

A higher porosity translates into a lower amount of reactive solid material and thus a lower energy density (expressed here based on the reactor volume). A variation of the porosity leads to numerically predicted energy density values that correspond well to the theoretically determined ones (Fig. 2b). When plotting the gas temperature at the outlet over time it can be seen that the reactor with a porosity of 90% is fully converted much earlier than the reactors with porosities of 80% or 70% (Fig. 2a). The pressure profiles in Fig. 2a exhibit the influence of the reaction on the inlet gas pressure over time and the pressure magnitudes clearly reflect the permeability variations performed in simulations 1–3. The lower permeability in the case of a porosity of 70% causes higher pressures and thus also higher reaction equilibrium temperatures (Fig. 2a).

![Graphs showing inlet pressure and outlet temperature, and energy density for different porosities.](image)

Figure 2: Results for the porosity and permeability variations (simulations 1–3) plotted over the length of the reactor at various time points during the discharge.

3.3 Influence of particle diameter

An increase of the particle diameter from 50 µm to 500 µm causes a more pronounced local thermal non-equilibrium around the reaction front as evidenced by significantly higher gas-solid temperature differences (Fig. 3). This is due to the lower interphase heat transfer coefficient for larger particle sizes. It remains to be noted that at the inlet large temperature deviations occur also for smaller particle sizes as long as a chemical reaction is ongoing in that part of the reactor.

4 DISCUSSION

A thermodynamically consistent theoretical model based on the Theory of Porous Media for reactive heat and mass transport in directly permeated thermochemical heat storage systems
was developed. The numerical implementation into the finite element software OpenGeoSys was presented and applied to study an experimental reactor set-up. Parameter variations revealed the impact of the porosity on energy density and conversion time. It was shown that permeability variations do not only cause differences in local pressures but affect the temperature profiles as well. The possibility of local thermal non-equilibrium and a directly calculated local reaction rate in a manner that is coupled to other process parameters allow the model to be used for an assessment of the parameter range within which more simpler models based on assumptions like local thermal equilibrium or a sharp reaction front are valid.

REFERENCES


THermo-CHemical MoDeling oF FIBer-PolyMeR CoMoSiteS IN FIlE For FLuID/STRucTure InterAcTion

V. Biasi*, G. Leplat* AND F. Feyel†

*ONERA - The French Aerospace Lab
F-31055 Toulouse, France
e-mail: valentin.biasi@onera.fr, web page: http://www.onera.fr/

†ONERA - The French Aerospace Lab
F-92322 Châtillon, FRANCE

Key words: Composite materials, Multiphysics Problems, Fire, Pyrolysis, Homogenization

Abstract. Fiber-polymer composite laminates decompose by pyrolysis or oxidation when submitted to high thermal fluxes. The phenomena of heat and mass transfer occurring in these materials are investigated for 2D geometries. A multi-components approach is used to determine the effective properties of those decomposing materials, depending on component fractions and temperature. The gas flow resulting from degradation reactions is driven by Darcy’s law. The degradation of a glass fiber - phenolic resin composite disc exposed to a steady gaussian laser beam is evaluated.

1 INTRODUCTION

Composite materials are being used at an increasing rate in aeronautical structures. Their high tenacity, lightweight and corrosion resistance make those materials appropriate in such applications. However their low durability to fire remains one of the major issues restricting their use. Understanding the decomposition of composite materials submitted to fire, with the support of relevant numerical models, will improve the design of composite structures with respect to safety, environmental and cost reduction constraints.

The degradation of such materials under fire is the result of several combined phenomena. Heat transfers can be considered as the result of conductive effects for limited fluxes (less than $20\,\text{kW/m}^2$) [1]. When thermal fluxes are higher, pyrolysis or oxidation transformations appear and modify the material composition. The effective properties of the composite evolve during the decomposition [2], depending on the temperature and the degradation growth. These transformations produce a solid residue and a gas mixture.
The latter is carried and ejected outside of the material where this gas phase interacts with the surrounding atmosphere (with possibility of ignition), as shown on figure 1.

Several models have been developed for years to represent the decomposition of composite materials under important heat fluxes, mostly based on Henderson et al. [3, 4]. In these models, a composite in degradation is considered as a mix of virgin and charred material. This approach has been modified and enhanced most notably by Florio et al. [5], Sullivan and Salomon [6, 7], Dimitrienko [8] and Galgano et al. [9]. These studies
bring essential informations about effects of non-thermal equilibrium, thermal expansion or effective stress modifications. Although these models can accurately describe temperature fields within the material, they are limited to 1D cases assuming constant and known heat fluxes.

These assumptions are restrictive for the study of composites in fire because decomposition gases are ejected out of the material and interact with flames. Inflammation of those gases produces an important heat emission [10] which has an impact on thermal flux received by the material, and so on decomposition. Multi-dimensional effects of degradation of composites are rarely investigated [11, 12] although these are essential to take into account fire interactions including accurate heat flux distribution on the impinged surface.

The objective of this work is to propose a model of degradation of composite materials for flame/structure interaction. A major step to achieve this goal is to develop a relevant model which describes thermo-chemical degradation of composite materials under steady heat flux. A bi-dimensional approach is used to model heat and mass transfer in orthotropic materials. The decomposing composite is considered in this model as a multi-components material forming a porous medium. Pyrolysis and oxidation make solid components react to form gaseous mixtures that are transported through the material up to surfaces. Unsteady interface conditions (heat and mass fluxes, gas mixture composition, temperature and pressure) are the input data that will be used to couple the solid decomposition solver with a Navier-Stokes combustion solver.

2 THERMO-CHEMICAL MODELING

The decomposing composite material is represented in this model as a mix of different solid and gas components. Solid components are usually fibers or matrix both at the virgin or charred state, or in a simpler representation the whole solid phase and the charred phase. The gas phase is considered as a mixture of several ideal gases. Throughout this work, a composite material is modeled by a set of $I$ components including $J$ gaseous components. The subscript $i$ is used to refer to any component and the subscript $j$ is used to refer only to a gaseous component.

$\varphi_i$ and $Y_i$ denote respectively the volume fraction and the mass fraction of each $i$ component. $\rho_i$ denotes the absolute density (mass of $i$ divided by volume of $i$) and not the bulk density (mass of $i$ divided by total volume, which could be expressed as $\rho Y_i$). Absolute density of each solid component is assumed constant ($\rho_i = \rho_{i0}$ if $i \in \{s\}$). The weighted density is given by the relation:

$$\rho = \sum_{i=1}^{I} \rho_i \varphi_i \tag{1}$$
The porosity \( \varphi_g \) is calculated as the volume fraction of the whole gas phase:

\[
\varphi_g = \sum_{j=1}^{J} \varphi_j \quad (2)
\]

### 2.1 Properties

The average molecular mass \( M \) of the gaseous phase is weighted by gaseous phase volume fractions:

\[
M = \frac{1}{\varphi_g} \sum_{j=1}^{J} \varphi_j M_j \quad (3)
\]

The surface emissivity \( \varepsilon \) and absorptivity \( \alpha \) are weighted by solid phase volume fractions, which correspond to surface area fractions of each \( i \) solid component:

\[
\varepsilon = \frac{1}{\varphi_s} \sum_{i=1}^{I} \varphi_i \varepsilon_i \quad , \quad \alpha = \frac{1}{\varphi_s} \sum_{i=1}^{I} \varphi_i \alpha_i \quad \text{with} \quad i \in \{s\} \quad (4)
\]

where \( \varphi_s \) is the solid volume fraction. The heat capacity \( C_P \) of the material depends on the \( C_{Pi} \) of each \( i \) component weighted by mass fractions:

\[
C_P = \sum_{i=1}^{I} Y_i C_{Pi}(T) \quad (5)
\]

and each \( C_{Pi} \) depends on temperature using polynomial laws. Sensible enthalpy is calculated, for each component or for the whole material, by integration of heat capacity between the initial temperature \( T_0 \) and the current temperature \( T \):

\[
h_i = \int_{T_0}^{T} C_{Pi}(\tau) d\tau \quad \text{and} \quad h = \int_{T_0}^{T} C_P(\tau) d\tau \quad (6)
\]

The averaged effective thermal conductivity is defined as a second order tensor. Any \( j \) gas component is considered as isotropic \( (k_j = k_j I) \). For every gas or solid component, each coefficient of the associated tensor depends on temperature using a polynomial law. The effective thermal conductivity depends on the fiber diameters and orientations, pore dimensions and distributions, and those particularities should be taken into account. Due to the lack of experimental characterizations, a mixture law (using volume fractions) is used to evaluate \( \overline{k} \):

\[
\overline{k} = \sum_{i=1}^{I} \varphi_i \overline{k_i} \quad (7)
\]

where \( \overline{k_i} \) is the thermal conductivity tensor of the \( i \) component. An ideal gas law is used to calculate the internal pressure as \( P = \rho g \frac{R}{M} T \). Since Reynolds number in these media in
degradation is relatively low [13], the momentum conservation’s law can be homogenized using Darcy’s law. The latter links the averaged gas velocity vector $v_g$ to the pressure gradient as:

\[ v_g = -\frac{K}{\mu_g} \nabla P \]  

(8)

where $\mu_g$ is the gas dynamic viscosity and $K$ the permeability. This value follows a geometric average between the initial value $K_0$ and the final value $K_f$:

\[ K = K_0^{1-\varphi_g} K_f^{\varphi_g} \]  

(9)

2.2 Chemical reactions

A set of $M$ chemical reactions is considered to model pyrolysis or oxidation transformations. It is assumed that only solid components can react and that formed gases are inert. The general form of the $m$ reaction is expressed as:

\[ \nu_{Rm} R_m + \nu_{O_{2m}} (O_2 + \lambda N_2) \xrightarrow{+Q_m} \nu_{Pm} P_m + \sum_{l=1}^{L} (\nu_{ml} G_{ml}) + \nu_{O_{2m}} \lambda N_2 \]  

(10)

where $R_m$ is a solid reactant, $O_2 + \lambda N_2$ is the air (if present and if oxidation reactions occur), $P_m$ is a solid product and $G_{ml}$ is a gas product (in the set of $L$ gas products emitted in the $m$ reaction). $\nu$ is used to denote stoichiometric mass coefficients and $\lambda$ is the mass ratio of $N_2$ on $O_2$ in the air, which equals to 3.26.

The heat of reaction $Q_m$ is introduced to express the generated or consumed heat in the $m$ reaction for each consumed quantity of $R_m$. The reaction rate of $R_m$ is driven by an Arrhenius law:

\[ \dot{\omega}_{Rm} = -f(O_2) (\rho R \varphi_R)_0 \left( \frac{\rho R \varphi_R}{(\rho R \varphi_R)_0} \right)^{n_m} A_m \cdot exp \left( \frac{-E_{A_m}}{RT} \right) \]  

(11)

where $A_m$, $E_{A_m}$ and $n_m$ are the Arrhenius parameters for the $m$ reaction. Terms with the subscript ”0” correspond to quantities at the initial state (before that any degradation reaction occurs). The factor $f(O_2)$ in equation 11 allows to take into account the oxygen concentration on the reaction rate $\dot{\omega}_{Rm}$, as:

\[ f(O_2) = 1 \text{ if } \nu_{O_{2m}} = 0, \quad f(O_2) = \left( \frac{\rho O_2 \varphi_{O_2}}{(\rho O_2 \varphi_{O_2})_0} \right)^{n_{O_{2m}}} \text{ otherwise} \]  

(12)

The reaction rate $\dot{\omega}_{Rm}$ is linked to other $\dot{\omega}_{im}$ reaction rates for each $i$ component which takes part in the $m$ reaction:

\[ \dot{\omega}_{im} = \delta_{im} \frac{\nu_{im}}{\nu_{Rm}} \dot{\omega}_{Rm} \]  

(13)
where $\delta_{im}$ sets the sign of the reaction rate in equation 13: $\delta_{im} = +1$ if $i$ is a reactant and $\delta_{im} = -1$ if $i$ is a product in the $m$ reaction. The total source term of the $i$ component (formation and destruction) is the sum of all reaction rates $\dot{\omega}_{im}$ where $i$ takes part in the set of $M$ reactions:

$$\dot{\omega}_i = \sum_{m=1}^{M} \dot{\omega}_{im}$$ (14)

Another remarkable value is $(\dot{\omega}_R Q)$, the total heat source produced or consumed by the whole $M$ reactions:

$$(\dot{\omega}_R Q) = \sum_{m=1}^{M} \dot{\omega}_{Rm} Q_m$$ (15)

where $Q_m$ is the heat of reaction in $m$.

### 2.3 Conservation laws

All conservation laws in this section correspond to averaged local equations. A mass conservation law is written for each $i$ component (solid or gas). The mass variation is caused only by degradation reactions for solid components, considered as a source term.

$$\frac{\partial}{\partial t} (\rho_i \varphi_i) = \dot{\omega}_i$$ (16)

Concerning gas components, mass variations result from gas production or consumption as well as bulk gas transport. This transport term is driven by the pressure gradient as expressed in equation 8. The gas component mass conservation is:

$$\frac{\partial}{\partial t} (\rho_j \varphi_j) = -\nabla \cdot \left( \frac{Y_j}{Y_g} \rho_g v_g \right) + \dot{\omega}_j$$ (17)

This equation is also valid for $O_2$ if needed. The local thermal equilibrium is assumed in the material, meaning that there are no local temperature difference between all components. Kinetic energy and pressure work are ignored in the energy conservation.

$$\frac{\partial}{\partial t} (\rho h) = -\nabla \cdot (-k \nabla T + h_g \rho_g v_g) + (\dot{\omega}_R Q)$$ (18)

In this conservation law, the left hand side represents the internal energy variation for all components. The first right hand side term is the contribution of fluxes. The latter is composed of the conduction term and of the energy transport term of the whole gas phase. The last term is the energy source term, contribution of all heat released for the $M$ reactions. These governing equations are forming a set of $I + 1$ partial differential equations that has to be solved.
3 NUMERICAL METHODS

A numerical solver has been developed for 2D (planar or axisymmetric) geometries, using a finite volume numerical method to discretize the equations developed in 16, 17 and 18. Those equations are integrated over small volumes in unstructured meshes, then divergence terms are converted into surface integrals using the divergence theorem. Natural variables \((T, P, \rho, \varphi_i, Y_i)\) are interpolated at surface centers and all properties developed in section 2.1 are calculated from those natural variables. Gradients are evaluated at cell centers using the gradient theorem from known values at adjacent cells. The integration in time is performed using an implicit theta-scheme. Since Arrhenius based source terms can cause stability problems due to non linearity and slow characteristic times, source terms are integrated explicitly using sub-time steps.

4 RESULTS

4.1 Study case

An evaluation of this bidimensional solver is performed using a glass fiber - phenolic resin composite material, designated as H41N. Material properties were characterized by Henderson et al. [4] and are summarized in table 1. These properties have been established for the virgin and the charred state, and as a consequence three components are used to describe the decomposing composite: "\(v\)" the virgin component, "\(c\)" the char component and "\(g\)" the gaseous phase. Thermal conductivities are arbitrarily modified to illustrate orthotropic heat transfers.

| \(\rho_v\) [kg/m\(^3\)] | 2040.6 | \(K_0\) [m\(^2\)] | 2.60 \times 10^{-18} |
| \(\rho_c\) [kg/m\(^3\)] | 1980.7 | \(K_I\) [m\(^2\)] | 1.14 \times 10^{-16} |
| \(M_g\) [kg/mol] | 18.35 \times 10^{-3} | \(\varepsilon_v - \alpha_v\) [-] | 0.6 |
| \(k_v\) [W/m/K] | 1.12 + 3 \times 10^{-3}T | \(\varepsilon_c - \alpha_c\) [-] | 0.9 |
| \(k_v\) [W/m/K] | 0.35 + 5 \times 10^{-4}T | \(\mu_g\) [kg/m/s] | 8.0 \times 10^{-6} + 2.5 \times 10^{-8}T |
| \(k_c\) [W/m/K] | 1.12 + 3 \times 10^{-3}T | \(\varphi_{g0}\) [-] | 0.113 |
| \(k_c\) [W/m/K] | 0.35 + 5 \times 10^{-4}T | \(\varphi_{gf}\) [-] | 0.274 |
| \(k_g\) [W/m/K] | -8.4 \times 10^{-3} + 1.4 \times 10^{-4}T | \(Q\) [J/kg] | -234 \times 10^{13} |
| \(C_p_v\) [J/kg/K] | 791.3 + 1.09T | \(A\) [-] | 1.98 \times 10^{29} |
| \(C_p_c\) [J/kg/K] | 600.4 + 1.02T | \(E_A\) [J/mol] | 2.6 \times 10^{5} |
| \(C_p_g\) [J/kg/K] | 2096 + 1.05T | \(n\) [-] | 17.33 |

The chosen domain is a 80mm diameter and 4.16mm thick disc. A gaussian laser beam impacts the upper face, centered on the symmetry axis. The computational domain is reduced to an axisymmetric rectangle of 40mm along the X axis (representing the disc radius) and 4.16mm along the Y axis. This domain is splitted into uniform rectangular cells (1mm along X and 0.208mm along Y) and the time step is fixed at 0.1s.

All domain boundaries are considered at atmospheric pressure \((P_0 = 101 325 Pa)\). On surfaces, convective heat transfers are neglected but radiative heat transfers are considered...
between surfaces and the surrounding atmosphere temperature (fixed at 293K). At the initial state, the whole domain is at \( T_0 = 293K \) and at atmospheric pressure \( P_0 \). The surrounding atmosphere is considered as inert (non-oxidative) and only one reaction drives the degradation (where Arrhenius coefficients \( E_A, A \) and \( n \) are detailed in table 1). The virgin material ”\( v \)” pyrolyses to form the char material ”\( c \)” and gaseous phase ”\( g \)”, with stoichiometric mass coefficients:

\[
v \xrightarrow{Q} 0.795 \ c + 0.205 \ g \quad (19)
\]

Between \( t = 0s \) and \( t = 400s \), a gaussian heat flux is applied on the upper surface, with a maximal intensity of 318.3\( kW/m^2 \) along the symmetry axis and a half-width of 10\( mm \) at \( 1/e^2 \).

4.2 Discussions

Figure 2 shows a mirror view of the domain at four different times after the degradation begins. Temperature fields are presented on the right hand side and pressure fields on the left hand side of the axis. Averaged gas velocities are represented by vectors only on the domain boundaries. Figure 3 shows temperatures and virgin mass fractions, and figure 4 shows internal pressures extracted along the symmetry axis at the same different times detailed in figure 2.

Temperature is quickly high enough on the front face (\( T = 940K \) at \( t = 50s \)) to cause pyrolysis. As a consequence, the virgin mass fraction on the front face has considerably decreased after \( t = 50 \) seconds (\( Y_v = 0.45 \) at \( Y = 4.16mm \)). The virgin component consumption produces a gaseous phase, and so the internal pressure increases significantly (\( P_{max} = 10.4 \times 10^5Pa \) at \( t = 50s \)). The formed gases are concentrated around the symmetry axis (see figure 2). The atmospheric pressure imposed at all boundary surfaces produces pressure gradient and as a consequence the gaseous phase flows through the material up to the front face where averaged gas velocity exceeds 3.2\( mm/s \) at \( t = 50s \) along the symmetry axis (see figure 5).

The pressure peak continues to increase up to \( P_{max} = 11.8 \times 10^5Pa \) at \( t = 100s \). It induces important pressure gradients in both directions along the symmetry axis (see figure 4) while decomposition gases ejection remains mainly located at the front face and negligibly at the rear face, as depicted in figures 2 and 5. This difference is explained by higher permeability values in decomposed regions which result in a higher gas velocity even if pressure gradients have the same order of magnitude on both sides.

Then the pressure peak begins to decrease after 100s and reaches 9.6 \( \times 10^5Pa \) at \( t = 200s \). At this time, the degradation reached the rear face as shown by the virgin mass fraction in figure 3 (\( Y_v = 0.98 \) at \( t = 200s \) and \( Y = 0mm \)). As a consequence, the burnthrough phenomenon causes a release of the gaseous phase with an increasing gas velocity at the rear face until \( t = 400s \) and an important decrease of the internal pressure (\( P_{max} = 3.6 \times 10^5Pa \) at \( t = 400s \)) within the material.
Figure 2: Temperature (right hand side) and pressure (left hand side) distributions associated with averaged gas velocities on surfaces extracted at different times.
Figure 3: Temperature and virgin mass fraction along the symmetry axis

Figure 4: Internal pressure along the symmetry axis

Figure 5: Norm of averaged gas velocity at the front and rear face at $X = 0\, mm$

Figure 6: Evolution of temperature on the front face

Afterwards, either on the front or rear faces, the gas outlet profile tends to spread out of the laser beam region (beyond $20\, mm$ width). The conductive heat transfer, higher along $X$ due to orthotropic conductivities, generates high enough temperature increase to cause pyrolysis in the peripheral region too (see figure 6). Finally, the gaseous phase production does not occur any longer in the symmetry axis region but moves away along the $X$ axis as shown by the pressure peak location over the domain at $t = 400\, s$ and the associated velocity vectors at the boundaries in figure 2.
5 CONCLUSIONS

A model has been developed with the aim to predict thermo-chemical degradation phenomena on most fiber-polymer composites submitted to non-uniform fluxes. The multi-components formulation, associated with a specific description of each component, allows describing detailed transformations within the material.

An experimental study will be performed to validate the developed model using a configuration similar to the study case developed in section 4 and from the assessment of all material properties. However, the more important the number of components is, the more significant the characterization process will be to determine all needed parameters. It is all the more important since the accuracy of the simulations are subject to experimental uncertainties that must be taken into account in further analyses and comparisons of results.

Nevertheless, the present approach integrated into a bidimensional solver turns out to be suitable to represent internal thermo-chemical behaviors and expected boundary conditions of a decomposing anisotropic composite material. The unsteady gas ejection profiles have been identified to provide essential information for the investigation of the interaction between the fire and composite surface including the effects of the heat generated by the ignition of those combustible gases.

REFERENCES


IMPLEMENTATION OF A WEAK COUPLING APPROACH BETWEEN A CFD AND AN FE SOFTWARE FOR FIRES IN COMPARTMENT

NICOLA TONDINI* AND JEAN-MARC FRANSSEN†

* Department of Civil, Environmental and Mechanical Engineering (DICAM)
  University of Trento
  Via Mesiano 77, 38123 Trento, Italy
  e-mail: nicola.tondini@unitn.it

† Structural Engineering Sector (ArGEnCo)
  Chemin des Chevreuils 1, 4000 Liège, Belgium
  e-mail: jm.franssen@ulg.ac.be

Key words: CFD-FE interface, Weak coupling, Compartment fires.

Summary. The paper presents the assumptions and the issues that arise when developing an integrated modelling methodology between a Computational Fluid Dynamics (CFD) software applied to compartment fires and a Finite Element (FE) software applied to structural systems subjected to fire. In particular, a weak coupling approach used to simulate a fire exposed structure by modelling the fire development in the compartment, the heat penetration in the structure and the mechanical response is described. The advantages and the disadvantages of such a technique are highlighted compared to a full coupling that conversely takes into account all mutual interactions. The favourable aspect of computing the thermal response of the structure in the FE model in order to avoid modelling the structure in the CFD model is underlined, namely a sensitive reduction of computational demand. Finally, the study is enriched by an application of this methodology that concerns the simulation of a pool fire in an open compartment the results of which are compared with those obtained by employing the simplified Hasemi model included in the Eurocodes.

1 INTRODUCTION

Integrated modelling methodologies applied to compartment fires in order to obtain the thermo-mechanical response of structures exposed to fire would represent a powerful tool to widen the application field of structural fire safety engineering by overcoming limitations associated to simplified procedures. The exploitation of such strategies based on the coupling between CFD and FE programs are already used in medicine, e.g. for modelling the blood flow in arteries [1]. However, in the fire engineering field, very few applications are available and the ones that have been developed are often limited to specific software pairs [2]. The Research Fund of Coal and Steel (RFCS) project called FIRESTRUCT [3] dealt with this issue by studying different coupling approaches and employing different software. In this paper is described the weak coupling approach developed here between the CFD software...
2 OVERVIEW ON INTEGRATED CFD-FE METHODOLOGY

2.1 Compartment fires: problem definition

Three problems have to be solved when modelling the behaviour of a structure subjected to a compartment fire, each of them being governed by different physical phenomenon and, hence, by different equations: 1) temperature development in the compartment; 2) thermal response of the structure and 3) mechanical response of the structure. The fire development analysis yields the temperatures of gasses and the radiative and mass flows in the compartment. By means of the thermal analysis, the temperatures in the structural elements are obtained. The mechanical response provides the behaviour of the structural system, i.e. stresses, deflections etc. Several differences distinguish these three processes. First, the spatial scale of the thermal analysis in the structure is an order of magnitude smaller than the spatial scale used for the compartment temperature development and the mechanical response. Second, the time scale may be different to solve the problem within CFD and FE method. Third, for the temperature development in the compartment and the mechanical response a 3D analysis is generally required, whereas for the thermal analysis a 2D analysis is usually sufficient. Thus, some issues arise when an integrated methodology CFD-FE is to be used to tackle the whole problem. It is natural to assign the task of performing the compartment temperature development analysis to the CFD model and the mechanical response to the FE model, but it is not so straightforward to decide where to carry out the thermal analysis. Both software may be exploited to fulfil the task. The advantages to perform the thermal analysis in the CFD model are first to get direct information from the compartment temperature development analysis and, second, to allow consideration of the energy absorbed by the structure to be considered in the analysis of the compartment. On the other hand, if the thermal analysis in the structure is carried out in the FE model, all necessary data are directly available by the FE code to determine the mechanical response. Whatever the choice, the difficulties arise when data have to be exchanged between the two software, because of different scales in space and time. Moreover, if the thermal analysis is performed in CFD software, the compartment model must include the structure as well. The latter aspect is not desirable as described later on.

2.2 Levels of coupling

From the description of the problem it is clear that coupling CFD to FE model is far from straightforward and that the selected level of coupling influences the complexity of the model. In reality all three problems are mutually coupled (full coupling or two-way coupling) as shown in Figure 1 and reported in Table 1, where the main phenomena involved in a compartment fire are listed along with their mutual interaction.
Figure 1: Full coupling strategy between the main phenomena involved in a compartment fire

Table 1: Mutual interactions of main phenomena involved in a compartment fire

<table>
<thead>
<tr>
<th>COUPLING</th>
<th>1 to 2</th>
<th>2 to 3</th>
<th>3 to 1</th>
<th>1 to 3</th>
<th>3 to 2</th>
<th>2 to 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convection and radiation to structural elements</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VELOCITY OF GASES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convection factors</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic pressure on walls, windows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>PRESSURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static pressure on walls, windows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE IN MATERIALS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermal elongation of elements</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation of mechanical properties</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absorption of energy from the compartment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>PLASTICITY AND CRACKING IN ELEMENTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generation of heat or heat leakage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Modification of material thermal properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>DISPLACEMENTS IN ELEMENTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modification of the gas flow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Modification of the element thermal exposure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

The implementation of a full coupling allows taking into account all phenomena and it guarantees a general field of application as well as a solution that tends to be exact. An example of full coupling is the interface developed between VESTA, a CFD software, and DIANA, an FE software developed by TNO in the Netherlands [3]. However, an integrated methodology that relies on full coupling is very complex to achieve. The first reason lies in various uncertainties that question the so-called exactness of the method. For instance, heat leakage through cracks in concrete or gypsum plaster boards enclosures are still very difficult to quantify because they do not follow deterministic rules. Moreover, from a programming point of view, the code of the selected CFD software and the code of the FE program have to be modified so that they can communicate for the exchange of data, but it means that in most cases the integrated methodology will not work if another CFD or another FE software is used. This is a clear drawback in terms of versatility and flexibility. Furthermore, for each simulation, a CFD specialist as well as an FE specialist are required since the two models cannot be run independently. Other typical issues that may occur in the design practice are related to possible modifications that the structure undergoes during the construction project as well as modifications of the structure that have to be applied because of an unsatisfactory
behaviour in terms of fire safety requirements. Since the structural elements must be included in the CFD model, any changes in the structural system imply that the whole analysis must be re-run, entailing large time consuming analyses.

From these considerations, a simplified approach, the so-called weak coupling, is proposed to overcome the major aforementioned issues with the aim to be applicable to a wide number of likely-to-occur scenarios in compartment fires.

3 PROPOSED WEAK COUPLING METHODOLOGY

3.1 Assumptions and general remarks

In the proposed weak coupling (or one-way) approach the mutual interactions are discarded, as illustrated in Figure 2. The CFD software models the fire development, while the FE program performs the thermal and the mechanical analyses. The fire development is calculated independently of the thermal response in the linear elements of the structure such as, for example, steel columns, beams or truss girder. If part of the structure is made of planar elements that also constitute boundaries of the compartment such as, for example, concrete walls or slabs, they must be modelled, perhaps with some degrees of approximation [4], in the fire development analysis. The detailed temperature field in these structural elements will nevertheless be computed subsequently by the FE software. As a consequence, if \( p \) variations of the structure must be evaluated under \( q \) fire scenarios, only \( q \) CFD analyses have to be performed, compared to \( p \times q \) coupled analyses in a full coupling approach.

In this strategy, the thermal response of the structure represents the input of the mechanical analysis. Hence, it can be performed first, over the whole time domain, and then the resulting data are transferred at the beginning of the mechanical analysis which is performed subsequently.

\[ \text{Figure 2: Weak coupling strategy between the main phenomena involved in a compartment fire} \]

Nonetheless, these simplifications imply some limitations:

i) the dimensions of the structural elements and their displacements perpendicular to their longitudinal axis must be small compared to the dimensions of the compartment in order not to significantly influence the temperatures and the air flow around the elements. According to this assumption a series of 2D thermal analyses rather than a unique 3D thermal analysis is made on the structural elements because the transverse dimension of the section of the elements is small with respect the longitudinal dimensions. For instance, a 1x1 m² concrete
columns in a 100 m$^2$ compartment must clearly be considered in the CFD model. This would also be the case for 1-meter deep concrete beams in a car park with a distance from the floor to the beams that is on the order of magnitude of 2 meters. Very flexible structures that are sensitive to air pressure variation are also not suitable for such integrated methodology because the effects of air pressure variation on the displacements of the structure cannot be neglected. Floor systems designed according to the tensile membrane action also exhibit very large displacements during the fire and may also not comply with this requirement if the floor to ceiling distance is small compared to the displacement.

ii) It is possible, for each 2D thermal analysis, to consider the boundary conditions at the surface of the section at the same point, namely the point of the section located on the node line of the beam element, for example at the centre of gravity of the section. The influence of the distance from the node line of the section to the border of the section is neglected. This is consistent with the fact that the structure is not present in the CFD analysis because the size of the section perpendicular to the longitudinal axis is negligible with respect to the size of the compartment; the distance from the centre of the section to the border (approximately ½ of the size of the section) is then also negligible.

iii) Generally, in the CFD model the dimensions of a parallelepipedic compartment correspond to the clear distances between opposite walls. However, in the FE model a slab is generally modelled in correspondence to its centreline as illustrated in Figure 3. Thus, the slab would fall outside the CFD domain and assumptions have to be made in order to determine thermal information at the slab centreline.

![Figure 3: Different compartment dimensions for the CFD and the FE model.](image_url)

iv) Since the structure is not included in the CFD model, the effect of shielding from any structural elements on others cannot be detected. For example, if a series of closely-spaced columns one behind the other is impinged by a radiant flux with direction parallel to the column series, the magnitude of the flux received by each column will only depend on the distance of each single column from the fire source and no effects of shielding will be taken into account on the columns behind the first one.

v) Irrigated structures in which water is circulating in order to keep the temperature of the structure within acceptable limits cannot be neglected because they may contribute in
evacuating important amount of energy from the compartment.

This procedure is thus particularly well adapted for metallic structures made of relatively thin members (frame, truss girders) and located in very large compartments (railway or airport entrance halls, exhibition halls) where a localised fire is developing and simplified thermal models, such as those proposed in EN1991-1-2 [5], cannot be employed because the geometry of the compartment is too complex or the position of the structure in the compartment or with respect to the position of the fire is not within the field of application of simplified model.

3.2 How it works

In this section, the practical issues that have to be solved when implementing such an approach are presented. The programs used in this paper to illustrate the proposed integrated methodology are FDS [6] and SAFIR [7]. The main steps needed to couple CFD and FE are:

1) at the end of the CFD analysis, a transfer file containing all information regarding the state in the compartment, i.e. temperature of gas, convection factors and radiant intensities from various directions, is produced. These quantities can be provided at each grid point of the CFD model (the grid that was required to allow a precise determination of the solution) or instead at grid points of a coarser mesh reckoned by the CFD user as sufficient to get a sufficiently accurate representation of the solution that has been obtained. The mesh of considered points can also be refined around the structure that will be analysed and a coarser away from the structure.

The format of the transfer file should be as standardised as possible so that in a future perspective it could be used for any choice of CFD and FE software. Hence, type of file (e.g. ASCII), syntax, type of reference system, type and format of numbers, presence of blank lines etc. have to be clearly stated. Such a format has been proposed within the FIRESTRUCT project and can be obtained from the authors. Radiant intensities from different directions are preferred to impinging radiant fluxes on predefined surfaces because the structural elements are not included in the CFD model and thus no information is available at that stage about the shape of the cross-sections. The fluxes at the surface of the structural elements will be computed within the FE software by integrating the radiant intensities which allows taking into account possible shadow effects in concave sections.

2) A 3D Cartesian spatial interpolation is needed because the points of the structure where the information is needed (called here “the structural points”) generally do not coincide with the points of the CFD grid where the information is provided. If any structural points fall outside the CFD domain for the reason described above, they are moved to the closest boundary of the CFD domain where the Cartesian interpolation can be made, namely the boundary corresponding to the centre of the outermost cells because FDS provides the information at the centre of cells (see Figure 4). A trilinear interpolation algorithm was successfully implemented in SAFIR to fulfill the 3D Cartesian interpolation.
3) An interpolation in the time domain is also necessary because the time steps of the CFD analysis and the time steps of the thermal analysis may not be the same. In this case a simple linear interpolation may be used.

4) In order to get the impinging fluxes \( q \) on the surface of the structural elements a spherical numerical integration of radiant intensities \( I \) has to be performed. A numerical integration can be performed according to Eq 1.

\[
q = \sum_{i} n I_i \cos \theta_i \omega_i
\]  

where \( n \) is the number of intensities considered for the integration, \( \omega_i \) is the solid angle associated to the direction \( i \) and \( \theta_i \) is the angle between the direction of the radiant intensity \( i \) and the normal to the surface.

The directions of the intensities which are required to perform the spherical integration are generally not the directions in which the intensities are given by the CFD analysis. This is particularly the case if the structural elements are not parallel to the axes of the system of coordinates used in the CFD analysis (e.g. for diagonals in a truss girder). A spherical interpolation is thus performed in order to obtain the radiant intensities in the directions required by the numerical integration. Rotations of local axes are required to find the surface system of coordinates taking into account the direction of the longitudinal axis and the shape of the cross-section.

It is essential that the type of mesh and type of system of coordinates used in the CFD analysis (step 1) be clearly defined and taken into account in steps 2 and 4. The format of the transfer file established within the FIRESTRUCT project is based on the hypotheses of a structured parallelepipedic mesh in a dextrorsum Cartesian system of coordinates. The position of the origin of the system of coordinates and the directions of the \( X \), \( Y \) and \( Z \) axes as well as the direction of gravity must be common in the CFD and in the FE analyses.

4 APPLICATION OF THE METHODOLOGY

In order to verify the methodology and to give insight to its capabilities, a comparison with the simplified Hasemi model included in EN1991-1-2 [5] by simulating a pool fire in an open compartment was carried out.

In detail, an open compartment (created to simulate for instance a small part of an open car park) with dimensions 5 x 3 x 3 m was modelled in FDS without including structural elements and the ceiling being an adiabatic surface. A pool fire of diameter \( D = 1.5 \) m, having a
constant Rate of Heat Release (RHR) \( q = 1727 \text{ kW/m}^2 \) (\( Q = 3051 \text{ kW} \)), was located in the compartment at coordinates \( x = 3.5 \text{ m} \) and \( y = 1.5 \text{ m} \), as illustrated in Figure 5 and Figure 6. The power of the pool fire was chosen so that the flames impacted the ceiling and thus, the Hasemi model could be applied. After the CFD analysis, the results, in terms of temperatures and radiant intensities, were transferred to SAFIR that performed the thermal analysis of an HE400M steel section of a beam located at a distance of 1 m from the plume axis. The same scenario was simulated in SAFIR by exploiting the Hasemi model. Since the flux in the Hasemi model is given at the ceiling level, the thermal analysis performed by the integrated modelling FDS - SAFIR also relied on thermal data obtained at the ceiling level, i.e. \( z = 3 \text{ m} \), so that to provide a more coherent comparison. In Figure 7 it is possible to observe the difference of temperature distribution in the section between the two methods and it is clearly visible the effect of the non-uniform radiation affecting the sides of the section by employing the integrated modelling FDS - SAFIR, whereas in the Hasemi model this is not possible. The comparison of the results is shown in Figure 8 in terms of the temperature at the bottom right corner of the bottom flange of the section and at the bottom right corner of the top flange of the section (see Figure 7). Points where on the basis of their location it could expect results quantitatively comparable with the Hasemi model in the view of a meaningful comparison of modelling were selected. The temperature obtained by employing the integrated modelling FDS - SAFIR is slightly lower and the reason can be traced to lower fluxes predicted by FDS compared to those given by the Hasemi model. With regard to the simulation time step, because of the Large Eddy Simulation (LES) model used in FDS, the fire plume oscillates constantly during the fire and so do the instantaneous values of the radiant intensities. The oscillations in the example are due to the fact that the section is in the intermittent region of the flame.
Figure 5: Compartment geometry.

Figure 6: Modelling of the pool fire in FDS.

Figure 7: Comparison of temperature distribution in the beam at 3600 s between a) the integrated modelling FDS-SAFIR and b) the HASEMI model.
Figure 8: Comparison in terms of temperature between the integrated modelling FDS-SAFIR and the HASEMI model at the bottom right corner of the bottom flange of the section.

5 CONCLUSIONS

This paper describes assumptions and issues of an integrated modelling methodology for the behaviour of a structure located in a fire compartment by implementing a weak coupling approach. It is suitable for localised fires in very large compartments built with relatively thin structures where the transverse dimensions of the structural elements can be neglected. In this respect, the comparison with the Hasemi model demonstrated the reliability of the methodology and the capability to overcome problems that cannot be analysed by simplified models. The choice to perform the fire development analysis independently from the thermal analysis of the structure and from the mechanical response analysis by neglecting the structural elements in the CFD model allows reducing the computational demand and enhances the versatility of the methodology.

6 ACKNOWLEDGEMENTS

The financial support of ArcelorMittal and of the Province of Trento through the Call 3 - post-doc 2009 (Outgoing) - FIRAS project are gratefully acknowledged.

REFERENCES

NON-SYMMETrICAL BE-FE PARTITIONED FORMULATION FOR ACOUSTIC FLUID-STRUCTURE INTERACTION PROBLEMS

LUIS RODRIGUEZ-TEMBLEQUE∗, JOSÉ ÁNGEL GONZÁLEZ, ANTONIO CERRATO AND RAMÓN ABASCAL

Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Camino de los Descubrimientos s/n, Sevilla E-41092, Spain
∗e-mail: luisroteso@us.es

Key words: Domain Decomposition, FETI, nsBETI, Fluid-structure interaction, localized Lagrange multipliers

Abstract. This work presents a non symmetric Finite-Boundary Element Tearing and Interconnecting (FE-BETI) formulation for acoustic FSI problems, where the finite element method is used to model the structure, while the acoustical fluid domain is represented by the boundary element method. The method interconnects fluid and structure domains using the localized Lagrange multipliers, allowing the use of non-matching meshes on the interfaces. Furthermore, the methodology proposes a preconditioned projected bi-conjugate gradient solver, that presents good scalability properties in the solution of large problems.

1 INTRODUCTION

This paper extends the recently proposed nsBETI [1] formulation to acoustics-FSI problems, where the finite element method is used to model the structure, while the acoustics fluid domain is represented by the boundary element method. This non-overlapping domain decomposition technique uses the classical non-symmetrical acoustics boundary element formulation, instead of a symmetric Galerkin boundary element one [2]. The method interconnects fluid and structure domains using the localized Lagrange multipliers [3, 4, 5, 6, 7, 8, 9, 10, 11], which also allow considering non-matching meshes on the interfaces. Furthermore, the methodology uses a preconditioned Bi-Conjugate Gradient
Stabilized (Bi-CGSTAB) algorithm, which presents a very good scalability in the solution of large problems.

2 ACOUSTIC FSI PARTITIONED FORMULATION

A FEM structure and a BEM fluid domain are considered, so the total virtual work of the system $\delta W_T$ can be expressed as the addition of the virtual work done by the FEM structure domain $\delta W^s$, the BEM fluid domain $\delta W^f$ and the interface coupling contribution $\delta W^c$,

$$\delta W_T = \delta W^s + \delta W^f + \delta W^c$$ (1)

2.1 Structure domain

The virtual work of a flexible structure, which is susceptible to the dynamic of the fluid, $\delta W^s$, is described by the principle of virtual work for a continuum body of domain $\Omega_s$ and surface $\Gamma_s$ that, assuming small displacements,

$$\delta W^s = \int_{\Omega_s} \sigma_s : \nabla \delta u_s d\Omega - \int_{\Omega_s} (\omega^2 \rho u_s + b_s) \cdot \delta u_s d\Omega - \int_{\Gamma_s} t_s \cdot \delta u_s d\Gamma$$ (2)

where $u_s$ are the structural displacements, $\sigma_s$ the Cauchy stress tensor, $t_s$ the applied surface tractions and $b_s$ the body forces. Finally, $\omega$ and $\rho$ are the angular frequency of the displacement oscillation and the density of the structures, respectively.

Next, the substructure is discretized using the classical FEM approximation, where the assembly of element contributions by the direct stiffness method leads to the semidiscrete equations of motion:

$$\delta W^s = \delta d_s^T \{ (K - \omega^2 M) d_s - f_s \}$$ (3)

where $K$ is the stiffness matrix, $M$ is the mass matrix, $d_s$ is the vector of nodal displacements and $f_s$ is the applied nodal forces. Equation (3) can be written in a more compact form

$$\delta W^s = \delta d_s^T \{ \bar{K} d_s - f_s \}$$ (4)

defining $\bar{K} = (K - \omega^2 M)$.

2.2 Fluid domain

The governing equation for the linear acoustic fluid domain $\Omega_f$ is known as the Helmholtz equation and can be written as

$$\Delta p + k^2 p = 0$$ (5)

where $\Delta$ represents the Laplace operator, $p$ is the acoustic pressure, $k = \omega/c$ is the wave number, and $\omega$ and $c$ are the angular frequency of the pressure oscillation and the speed of sound traveling in the fluid, respectively. On the boundary $\Gamma_f$ can be prescribed the following boundary conditions:
- Neumann boundary condition
\[ \frac{\partial p}{\partial n} = -i\rho \omega v_n \] (6)

- Rigid boundary
\[ \frac{\partial p}{\partial n} = 0 \] (7)

where \( n \) denotes the unit normal at the surface point, \( \rho \) is the density of medium, \( v_n \) represents the velocity on the boundary surface and \( i = \sqrt{-1} \).

The BEM formulation for acoustic problems is well known and can be found in many classical texts like [12], where the Helmholtz equation (5) is transformed into a boundary integral equation. First, Helmholtz equation is written in a weak form using a weighted residual approach with the following fundamental solution as the weight function
\[ G(x, y) = \frac{e^{-ik|x-y|}}{4\pi|x-y|} \] (8)

In the expression above, \( |x - y| \) is the distance between the collocation point \( x \) and the source point \( y \). Applying Green’s second theorem on the weighted residual and defining the collocation points at the boundary, the resulting boundary integral equation can be expressed as
\[ C(x)p(x) + \int_{\Gamma_f} p(y) \frac{\partial G(x, y)}{\partial n} d\Gamma = \int_{\Gamma_f} G(x, y) \frac{\partial p(y)}{\partial n} d\Gamma \] (9)

where \( C(x) \) is a coefficient which depends on the position of point \( x \): \( C(x) = 1 \) for an internal point, \( C(x) = 1/2 \) for \( x \) on a smooth boundary \( \Gamma_f \), and \( C(x) = 0 \) for a external point.

Equation (9) can be rewritten, taking into account the Neumann boundary condition (6), as
\[ C(x)p(x) + \int_{\Gamma_f} p(y) \frac{\partial G(x, y)}{\partial n} d\Gamma = -\int_{\Gamma_f} G(x, y)i\rho \omega v_n(x) d\Gamma \] (10)

The boundary \( \Gamma_f \) is divided into \( N_e \) elements, \( \Gamma_e \in \Gamma \), so: \( \Gamma_f = \bigcup_{e=1}^{N_e} \Gamma_e \) and \( \bigcap_{e=1}^{N_e} \Gamma_e = \emptyset \). The fields \( p \) and \( v_n \) are approximated over each element \( \Gamma_e \) using shape functions, as a function of the nodal values
\[ p = \sum_{j=1}^{m} N_j p_j = N_f p \quad v_n = \sum_{j=1}^{m} N_j v_{nj} = N_f v \] (11)

\( p_j \) and \( v_{nj} \) being the nodal values acoustics pressure and particle normal velocity at node \( j \), respectively, and \( N_f \) being the shape functions approximation matrix. A discrete linear
The virtual work of a BE fluid domain can be computed using a weak statement for static equilibrium reduced to the boundary, using the Clapeyron formula, 

$$\delta W' = \int_{\Gamma_f} (p - \bar{t}) \cdot \delta u_f d\Gamma$$  

(17)

combined with the discretized BE equation (16). Note that symbol $\delta W$ is used instead of $\delta \Pi$ to express virtual work, emphasizing that in general this variational statement does not derive from an energy functional.

Equation (17) is discretized using the BE mesh to obtain a discrete approximation of the virtual work: 

$$\delta W' = \delta d_f^T (\int_{\Gamma_f} N^T N, d\Gamma) (p - \bar{t}) = \delta d_f^T M (p - \bar{t})$$  

(18)

with

$$M = \int_{\Gamma_b} N^T N d\Gamma$$  

(19)

Substituting the discrete tractions $p$ coming from the BE equations (16) a final expression for the discrete variation is obtained

$$\delta W' = \delta d_f^T \{H^{-1} i\omega G d_f - \bar{t}\}$$  

(20)

As it has been done in previous section, Eq.(20) is written in a more compact form

$$\delta W' = \delta d_f^T \{\tilde{A} d_f - f_f\}$$  

(21)

defining $\tilde{A} = i\omega M H^{-1} G$. 

The particles normal velocity at every node can be computed as

$$v = i\omega d_f$$  

(15)

so Eq. (13) can be expressed in terms of the nodal acoustic pressure and particles normal displacement:

$$H p = \omega G d_f$$  

(16)
3 LOCALIZED LAGRANGE MULTIPLIERS

The virtual work for the interface frame $\delta W_c$ can be also evaluated applying the variational-based formulation proposed by Park and Felippa [3, 4], and González et al. [14]. The virtual work variation of the total system $\delta W_T$ consists of those of both FE structure and BE fluid, $\delta W_s$ and $\delta W_f$, plus of the interface frame $\delta W_c$. This formulation enforces the kinematical positioning of the frame in a weak sense with the following expression

$$W_c = \int_{\Gamma_c} \{\lambda_s^n(u^n_s - u_f) + \{\lambda_f^n(u^n_f - u_f)\}d\Gamma$$

where both integrals are extended to the boundary interface $\Gamma_c$. The localized Lagrange multipliers and the displacements on the structure interface are represented by $(\lambda_s^n, u^n_s)$, and the fluid localized Lagrange multipliers and displacements by $(\lambda_f^n, u^n_f)$. Finally, the frame displacements are represented by $u_f$.

Equation (22), can be written in a matrix form as follows

$$\delta W_c = \delta \{\lambda_s^T(B_s^Td_s - L_su_f)\} + \delta \{\lambda_f^T(B_f^Td_f - L_fu_f)\}$$

$$\delta u_f^T\{L_f^T\lambda_f + L_f^T\lambda_f\}$$

The stationary point of the above virtual work expression leads to the following coupled equilibrium equation set:

$$\begin{bmatrix} \bar{K} & 0 & B_s & 0 & 0 \\ 0 & \bar{A} & 0 & B_f & 0 \\ B_s^T & 0 & 0 & 0 & L_s \\ 0 & B_f^T & 0 & 0 & L_f \\ 0 & 0 & L_f^T & L_s^T & 0 \end{bmatrix} \begin{bmatrix} d_s \\ d_f \\ \lambda_s \\ \lambda_f \\ u_f \end{bmatrix} = \begin{bmatrix} f_s \\ f_f \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The total virtual work of the coupled BEM-FEM-Frame system can now be expressed as

$$\delta W_T = \delta d_s^T\{\bar{K}d_s + B_s\lambda_s - f_s\} + \delta d_f^T\{\bar{A}d_f + B_f\lambda_f - f_f\}
+ \delta \lambda_s^T\{B_s^Td_s - L_su_f\} + \delta \lambda_f^T\{B_f^Td_f - L_fu_f\}
+ \delta u_f^T\{L_f^T\lambda_f + L_f^T\lambda_f\}$$

The stationary point of the above virtual work expression leads to the following coupled equilibrium equation set:

$$\begin{bmatrix} K & B & 0 \\ B^T & 0 & L \\ 0 & L^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \lambda \\ u_f \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

If we have $N_p$ fluid and structure partitions, equation (25) can be written as
where

\[ \mathbf{B} = \text{diag} [\mathbf{B}^{(1)} \ldots \mathbf{B}^{(N_p)}] \quad \text{with} \quad \mathbf{B}^{(p)} = \begin{cases} \mathbf{B}_i^{(p)} \quad \text{(FEM)} \\ \mathbf{B}_f^{(p)} \quad \text{(BEM)} \end{cases} \] (27)

\[ \mathbf{L} = \begin{bmatrix} \mathbf{L}^{(1)} \\ \vdots \\ \mathbf{L}^{(N_p)} \end{bmatrix} \quad \text{with} \quad \mathbf{L}^{(p)} = \begin{cases} \mathbf{L}_i^{(p)} \quad \text{(FEM)} \\ \mathbf{L}_f^{(p)} \quad \text{(BEM)} \end{cases} \] (28)

\[ \lambda = \begin{bmatrix} \lambda^{(1)} \\ \vdots \\ \lambda^{(N_p)} \end{bmatrix} \quad \text{with} \quad \lambda^{(p)} = \begin{cases} \lambda_s^{(p)} \quad \text{(FEM)} \\ \lambda_f^{(p)} \quad \text{(BEM)} \end{cases} \] (29)

Eliminating \( \mathbf{d} \) from the first row of (26) using the relation

\[ \mathbf{d} = \mathbf{K}^{-1}(\mathbf{f} - \mathbf{B}\lambda) \] (30)

the following flexibility matrix equation is obtained

\[ \begin{bmatrix} \mathbf{F}_{bb} & \mathbf{L} \\ \mathbf{L}^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{u}_f \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \] (31)

being \( \mathbf{F}_{bb} = \mathbf{B}^T\mathbf{K}^{-1}\mathbf{B} \) and \( \mathbf{b} = \tilde{\mathbf{B}}^T\mathbf{K}^{-1}\mathbf{f} \).
4 ITERATIVE SOLUTION ALGORITHM FOR THE INTERFACE PROBLEM

In this section the solution strategy is presented to solve the flexibility system obtained for the FSI localized Lagrange multipliers formulation, together with some studies of convergence and scalability.

4.1 LLM coupled system

The algorithm uses a decomposition of the interface solution vector in the form:

\[ \lambda = P_L \lambda_d \]  

with symmetric projector

\[ P_L = I - L(L^T L)^{-1} L^T \]  

such as \( P_L L = 0 \)

Substituting this decomposition into the flexibility formulation of the interface (31) yields the following equation set:

\[ P_L F_{bb} P_L \lambda_d = P_L b \]  

The projected residual is finally given by

\[ r = P_L (b - F_{bb} P_L \lambda_d) \]

and it is solved for \( P_L \lambda_d \).

Iterate on the projected residual: \( r = P_L (b - F_{bb} \lambda) \), using the proposed preconditioned projected bi-conjugate gradient algorithm:

(I) Initialize: \( \lambda_0, \ r_0 = P_L (b - F_{bb} \lambda_0), \ x_0 = 0, \ p_0 = 0 \).

(II) Iterate \( i = 1, 2, 3... \) until convergence:

- Compute: \( p_i = r_{i-1} + \omega_i (p_{i-1} - \alpha_{i-1} x_{i-1}) \) being \( p_1 = r_0, \ \beta_i = \text{Re}\{(r_0^T r_{i-1})\}, \) and \( \omega_i = \beta_i \gamma_{i-1}/(\alpha_{i-1} \beta_{i-1}) \).
- Precondition: \( a_i = \tilde{F}_{bb}^t p_i \).
- Projection: \( z_i = P_L a_i \).
- Compute: \( v_i = r_{i-1} - \gamma_i x_i \), being \( b_i = F_{bb} z_i, \ x_i = P_L b_i, \) and \( \gamma_i = \beta_i / \text{Re}\{(x_i^T r_0)\} \).
- Precondition: \( c_i = \tilde{F}_{bb}^t v_i \).
- Projection: \( y_i = P_L c_i \).
- Update solution: \( \lambda_i = \lambda_{i-1} + \gamma_i z_i + \alpha_i y_i \), being \( g_i = F_{bb} y_i, \ w_i = P_L g_i, \) and \( \alpha_i = \text{Re}\{(w_i^T v_i)\}/\text{Re}\{(w_i^T w_i)\} \).
Update residual: \( r_i = v_i - \alpha_i w_i \).

(III) If \( \|r_i\|/\|r_0\| > \epsilon \), \( i \leftarrow i + 1 \) go back to (II).

The preconditioner proposed is based on extensions of the well-known lumped and Dirichlet preconditioners in the standard FETI and AFETI algorithms. These preconditioners are defined as

\[
\tilde{F}_{bb}^+ = \begin{cases} 
\tilde{K}_{bb} & \text{(FEM)} \\
\tilde{A}_{bb} & \text{(BEM)} 
\end{cases}
\]  \( (36) \)

where subscript (\( bb \)) refers to boundary extraction, i.e. pre and post multiplication by \( B^T \) and \( B \) respectively.

4.2 Benchmark application: flexible wall and acoustic cavity

The coupling possibilities of LLM methodologies are studied in the section, solving a benchmark problem: a two dimensional \( L_x \times L_y \) cavity (\( L_x = 10m \) and \( L_y = 4m \)), with one flexible side (see Fig. 2). The flexible wall is modeled by the FEM, using beam elements, and simply supported on both edges. The properties of this structural domain are: Young modulus \( E = 2.1 \times 10^{11} \) \( Pa \), beam section inertia \( I = 1.59 \times 10^{-4} \) \( m^4 \) and cross section area \( A = 0.02 \) \( m^2 \), and a mass per unit length \( m_s = 50 \) \( kg/m \). The remaining edges of the cavity are reverberant walls, i.e. homogeneous Neumann boundary conditions are applied (\( v_n = 0 \)). The acoustics fluid is water being \( c_f = 1500 \) \( m/s \) and \( \rho = 1000 kg/m^3 \). This problem is presented in Fig. 2 where an oscillatory moment \( M_{exc} = 1 Nm \) is applied on one edge. In Fig. 3 is presented an scheme of the meshes and the coupled BEM-FEM subdomains using LLM.

The acoustic cavity with a flexible wall is solved using the nsBETI iterative algorithm. The results agree with the vibration modes of the flexible wall acoustic cavity presented...
The influence of different factors like the number of elements per subdomain, the frequency of the excitation, and the presence of non-matching interfaces are examined in the convergence of nsBETI algorithm. The BiCGSTAB error evolutions in every case are presented in Fig. 4 for different number of elements. It can be observed how the number iteration is not affected by the number of degrees of freedom, but it is very effected by the harmonic excitations frequency. Initially, each subdomain is discretized using BEM-FEM matching meshes with \( L/h = 32, 64, 128, \) and \( 256 \) divisions at the interface. Figures 4(a) and 4(b) shows the error evolutions for a low frequency excitation of 5 Hz and high frequency of 80 Hz, respectively, with the number of iterations needed by nsBETI to solve these problems with a tolerance of \( 10^{-10} \). It can be observed, for the cases considered, that an exponential increase of the type \( H/h = 2^n \) translates into a constant increment of the number of iterations, for every excitation frequency. The difference between the number of iteration in every case (5 Hz and 80 Hz) is due to the differences between their deflections, as Fig.5 shows.

Finally, the non-matching case is considered by changing the discretization of the structure to produce dissimilar meshes at the interface. Figure 6 presents the error evolutions for 5 Hz (Figure 6(a)) and 80 Hz (Figure 6(b)) excitation. The results obtained when the mesh of structure ranged from \( (L/h) = 64 \) (highly non-matching case) to \( (H/h)_{BEM} = 256 \) (matching case), maintaining the mesh of fluid fixed with \( (L/h) = 256 \) divisions. It is noted that the introduction of dissimilar meshes, maintaining a constant \( (L/h)_{max} \), increases the number-of-iterations needed by nsBETI to solve the problem for low and high excitation frequencies.

As a summary, we can say that, in the matching case, the convergence of nsBETI algorithm is governed by \( (L/h)_{max} \) but the introduction of non-matching interfaces destructs this property, producing a negative effect in the convergence that is controlled by the
Figure 4: BiCGSTAB error evolution for: (a) 5 Hz and (b) 80 Hz, considering a LLM coupling of matching meshes.

interface mesh-dissimilarity parameter $h_{\text{max}}/h_{\text{min}}$.

5 CONCLUSIONS

A FETI-type solution algorithm (nsBETI algorithm) has been extended to treat non-matching and non-symmetrical FE-BE acoustic FSI problems, which enjoys similar scalability properties of those of classical FETI and symmetrical-BETI algorithms. This scheme of resolution is based the LLM methodology which allows to consider non-matching interfaces, and preserves software modularity. The scalability studies have been studied for the cases of dissimilar meshes at the interfaces. In the matching interface case, convergence of nsBETI algorithm is governed by $(L/h)_{\text{max}}$, but the introduction of non-matching interfaces produces a negative effect in the convergence that is controlled by the interface mesh-dissimilarity parameter $h_{\text{max}}/h_{\text{min}}$.

6 ACKNOWLEDGEMENTS

This work was co-funded by the DGICYT of Ministerio de Ciencia y Tecnología, Spain (research project DPI2010-19331) and by the Consejería de Innovación Ciencia y Empresa de la Junta de Andalucía, Spain (research project P08-TEP-03804), which are co-funded with the European Regional Development Fund (ERDF) (Fondo Europeo de Desarrollo Regional, FEDER).

REFERENCES

Figure 5: Beam deflection due to harmonic excitations: (a) 5 Hz and (b) 80 Hz, considering a LLM coupling of matching meshes.

Figure 6: BiCGSTAB error evolution considering non-matching meshes and harmonic excitation of: (a) 5 Hz and (b) 80 Hz.


NUMERICAL COUPLING PROCEDURE IN STEADY CONJUGATE HEAT TRANSFER PROBLEMS

M.-P. ERRERA AND S. CHEMIN

* ONERA, The French Aerospace Lab
29, Avenue de la Division Leclerc – 92322 Châtillon Cédex, France
e-mail: marc.errera@onera.fr, www.onera.fr

Key words: Coupled Problems, Multiphysics, Heat Transfer, Stability, Convergence.

Abstract. This paper analyses the numerical stability of a coupling procedure between a CFD code and a conduction solver in a partitioned approach. A finite volume method is used in the fluid partition and a finite element method in the solid partition. Since our goal is to get a global fluid-solid solution, the analysis of the transient in the solid is not of particular interest. Consequently, the numerical method is based on the coupling of a steady state in the solid with a time-dependent solution in the fluid. At the shared interface, Dirichlet (on the fluid side) and Robin (on the solid side) conditions are applied. An interface stability study is performed according to the normal-mode analysis of the theory of Godunov-Ryabenkii. The existence of an optimal coupling parameter is highlighted.

1 INTRODUCTION

Conjugate heat transfer (CHT) problems are encountered in many real-world applications such as cooling systems, building heating, ventilating flows, heat-exchanger equipment, etc. They arise in situations where considering only heat transport in the fluid is not sufficient. In such cases we must take into account the fully coupled problem, including conduction in the solid wall adjacent to the fluid.

CHT is now widely used in engineering applications, but in most cases, arbitrary relaxation parameters or reference temperatures are used to stabilize the coupling procedure. This may have a significant negative impact on the convergence rate. Our intention in this paper is to recall that in a CHT problem based on Dirichlet-Robin interface conditions, there is an optimal coefficient in terms of stability and convergence. This coefficient has been highlighted recently [1] using a 1D model equation.

This article is organized as follows. First the model equations of the CHT procedure are described briefly. Then, the coupling interface conditions and the coupling algorithm are presented. After that, the stability analysis according to the theory of Godunov-Ryabenkii is summarized. This analysis provides an optimal coupling coefficient. This coefficient is given and its influence on stability and convergence is discussed.

2 MODEL EQUATIONS

2.1 Time-scale discontinuity

The CHT strategy presented in this paper is motivated by the desire to obtain rapidly a global fluid-structure steady solution. There is a significant discrepancy of the characteristic times, namely a fast transient process in the fluid, a very slow one in the structure. Hence, the thermal response of the solid can be very long (several hours or minutes). Since our goal is to
get a global fluid-solid solution, we must recognize that the analysis of the transient in the solid is not of particular interest.

2.2 **Solid solution**

If we want to obtain a steady-state solution of the heat equation in the solid domain, it is always possible to choose some initial data and march forward in time. But this approach is typically not an efficient way to compute the steady state solution, if this is all we want. Instead, if the boundary conditions are time independent and if we are interested only in computing the steady-state solution itself, then we can solve directly a second order ordinary differential equation (Laplace equation). As this procedure performs well in the solid when it is considered independently as a single subsystem, we should be able to use it in a coupled system.

2.3 **Fluid solution**

The same cannot be said for the fluid subsystem. In the vast majority of cases, the Navier-Stokes (NS) equations are solved to steady-state by a time marching scheme. As a consequence, a time marching scheme will be employed in the fluid domain in our CHT model. As physically realistic time-dependent solutions are not sought, the unsteady fluid and steady solid procedures will be interactively solved and coupled until a global fluid-solid solution is obtained.

2.4 **Fluid-solid solution**

The simulation of this type of problems is generally accomplished by partitioned staggered schemes [2][3]. As mentioned above, we just want to obtain a stable global solution by coupling a transient fluid solution with a steady solid state. But solving a Laplace equation at each time step could change the solution too rapidly for stability to be maintained. The goal of this paper is to provide an optimal coupling coefficient that never introduces stability restrictions.

3 **COUPLING CONDITIONS AT THE FLUID-SOLID INTERFACE**

3.1 **Discretized model**

The 1D discrete model is presented in Figure 1. This model is based on a finite-volume (FV) procedure on the fluid-side \((x > 0)\) and a finite element (FE) procedure on the solid-side \((x < 0)\).
3.2 Dirichlet-Robin procedure

At the common interface (x = 0), coupling conditions are applied. Roux & Garaud [4] studied recently the behavior of interface conditions in a steady CHT problem. They first generalized the result obtained by Giles [5] and confirmed that Dirichlet conditions must be imposed in the fluid domain. Then, they pointed out from the Schur complement that the Dirichlet-Neumann condition may suffer considerably from destabilizing effects. On the contrary, it is indicated in the same work, that the Robin condition has many attractive features. First it can always be formulated in such a way that the associated local problem is well posed even though the Neumann problem is not. Second, the use of such a condition introduces an interface stiffness forcing the boundary to behave in the same way as the boundary of the other domain. Finally, it provides much better stability properties. Consequently a Dirichlet-Robin procedure will be considered here. In other words, in the first step of the coupling procedure in the interval \([0, N]\), the temperature coming from the solid is applied on the interface “0+“ of the fluid domain

\[ T_{0+}^{n+1} = T_{0-}^n \] (1)

A numerical Robin condition is in turn used as a boundary condition for the interface “0-“ of the solid domain

\[ q_j^{n+1} - \alpha_f^{n+1} T_{0-}^n = -q_j^{n+1} - \alpha_f^{n+1} T_{0+}^{n+1} \] (2)

The subscripts \(f\) and \(s\) denote the fluid and solid domain respectively. \( q \) is the heat flux (W.m\(^{-2}\)). The general Dirichlet-Robin condition considered in this paper introduces the numerical coupling parameter \( \alpha_f \) (W.m\(^{-2}\).K\(^{-1}\)) the choice of which directly influences the stability of the coupled process. The expression of an optimal coefficient will be given in this paper. The coupling procedure is briefly described in the next section.
3 CSS PROCEDURE

Many partitioned staggered procedures have been proposed to solve fluid-structure interaction problems. They can be categorized as either strongly-coupled or loosely-coupled. The strongly-coupled schemes involve predictor-corrector iterations and then increase the complexity of the implementation of a CHT problem as well as the computational cost at each time step. In this paper, we have employed the basic loosely-coupled conventional serial staggered (CSS) algorithm \[6\] whose generic cycle is described in Fig. 1 in the time interval \([t^n, t^{n+1}]\) with time step size \(\Delta t = t^{n+1} - t^n\), where \(n\) corresponds to the coupling time level. This procedure is repeated until a global steady-state solution is obtained.

Our goal is not to discuss the pros and cons of coupling algorithms, but to present a stability analysis and a resulting ideal coupling parameter. For this, we need a basic algorithm and naturally, we choose the simplest. Developments that follow remain valid if we adopt another one. However, the results may vary slightly.
4 NUMERICAL TREATMENTS

4.1 Numerical schemes

As mentioned previously, the fluid domain is discretized with a FV scheme and the solid domain with a FE scheme. On the interior mesh points we employ an implicit Euler-backward scheme for the time derivative. The diffusive term (heat flux) is computed in the same manner as the viscous terms in the NS equations, that is a second order accurate central difference formulation. The numerical treatments are summarized in Table 1.

Table 1: Numerical treatments of the CHT procedure

<table>
<thead>
<tr>
<th>FLUID DOMAIN</th>
<th>SOLID DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time stepping</td>
<td>Implicit backward</td>
</tr>
<tr>
<td>Numerical approach</td>
<td>Finite Volume</td>
</tr>
<tr>
<td>Spatial scheme</td>
<td>2nd order central</td>
</tr>
<tr>
<td>Fluid-Solid Interface</td>
<td>Dirichlet</td>
</tr>
</tbody>
</table>

4.2 Discretized system

The numerical discrete CHT system is obtained after discretization of the diffusion equation & Dirichlet conditions on the fluid side and the Laplace equation & Robin conditions on the solid side. The resulting discrete system is obtained

\[
\begin{align*}
\rho C_f \frac{\Delta T_j}{\Delta t} & = \frac{k_f}{\Delta x_j} (T_{j+1}^{n+1} - 2T_j^{n+1} + T_{j-1}^{n+1}) & (a) \\
T_{0+}^{n+1} & = T_0^n & (b) \\
T_{j+1}^{n+1} & - 2T_j^{n+1} + T_{j-1}^{n+1} = 0 & (c) \\
k_f \frac{1}{\Delta x_s} (T_{0+}^{n+1} - T_{0-}^{n+1}) + \alpha_f T_{0+}^{n+1} & = \frac{k_f}{\Delta x_f / 2} (T_1^{n+1} - T_{0+}^{n+1}) + \alpha_f T_{0+}^{n+1} & (d)
\end{align*}
\]

where \( T_j^n = T(j \Delta x, n \Delta t) \). \( \rho \) is the density, \( C_f \) is the specific heat, \( k_f \) is the thermal conductivity of the fluid.
5 OPTIMAL COUPLING COEFFICIENT

5.1 Stability analysis and amplification factor

This stability analysis is very similar to the standard Fourier stability method except that the Fourier analysis ignores boundary conditions and as these may affect the stability, the theory of Godunov & Ryabenkii (G-R) [7][8] is preferable. We introduce the normal mode solution for the case defined by the discretized system (3) by considering eigensolutions of the form

\[
T^n_j = \begin{cases} 
  z^{n-1} \kappa_j^f, & j > 0 \\
  z^n \kappa_j^s, & j \leq 0
\end{cases}
\]

(4)

where \( z \) is the “temporal amplification factor” and \( \kappa \) is the “spatial amplification factor”. The discretized model (3) is stable in the sense of G-R if no solutions of the form (4) are admitted with \( |\kappa_j^f| < 1, |\kappa_j^s| > 1 \), and \( |z| > 1 \) as \( j \to \pm \infty \). Moreover, we exclude modes (4) with \( |z|=1 \) (neutrally stable modes).

Introducing (4) into the interface conditions (3), after elementary transformations, we obtain the following temporal amplification factor

\[
z = g(z) = \frac{k_j}{\Delta x_j} \left[ \frac{z - 1}{Dz} - \sqrt{\left( 2 + \frac{z - 1}{Dz} \right)^2 - 4} + \alpha_f \right] \left/ \left( \frac{\beta k_s}{\Lambda_s} + \alpha_f \right) \right.
\]

(5)

where \( D \) is the Fourier number defined as follows

\[
D = \frac{k_j \Delta t}{\rho C_j \Delta x_j^2}
\]

(6)

Without going into too much detail, let us mention that the parameter \( \beta \) in Eq. (5) accounts for the contributions to physical and geometric solid characteristics and controls the external boundary condition (see Fig. 1). The influence of this parameter on the global CHT procedure is crucial.

5.2 Stability zones

The function \( \max_{z \geq 1} \{ g(z) \} \) has been plotted in Figure 3 in terms of \( \alpha_f \) for different Fourier numbers \( D \). We have chosen typical fluid and solid physical parameters corresponding to air and steel respectively. We can observe that this function is defined and continuous and that each curve is composed of two half-lines with a singular point similar to a cusp at the intersection. At this point, \( \max_{z \geq 1} \{ g(z) \} \) attains its minimum value.
In other words, the existence of a transition value for $\alpha_f$ is highlighted. At this transition value, the shape of the curve switches and turns back abruptly. This value is denoted $\alpha_f^{(opt)}$.

### 5.3 Optimal coefficient

It can be shown [1] that the transition occurs at a unique and remarkable value $\alpha_f^{(opt)}$ whose exact expression is given by

$$\alpha_f^{(opt)} = k_f \Delta x_f \sqrt{(1 + \frac{1}{D})^2 - 1 - \frac{1}{D}}$$  \hspace{1cm} (7)

When the optimal value defined by (7) is employed, we obtain the best-case scenario with no additional computational effort. The point $\alpha_f^{(opt)}$ is the intersection of two opposite zones. The left half-line ($\alpha_f < \alpha_f^{(opt)}$) is controlled by Neumann conditions. It is a fast process prone to instability. The right half-line ($\alpha_f > \alpha_f^{(opt)}$) is controlled by Dirichlet conditions. It is a low but always stable process. The intersection $\alpha_f^{(opt)}$ is a perfect equilibrium between both.
5.4 Energy, stability and optimal coefficient

First of all, let us recall that the Fourier number used in unsteady-state flow problems, whose expression is given by (6), is a dimensionless number that characterizes heat conduction:

\[
D = \frac{\text{heat conduction rate}}{\text{rate of thermal energy storage}}
\]

The various curves in Fig. 3 can be interpreted easily in terms of the Fourier number \(D\) and the coupling coefficient \(\alpha_f\).

Low values of \(\alpha_f (\alpha_f < \alpha_f^{(opt)})\) will result in a rapid convergence, but in this case however \(D\) must be sufficiently large to allow heat diffusion on the fluid side. Otherwise, a low Fourier number \(D\) will soak up a lot of heat. It will be then necessary to enhance stability by increasing the coupling parameter. Thus, the energy transfer will be “frozen” by an increase of \(\alpha_f\).

But, likewise, large values of \(\alpha_f (\alpha_f >> \alpha_f^{(opt)})\) will always lead to an extremely slow convergence. This is only consistent with small values of \(D\) and corresponds to a slow diffusion of heat through the fluid subsystem. But it should be pointed out that relatively large Fourier numbers \(D\) indicate fast propagation and energy in this case will be unnecessarily frozen by \(\alpha_f\). As a result, this will have a very negative impact on the computing time.

All these situations might happen in the same coupled computation. It is the reason why it is crucial to use a local coupling coefficient and it has been shown that \(\alpha_f = \alpha_f^{(opt)}\) is the optimal choice in the case of the model equation adopted herein.

6 CONCLUDING REMARKS

We have shortly presented a stability analysis of a CHT problem using a Dirichlet-Robin procedure. An optimal coupling coefficient has been highlighted and discussed. It is a dynamic parameter with no increase in computational effort. The resulting coupling method can be regarded as an adaptive procedure to always obtain the fastest rate of convergence and the best stability properties.

There are many other points that would have deserved special attention. First of all, the influence of the parameter \(\beta\) that governs the external boundary condition and mimics the geometric and physical behavior of the solid domain. This key parameter has a direct impact on the global stability. The other important study to be carried out is to extend the present analysis to a general Robin-Robin procedure. These points, not presented here, have been thoroughly considered in a recent paper in the Journal of Computational Physics [1].

REFERENCES


RESEARCH ON ELECTROMECHANICAL FIELD COUPLING MODEL AND SOLUTION STRATEGY OF LARGE REFLECTOR ANTENNAS

Meng Wang and Wei Wang

Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi’an, Shaanxi 710071, China
e-mail: lemonwm@163.com

Key words: reflector antennas, electromechanical coupling model, solution strategies, finite element analysis

Abstract. The overriding design goal of reflector antennas is accomplishing an acceptable electromagnetic performance. However, the success of this goal lies — for a substantial part — in the construction of mechanical structure, which ensures and restricts the achievement of its electrical properties simultaneously. The reflector antenna is a system that combines electromagnetic fields with structural deformation fields, in which electromechanical coupling problems exist. The high precision expected makes the coupling phenomenon more serious as reflector antennas develop towards a situation of large diameter and high band. Therefore it is necessary to do some research about the electromechanical coupling problems, establish precise field coupling model and get the solution strategies. This paper presents a electromechanical field coupling model based on the method of surface current. The model includes the systematic error, with considering all-around displacement of each node, based on which the interior surface deformation of elements could be got by the interpolation with shape function. The vertex normals of distorted curve are then applied to the surface current density to investigate its impact on the far-field electrical performance. The investigation of the coupling model and solution strategy led to an improvement of design level and to a better performance of reflector antennas.

1. INTRODUCTION

Reflector antennas, as the microwave and millimeter wave high-gain antenna, are widely used in the field of communication, radar, navigation, radio astronomy and meteorology, etc. [1]. The reflector antenna is a system that combines electromagnetic fields with structural deformation fields, in which electromechanical coupling problems exist. With the amplitude and scope of deformation becoming closer to the wavelength, the coupling phenomenon gets more significant when the antenna towards the direction of larger diameter and higher band. So it is necessary to study the electromechanical coupling problems, establish accurate
electromechanical coupling model, find the solving strategies and methods close to the actual case, accurately predict the effect made by the deformation to electrical properties, in order to provide a reference for its compensation.

The errors of reflector surface include random error and systematic error. The systematic error is the deformation of antennas under the influence of weight, environmental temperature, wind and other external environmental loads. It would be a deterministic error if not considering the random wind load and could be gained by finite element analysis. Random error is brought in the process of manufacturing and assembly of panel, back-up structure and central body. Some researchers home and abroad have studied the effect of errors to the electrical performance of reflector antennas. For example, the literature [2-7] studied the effect of random error to the average power pattern of reflector antennas and drew some meaningful conclusions. As to systematic errors, the literature [8-9] analyzed the deformed surface with the best-fitted parabolic method (BFP). However, but this approach does not fully reflect the local deformation, not to mention its special effect to electrical performance of the antenna, based on which, the Coons-surface blocked fitting method proposed by the literature [10-11] reflects the deformation of the reflecting surface comprehensively. Literature [12] investigated the effect of partial deformation to the reflector antenna’s electrical performance, but it is not general without from the view of field coupling. Literature [13] studied the impact of deformation to the electrical performance of antennas under different Taylor series, with approximating the surface errors using the superposition by a series of surface-expanded function, unfortunately, the method is effective only in the accuracy range of 0.1\(\lambda\). The above literature, or being fitted based on the structural deformation data, or analyzing the surface deformation using the experience approximately, all caused errors more or less to the electrical performance forecast. Literature [14] studied the effect of random error and systematic error to the reflector antenna electrical performance from the view of electromechanical coupling, but limited on the impact of antenna efficiency.

The electromechanical field coupling model is presented in this paper based on the method of surface current. The deformation of antennas under the environmental loads could be got with element analysis software. Then we can obtain the interior surface deformation of elements by the interpolation with shape function to meet the grid accuracy when calculating the electrical performance. Finally, the structural displacement field of the whole reflector surface could be got. Therefore it is helpful for the electromagnetic design and structural design of antennas to analyze the effect of structural errors to electrical performance quantitatively based on the electromechanical coupling model established before. So we can accomplish general excellence of mechanical and electromagnetic performance of microwave antenna. In addition, the model could also provide certain reference for other electronic equipments.

2. ESTABLISHMENT OF ELECTROMECHANICAL TWO COUPLING MODEL

The surface current method and aperture field method are often used to analyze the
radiation patterns of reflector antennas. Some scholars [6] have established the electromechanical two coupled model based on the latter method. In this study, the two coupling model established based on the surface current method is addressed, and the corresponding solution strategy. As is shown in Fig. 1, the radiation integral formulation for the ideal reflector antenna is given by

$$\vec{E} = \iint_{S} \vec{J}^e_s \cdot e^{jkr} \hat{n} dS$$

Where $\vec{J}^e_s = 2\hat{n} \times \vec{H}$ is the current density of the reflector surface. The vector $\vec{r}'$ locates the integration points, the unit vector $\hat{n}_0$ is in the observation direction.

![Figure 1: geometric relation of paraboloid](image1)

However, because of the working environment of antennas, the design structure of antennas will be influenced by the external loads, thus affecting the electrical properties, which is the coupling of structural displacement field and electromagnetic field. Therefore, the additional items caused by structural displacement should be considered in the formula (1). The reflector surface will be deformed under the influence of external environmental loads (see Fig. 2), which deviate from the original ideal design surface. The vector that locates the integration point on the distorted surface is $\vec{r}'_d = \vec{r}' + \vec{d}$. Where, $\vec{d}$ is the difference of the corresponding points between the distorted surface and the desired undistorted reflector, which could be got by the stiffness equation $Kd = F$.

![Figure 2: defining the vector that locate the points on the distorted reflector and the undistorted reflector](image2)
Since \( |r_d'| \gg |d| \), \( r_d \) can be approximated as

\[
r_d = |r_d'| = |r'| + \frac{d \cdot r'}{|r'|}
\]

Using the amplitude approximation, the current-term \( \mathbf{j}_{sd}' \) could be approximated by

\[
\mathbf{j}_{sd}' = 2\mathbf{n}_d \times \mathbf{H}_s (r') \approx 2\mathbf{n}_d \times \mathbf{H}_s \cdot e^{-\frac{j \cdot d \cdot r'}{|r'|}} = \mathbf{j}_{sd} \cdot e^{-\frac{j \cdot d \cdot r'}{|r'|}}
\]

Using (2), (3) in (1) gives

\[
\mathcal{E} \approx \int_S \mathbf{j}_{sd} \cdot e^{-\frac{j \cdot d \cdot r'}{|r'|}} \cdot e^{j \cdot \mathbf{n}_d \cdot \mathbf{r}'} dS
\]

Also, because

\[
e^{j \cdot \mathbf{n}_d \cdot \mathbf{r}'} = e^{j \cdot (\mathbf{r}' \cdot \mathbf{n}_d + d \cdot \mathbf{n}_d)} = e^{j \cdot (\mathbf{r}' \cdot \mathbf{n}_d + d \cdot \mathbf{n}_d)}
\]

Using this in (5) and substituting into (4) gives

\[
\mathcal{E} \approx \int_S \mathbf{j}_{sd} \cdot e^{j \cdot \mathbf{n}_d \cdot \mathbf{r}'} \cdot e^{j \cdot (\mathbf{r}' \cdot \mathbf{n}_d + d \cdot \mathbf{n}_d)} dS
\]

which is the electromechanical two coupled model for the distorted reflector antenna, where \( \mathbf{j}_{sd} \) takes into account the impact of changes in surface normal vector and the irradiation magnetic field distribution from the feed caused by the deformation.

3. SOLVING STRATEGIES AND METHODS OF ELECTROMECHANICAL COUPLING MODEL

The establishment of electromechanical coupling model is just a basis, how to solve it accurately is still a key issue. Because the structural displacement data in the model is from the finite element analysis, the results of which are discretized, the double integral expression can only be calculated approximately using numerical integration methods. This calculation will need two sets of different grids, the structural grids and electromagnetic grids, the requirements to which in the form and number are different, resulting in the mismatch between grids. The electromagnetic calculating couldn’t be done based on the grid data got from the structural analysis, which can preserve the details of the structural deformation, but the accuracy certainly can not meet the requirements, because the latter calculation requires a denser and more uniform grid. Some scholars use the data from the structural analysis to fit the deformed reflector surface, upon which the electromagnetic grid is re-divided, then the calculation is done. The disadvantage of this approach is the introduction of a new fitting
error to cause the mismatch of the calculated with actual deformation, which makes the calculated results of electrical performance different from the actual case. The innovation of this paper is data interpolation to the structural displacement field with the shape function of elements in the element analysis, which is not only in line with the actual deformation of the structure, but also to meet the precision requirements of electromagnetic analysis. The specific process is detailed below:

(1) The finite element model of the reflector antenna need to be built first, then we can get the deformation information of the antenna structure after applying the load and structure analysis according to the working environment. This deformation information should contain the full displacement \( \bar{\mathbf{d}} \) of all nodes.

(2) First, the structural deformation information \( \mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k \) should be extracted (see Fig. 3), then we can get the actual surface deformation of all elements by selecting the corresponding shape function to interpolate to obtain all the interior displacement

\[
\mathbf{d}_{ii} = N_i \mathbf{d}_i + N_j \mathbf{d}_j + N_k \mathbf{d}_k.
\]

(3) Based on the original element, the following rules is abided to fractionize the triangular elements automatically. That is, on the basis of original element, node number and node coordinate, we can fractionize each triangular element to four smaller and similar triangular elements again and again, until meeting the precision requirements of electromagnetic calculation. The vertex normals of distorted curve and feed are then applied to get the distribution of the surface current density.

(4) Set the primary feed pattern function or caliber field distribution function in accordance with the discrete accuracy of the far-field determined by the main beam width of the antenna.

(5) With the fields coupling model established before, we can get the get the far-field pattern and main indexes of electrical performance.
4. SIMULATION ANALYSIS AND EXPERIMENTAL VERIFICATION

4.1 Experimental object

We use a Cassegrain dual-reflector antenna with 3.7m-diameter (see Fig. 4), C/Ku dual frequency band and ring focus to verify the correctness of two-field coupling model. The sub-reflector surface has a diameter of 0.44m, focus-diameter ratio of 0.35, operating frequency of 12.5GHz, feed of dielectric loaded horns, the distribution function of caliber field is

\[
f_{r'}(\rho') = \begin{cases} 
1 - 0.9 \exp(\rho'^2 - 1), & \rho' \geq 0.5 \\
1 - 0.85 \exp(0.13 - \rho'), & \rho' < 0.5 
\end{cases}
\]

Where \( \rho' \in [0,1] \) is the normalized radius. Therefore, we can get the equivalent normalized pattern function of the feed

\[
F_\xi(\xi) = \begin{cases} 
\frac{1}{\cos^2(\xi/2)} \left( 1 - 0.9 \exp \left( \frac{\tan \xi}{\tan \xi_0} \right)^2 - 1 \right), & \xi \geq \xi' \\
\frac{1}{\cos^2(\xi/2)} \left( 1 - 0.85 \exp \left( 0.13 - \frac{\tan \xi}{\tan \xi_0} \right) \right), & \xi < \xi' 
\end{cases}
\]

Where \( \xi_0 = 2 \arctan \left( \frac{D}{4f} \right) \) is the aperture angle of the antenna and \( \xi' = \arctan \left( 0.5 \tan \xi_0 \right) \)

(a) front view                      (b) back view
Figure 4: a C/Ku band 3.7m-diameter reflector antenna

The main antenna reflector surface consists of 12 the same fan-shaped panels (see Fig. 5) and its back frame includes 12 radial beam, 1 ring beam and central body. Each fan-shaped panel connects with the radial beam by 13 bolts and L-ring beam with 5 bolts (see Fig. 6).
4.2 Experimental procedures and analysis of the results

To verify the correctness of the electromechanical coupling theory, the premise is to get the deformation information of the reflecting surface deformation under its own weight and other environmental loads. As the self-weight deformation of the selected test subjects, the 3.7m antenna, is very small, the difference of the electrical properties is not obvious with the direct measurement and calculation.

So, we draw the idea in the literature [15] that makes the deformation with pasting metal rings to simulate the impact of different surface errors to the electrical performance of antenna. The reflecting surface is divided into 3 rings from inside to outer and every ring is given the metal shim with different thickness to make the panel away from its original position, thereby generating different distribution of deformation errors. As shown in Fig. 7, there are gaskets with 0.5mm, 1mm and 2mm thickness to make the experiment, which is measured with a micrometer (see Fig. 8).

The far-field pattern of reflector antennas could be got with the electromechanical
two-field coupling model. Table 1 shows the comparison of the calculated and measured results when the inner, middle and outer panels are given the gaskets with the 0.5mm, 1mm, 2mm thickness separately. The figures show that the measured and calculated results are very close in the vicinity of the main lobe, slightly different around the first side lobe, and the difference is gradually increased with the extension of the side lobe.

Figure 9: calculated and measured radiation pattern for a reflector antenna
<table>
<thead>
<tr>
<th>Case</th>
<th>Simulated (Gain(dB))</th>
<th>Simulated (3dB)</th>
<th>Simulated Left Side Lobe</th>
<th>Simulated Right Side Lobe</th>
</tr>
</thead>
<tbody>
<tr>
<td>inner</td>
<td>1mm</td>
<td>52.94</td>
<td>0.399</td>
<td>-13.180</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>51.96</td>
<td>0.396</td>
<td>-16.07</td>
</tr>
<tr>
<td></td>
<td>Simulated</td>
<td>52.77</td>
<td>0.396</td>
<td>-12.240</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>52.08</td>
<td>0.396</td>
<td>-13.660</td>
</tr>
<tr>
<td>middle</td>
<td>1mm</td>
<td>52.60</td>
<td>0.398</td>
<td>-15.365</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>51.66</td>
<td>0.408</td>
<td>-15.030</td>
</tr>
<tr>
<td></td>
<td>Simulated</td>
<td>52.26</td>
<td>0.400</td>
<td>-13.687</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>51.80</td>
<td>0.408</td>
<td>-14.860</td>
</tr>
<tr>
<td>outer</td>
<td>1mm</td>
<td>52.94</td>
<td>0.399</td>
<td>-13.092</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>52.01</td>
<td>0.382</td>
<td>-16.160</td>
</tr>
<tr>
<td></td>
<td>Simulated</td>
<td>52.81</td>
<td>0.399</td>
<td>-13.092</td>
</tr>
<tr>
<td></td>
<td>Measured</td>
<td>52.09</td>
<td>0.390</td>
<td>-15.750</td>
</tr>
</tbody>
</table>

Table 1: comparison of analysis and measurements for a 3.7m-diameter reflector antenna in different case

Table 1 shows the measured and calculated index results of electrical performance of the antenna in various conditions. The conclusion can be drew when the calculated results of electromechanical coupling model compare with the actual measured results, the maximum relative error of which is 1.89%, 4.45%, 19.55% and 15.70% in gain of antenna, main beam width, left and right first side lobe level. So the rightness and rationality of electromechanical two-field coupling theory model can be acquired. From the table 2, we can draw the conclusion that the error between the calculated and measured results get smaller with the increase of the gasket thickness, which is mainly due to the existence of initial random errors on the main reflector surface of the antenna.

5 Conclusion

Large reflector antennas are more vulnerable to the impact of the external environment, load, and therefore require high-precision calculations. The electromechanical two-field coupling model of reflector antennas is established in this paper using surface current method, based on which the solving strategies and methods are given also. Finally, the correctness and calculation accuracy of this model is verified using a 3.7m diameter antenna, which provides a certain referential significance to the electromechanical coupling analysis of reflector antennas and other electronic equipment.

REFERENCES

Meng Wang and Wei Wang

ROSENBROCK TIME INTEGRATION FOR UNSTEADY FLOW SIMULATIONS

D. S. BLOM*, H. BIJL*, P. BIRKEN†, A. MEISTER† AND A. H. VAN ZUIJLEN*

*Faculty of Aerospace Engineering
Delft University of Technology
P. O. Box 5058, 2600 GB Delft, The Netherlands

†Institut für Mathematik
Universität Kassel
Heinrich Plett Str. 40 (AVZ), D-34132 Kassel, Germany

Key words: Rosenbrock time integration, Higher order time integration, Unsteady flow simulations, ESDIRK, Rosenbrock-Wanner methods

Abstract. This contribution compares the efficiency of Rosenbrock time integration schemes with ESDIRK schemes, applicable to unsteady flow and fluid-structure interaction simulations. Compared to non-linear ESDIRK schemes, the linear implicit Rosenbrock-Wanner schemes require subsequent solution of the same linear systems with different right hand sides. By solving the linear systems with the iterative solver GMRES, the preconditioner can be reused for the subsequent stages of the Rosenbrock-Wanner scheme. Unsteady flow simulations show a gain in computational efficiency of approximately factor three to five in comparison with ESDIRK.

1 INTRODUCTION

Efficient time integration methods applicable to fluid dynamics and fluid-structure interaction simulations are of high importance. High order time integration methods are employed in order to increase the efficiency of unsteady computations. Currently, second order implicit schemes are commonly used in engineering codes [1]. The use of implicit methods is advised, since explicit methods impose strict stability constraints on the time step used by the method. Contrary to explicit methods, the time step for implicit methods can be chosen based on accuracy considerations. For fluid flows, large differences in length and time scales are present, namely in the boundary layer, which increase the stiffness of the system. Therefore, implicit schemes are preferred over explicit schemes for fluid solvers.
This contribution compares the computational efficiency and accuracy of implicit Runge-Kutta schemes, namely ESDIRK, with Rosenbrock-Wanner schemes for a non-linear convection-diffusion problem and a two-dimensional laminar flow problem [2]. Rosenbrock-Wanner methods follow from a linearisation of a DIRK scheme, thereby loosing some stability and accuracy properties of the implicit Runge-Kutta scheme, but the computational costs per time step are reduced. As a result of the linearisation step, the linear implicit Rosenbrock-Wanner scheme consists of solving the subsequent stages with constant system matrix and varying right hand sides. The preconditioner can be reused for all the stages of the scheme.

The paper is structured as follows. ESDIRK and Rosenbrock-Wanner time integration schemes are discussed in Section 2. The Newton-Krylov method is shortly discussed in Section 3. The results for a non-linear convection-diffusion problem and for a uniform flow around a cylinder are shown in Section 4 and 5, respectively. Section 6 finalises this paper with the conclusions.

2 TIME INTEGRATION SCHEMES

The method of lines paradigm is followed in this paper. Therefore, the non-linear convection-diffusion problem, and the Navier-Stokes equations are discretised in space and in time. The initial value problem of the form

$$\frac{d}{dt}w(t) = F(w(t)), \quad w(0) = w^0$$

is solved for a known initial solution $w^0$, where $F$ represents the spatial discretisation of the convection-diffusion problem or Navier-Stokes equations. ESDIRK and Rosenbrock-Wanner time integration schemes are considered for this study, and discussed in Section 2.1 and 2.2. The used adaptive time step control method is shortly laid out in Section 2.3.

2.1 ESDIRK schemes

Explicit first stage, single diagonal, diagonally implicit Runge-Kutta (ESDIRK) methods are a subclass of SDIRK methods, and as a consequence are L-stable for any design order. The ESDIRK method is given by

$$w^{(i)} = w^n + \Delta t \sum_{j=1}^{s} a_{ij} F^{(j)}, \quad F^{(j)} = F(w^{(j)}), \quad i = 1, \ldots, s$$

wherein $w^{(i)}$ are the stage values, and $\Delta t$ is the used time step. The solution at the next time level $w^{n+1}$ is computed with

$$w^{n+1} = w^n + \Delta t \sum_{j=1}^{s} b_j F^{(j)}.$$
A lower order solution $\hat{w}^{n+1}$ can be determined with

$$\hat{w}^{n+1} = w^n + \Delta t \sum_{j=1}^{s} \hat{b}_j F^{(j)},$$

which is used by the adaptive time step control algorithm to efficiently calculate an error estimate for the time step. The coefficients $a_{ij}$, $b_j$ and $\hat{b}_j$ can be found in a Butcher tableau. The coefficients of the used schemes are shown in [3, 4].

The first stage of an ESDIRK method is explicit, i.e. $a_{11} = 0$. Hence, $s-1$ non-linear systems need to be solved. Also, the solution at the last stage of the method is equal to the solution at the next time step, thus $a_{sj} = b_j$. The implicit stages can be solved with a multi grid method or a Newton-Krylov method.

### 2.2 Rosenbrock-Wanner schemes

Rosenbrock, Rosenbrock-Wanner or ROW-schemes are part of a class of linearly implicit Runge-Kutta methods. Rosenbrock methods replace non-linear systems with a sequence of linear systems, and are derived by linearizing a DIRK scheme. As a result, some stability and accuracy properties are lost, but the computational costs per time step are reduced: $s$ linear equation systems with a constant coefficient matrix and different right hand sides need to be solved, instead of $s$ non-linear systems.

An $s$-stage Rosenbrock method is described with the following relation:

$$(I - \Delta t \gamma_{ii} J) w^{(i)} = \Delta t F \left( w^n + \sum_{j=1}^{i-1} \alpha_{ij} w^{(j)} \right) + \Delta t J \sum_{j=1}^{i-1} \gamma_{ij} w^{(j)}, \quad i = 1, \ldots, s. \quad (5)$$

The solution at the next time step $w^{n+1}$ is determined with

$$w^{n+1} = w^n + \sum_{j=1}^{s} b_j w^{(j)}. \quad (6)$$

A lower order estimation $\hat{w}^{n+1}$ can be found by evaluating

$$\hat{w}^{n+1} = \hat{w}^n + \sum_{j=1}^{s} \hat{b}_j w^{(j)}. \quad (7)$$

The coefficients $\alpha_{ij}$, $\gamma_{ij}$ and $b_j$ are generally shown in a Butcher tableau.

In order to accelerate the computations, the Jacobians $J$ ($w^{(i)}$) are replaced by $J = J (w_n)$, such that the Jacobian needs to be evaluated only once during the Rosenbrock computation [5]. W-methods are obtained, if an approximation for the Jacobian is used. W-methods have additional order conditions. Krylov-ROW schemes are applied, if a Krylov subspace method is used to compute a solution for the linear system.
For implementation purposes, Equations (5) and (6) can be rewritten by introducing the new variables \( u^{(i)} \). This approach is applied, since a direct implementation of Equations (5) and (6) requires the solution of a linear system with the matrix \( I - \Delta t \gamma_{ii} J \) and the matrix-vector multiplication \( J \cdot \sum \gamma_{ij} w^{(j)} \). This matrix-vector multiplication is avoided by introducing the new variables \( u^{(i)} \):

\[
\begin{align*}
    u^{(i)} &= \sum_{j=1}^{i} \gamma_{ij} w^{(j)} & i = 1, ..., s. \\
\end{align*}
\]

If \( \gamma_{ij} \neq 0 \) for \( j \leq i \), then the matrix \( \Gamma = (\gamma_{ij}) \) is invertible and \( w^{(i)} \) can be determined from \( u^{(i)} \) with

\[
    w^{(i)} = \frac{1}{\gamma_{ii}} u^{(i)} - \sum_{j=1}^{i-1} c_{ij} u^{(j)},
\]

wherein \( C \) is given by \( C = \text{diag } (\gamma_{11}^{-1}, ..., \gamma_{ss}^{-1}) - \Gamma^{-1} \).

Thus the following formulation of the Rosenbrock method is found for practical implementations,

\[
    L u^{(i)} = F \left( w^n + \sum_{j=1}^{i-1} a_{ij} u^{(j)} \right) + \frac{1}{\Delta t} \sum_{j=1}^{i-1} c_{ij} u^{(j)}, \quad i = 1, ..., s,
\]

with \( L = \left( \frac{1}{\Delta t \gamma} - I - J \right) \), and \( \gamma = \gamma_{ii} \), thus \( L \) is constant for the consecutive stages of the Rosenbrock scheme. The solution at the next time step \( w^{n+1} \) is given by

\[
    w^{n+1} = w^n + \sum_{j=1}^{s} m_j u^{(j)},
\]

wherein the coefficients \( a_{ij} \) and \( m_j \) are given by \( (a_{ij}) = (\alpha_{ij}) \Gamma^{-1} \) and \( (m_1, ..., m_s) = (b_1, ..., b_s) \Gamma^{-1} \). The coefficients of the used ROW-schemes can be found in \([6, 7, 8, 9]\).

Concluding, Rosenbrock-type methods are presented as an alternative to ESDIRK time integration schemes. Rosenbrock-Wanner methods can be used, which use an approximation for the Jacobian, thus effectively reducing the computational costs per time step. However, the accuracy and stability are also reduced per time step. When Krylov-ROW schemes are applied, a Krylov subspace method is used to compute the solution for the linear system resulting from the Rosenbrock scheme.

2.3 Adaptive time step control

Time step control is an important measure to increase the efficiency and robustness of a time integration method. A constant time step often results in a large number of small steps, increasing the computational costs of a simulation significantly.
For Runge-Kutta and Rosenbrock schemes, an embedded scheme can be used as an error estimator:

\[ r^n = ||\hat{w}^{n+1} - w^n||. \]  

(12)

A digital filter is used for the selection of the time step, as discussed in [10]. The next time step is computed with

\[ \Delta t^{n+1} = (\epsilon \frac{\beta_1}{\epsilon})^\frac{1}{p} (\epsilon \frac{\beta_2}{\epsilon})^\frac{1}{p-1} (\Delta t^n)^{-\alpha_2} \Delta t^n, \]  

(13)

where \( \epsilon \) is determined with \( \epsilon = c \cdot TOL \), and \( p \cdot \beta_1 = p \cdot \beta_2 = \alpha_2 = \frac{1}{4} \) with \( p \) being the order of the embedded method of the ESDIRK or ROW-scheme. \( c \) is included as a safety margin, a typical value is 0.9.

The controller needs to be started with the classic controller:

\[ \Delta t^{n+1} = (\epsilon \frac{1}{p})^\frac{1}{p} \Delta t^n. \]  

(14)

Step size rejections may be reduced by basing the test on the requested change \( \rho^n \) instead on the error estimate. Also, discontinuities in the step size change ratio \( \Delta t^{n+1} / \Delta t^n \) are removed by applying a smooth limiter. Thus, the new step size is determined via \( \Delta t^{n+1} = \hat{\rho} \Delta t^n \), where \( \rho^n \) is given by

\[ \rho^n = (\epsilon \frac{\beta_1}{\epsilon})^\frac{1}{p} (\epsilon \frac{\beta_2}{\epsilon})^\frac{1}{p-1} (\rho^n)^{-\alpha_2}, \]  

(15)

and the smooth limiter gives \( \hat{\rho} \) with \( \kappa = 2 \):

\[ \hat{\rho} = 1 + \kappa \arctan \left( \frac{\rho^n - 1}{\kappa} \right). \]  

(16)

3 NON-LINEAR SYSTEMS OF EQUATIONS

The implicit Runge-Kutta schemes lead to a nonlinear system of equations of the form

\[ u = \hat{u} + \alpha \Delta t \hat{f}(u), \]  

(17)

where \( u \in \mathbb{R} \) is the unknown vector, \( \alpha \) is a parameter, and \( \hat{u} \) is a given vector. The function \( \hat{f}(u) \) performs the temporal and spatial discretisation of the computational domain. This problem is solved by the Newton-Raphson method, which effectively solves the root problem

\[ F(u) = 0. \]  

(18)

An inexact Newton method is employed, thus the following procedure is followed repeatedly until the convergence criteria are satisfied:

\[ \left| \left| \frac{\partial F(u)}{\partial u} \right| \right|_k \Delta u + F(u_k) \leq \eta_k ||F(u_k)|| \]

\[ u_{k+1} = u_k + \Delta u, \quad k = 0, 1, 2, 3, \ldots, \]  

(19)
The linear system of the Newton method is solved by a Krylov subspace solver such as GMRES. Eisenstat and Walker’s method [11] is used to determine the cut-off criterion \( \eta_k \) of the GMRES method with the parameters

\[
\eta_k^A = \gamma \frac{||F(u_k)||^2}{||F(u_{k-1})||^2},
\]

and

\[
\eta_k^B = \begin{cases} 
\eta_{max}, & k = 0, \\
\min (\eta_{max}, \eta_k^A), & k > 0, \gamma \eta_{k-1}^2 \leq 0.1, \\
\min (\eta_{max}, \max (\eta_k^A, \gamma \eta_{k-1}^2)), & k > 0, \gamma \eta_{k-1}^2 > 0.1 
\end{cases}
\]

wherein \( \gamma = 0.9 \). In order to avoid over solving of the final step of Newton method, the norm of the current nonlinear residual \( ||F(u_k)|| \) is compared to the nonlinear residual norm at which the iterations would stop

\[
\tau_t = \tau_a + \tau_r ||F(u_k)||,
\]

for an absolute \( \tau_a \) and a relative \( \tau_r \) convergence criterion. \( \eta_k \) is bounded from below by a constant multiple of \( \frac{\tau_t}{||F(u_k)||} \). The cut-off criterion \( \eta_k \) is determined with

\[
\eta_n = \min \left( \eta_{max}, \max \left( \eta_k^B, \frac{0.5 \tau_t}{||F(u_k)||} \right) \right).
\]

The matrix-vector products required by GMRES are estimated via a second order finite difference scheme

\[
\frac{\partial F(u)}{\partial u} v \approx \frac{F(u + \epsilon v) - F(u - \epsilon v)}{2 \epsilon},
\]

where the finite difference step \( \epsilon \) is determined with

\[
\epsilon = \sqrt{1 + ||u||} \sqrt{\frac{\epsilon_{mach}}{2}},
\]

with \( \epsilon_{mach} \) being the machine precision [12]. A first order approximation can also be used in order to decrease computational costs. However, preliminary computations showed that a second order approximation was necessary in order to increase the robustness of the method when applied to the Rosenbrock schemes.

An ILU preconditioner is used in order to decrease the condition number of the system matrix, and thus accelerating the convergence of GMRES.

4 RESULTS FOR A NON-LINEAR CONVECTION-DIFFUSION PROBLEM

This first test case consists of a generalised non-linear convection-diffusion equation:

\[
u_t = \beta u^n \cdot \nabla u + \alpha \nabla \cdot (u'' \nabla u), \quad x \in \Omega := (0, 1) \times (0, 1),
\]
where \( u(x, y, 0) \) is given by
\[
    u(x, y, 0) = u_0(x, y) \quad (27)
\]
The strength of the diffusion is determined by the parameter \( \alpha \in \mathbb{R} \). The strength and the direction of the convection is determined with
\[
    \beta = \tilde{\beta} \left( \begin{array}{c}
    \sin(\gamma) \\
    \cos(\gamma)
    \end{array} \right), \quad (28)
\]
where \( \tilde{\beta} \in \mathbb{R} \) is a user specified parameter, and \( \gamma \) determines the angle of the direction of the convection. The degree of non linearity is determined with the parameters \( m \) and \( n \).

The initial solution used for this test case is shown in Figure 1(a). The initial solution is one in the complete domain, except on the square \([0.1, 0.3] \times [0.1, 0.3] \), where the initial value is 1.1. The values for \( \alpha, \beta \), and \( \gamma \) are set to \( \alpha = 1, \beta = 200 \) and \( \gamma = 0.35 \pi \). The reference solution at the end of the simulation is shown graphically in Figure 1(b) for \( n = m = 1 \).

A non-uniform mesh is used for the computations shown in this section \((50 \times 50)\). As shown in Figure 1, the mesh is refined close to \( x = 0.5 \) and \( y = 0.5 \) resulting in cells with a high aspect ratio. The maximum aspect ratio of the non-uniform mesh is 9.8. The condition number of \( L \) is relatively low for a uniform mesh. Hence, preconditioning is not necessary. Therefore, there is no obvious advantage of a constant system matrix \( L \) for a uniform mesh.

![Figure 1](image-url)

**Figure 1:** Initial solution and reference solution for a non-linear convection-diffusion simulation. The reference solution has been obtained with ESDIRK5 and \( \Delta t = 10^{-7} \).

### 4.1 Effect of the Newton-Krylov method on accuracy and efficiency

The effect of the use of the Newton-Krylov method for the ESDIRK and Rosenbrock schemes is investigated. The results of simulations are shown in Figure 2. The ROW-schemes show a gain in efficiency compared to ESDIRK for the range of time step sizes. A small difference in accuracy is observed between the ESDIRK and ROW-schemes, where
the ESDIRK schemes show a slightly higher accuracy. RODASP has the greatest potential for use in a flow solver, since the computational time is reduced by approximately factor 2.5 compared to ESDIRK4.

![Figure 2](image_url)  
(a) Accuracy  
(b) Computational work

**Figure 2:** Fixed time step study performed with the Newton-Krylov method for the non-linear convection-diffusion problem

### 4.2 Effect of adaptive time step control on accuracy and efficiency

The effect of the use of the adaptive time step selection on accuracy and efficiency for the non-linear convection-diffusion case is shown in Figure 3. A difference in accuracy is observed between the different schemes for equal adaptive tolerance settings. Therefore it is necessary to perform a tolerance calibration in order to get the same accuracies for the different methods for a given tolerance [10].

Comparing the computational efficiency of ESDIRK and Rosenbrock shows a gain in efficiency for the Rosenbrock schemes. RODASP is the most computationally efficient scheme in comparison with the other time integration schemes for which the computational time reduced by a factor 3 to 4 compared to ESDIRK4.

### 5 RESULTS FOR A UNIFORM FLOW PAST A CIRCULAR CYLINDER

Following the previous non-linear convection-diffusion test case, the question remains how the ROW-scheme compares to the ESDIRK scheme in terms of computational efficiency and stability when applied to viscous flows. The second test case consists of a two-dimensional flow around a cylinder. The circular cylinder is held fixed in a uniform flow field, resulting in a vortex-street behind the cylinder. When the initial transient has disappeared, an unsteady periodic flow is present. This test case has been used in [13]
5.1 Description of the flow past a circular cylinder

The cylinder with diameter \( D \) is located on a fixed position in a uniform flow field with Mach number \( M_\infty = 0.3 \) and Reynolds number \( Re_\infty = 1.0 \cdot 10^3 \), simulating a laminar flow. The radius of the cylinder is used as the characteristic length to determine the Reynolds number.

The computational domain consists of \( 2.5D \) upstream of the centre of the cylinder, \( 4.5D \) above and below the cylinder centre, and \( 16.5D \) downstream of the centre of the cylinder. The mesh is refined in twelve steps to obtain a highly refined region close to the cylinder and in the wake downstream, resulting in a mesh with 10,608 cells. Close to the cylinder five extra layers of body conformal cells are generated resulting in an accurate representation of the boundary layer. The smallest cells which are located in the boundary layer, are of size \( 6.6 \cdot 10^{-5} D \). The maximum aspect ratio of the cells in the mesh is \( 6.3D \), and the minimum aspect ratio is \( 1.0D \). Refinement in the wake is performed, since the vortex street needs to be resolved accurately to obtain a good accuracy for the simulations. The generated mesh is shown in Figure 4.

5.2 Effect of the Newton-Krylov method on accuracy and efficiency

A fixed time step study has been performed for the ESDIRK and Rosenbrock time integration schemes. A non-linear multi grid solver [14] and Jacobian-free Newton-Krylov solver are used to solve the implicit stages of the ESDIRK scheme, and GMRES for the stages of the Rosenbrock scheme.

Figure 3: Comparison of efficiency and accuracy with adaptive time step control applied and [1] to study the order of the ESDIRK schemes in comparison with BDF2.
The accuracy of the other computations performed with a Rosenbrock scheme lie close to the computations performed with an ESDIRK scheme. The order of the fourth order schemes reduces for small time steps, which is caused by the iterative error originating from the use of the Newton-Krylov and GMRES solvers.

A gain in computational efficiency is observed for the ROW-schemes in comparison with ESDIRK. The JFNK solver shows a significant increase in computational efficiency compared to the multi grid solver. The performance of RODASP is close to ESDIRK4.

5.3 Effect of adaptive time step control on accuracy and computational stability for a uniform flow around a cylinder

Figure 6 shows the results of the numerical study comparing the accuracy and efficiency of ESDIRK and Rosenbrock when an adaptive time step control algorithm is employed. A large difference in accuracy for the same tolerance setting is observed when Rosenbrock and ESDIRK are compared. For the ROW-schemes, the accuracy of the solution is
more than one magnitude higher compared to ESDIRK3 and ESDIRK4. This difference in accuracy is comparable to the convection-diffusion case, indicating the possibility to perform a tolerance calibration and reuse the found coefficients for different problems. Also, a significant gain in efficiency is observed for the Rosenbrock schemes in comparison with ESDIRK.

![Graphs showing accuracy and computational work comparison](image)

(a) Accuracy (b) Computational work

Figure 6: Computational work and accuracy with adaptive time step selection

6 CONCLUSIONS

The computational efficiency and accuracy of ROW-schemes are compared with ESDIRK schemes for a non-linear convection-diffusion problem, and a laminar 2D flow around a cylinder. The numerical studies focused on the effects of using a fixed time or adaptive time step control algorithm on efficiency and accuracy of the simulations.

The main observation is that the ROW-schemes ROS34PRW and RODASP outperform the ESDIRK schemes in terms of computational efficiency. A difference in accuracy is observed when an adaptive time steps control method is used indicating that a tolerance calibration is necessary.

REFERENCES


SIMULTANEOUS FLOW OF WATER AND AIR ACROSS THE LAND SURFACE DURING RUNOFF

JENS-OLAF DELFS∗†, ASHOK K. SINGH†, THOMAS KALBACHER†, WENQING WANG† AND OLAF KOLDITZ‡

∗Water-Earth System Science (WESS) University of Tübingen Keplerst. 17, 72074 Tübingen, Germany
E-mail: jens-olaf.delfs@ufz.de, web page: http://www.wess.info

†Department Environmental Informatics
Helmholtz Centre for Environmental Research UFZ Leipzig Permoserstr. 17, 04318 Leipzig, Germany
Web page: http://www.ufz.de

‡Environmental System Analysis
Technical University of Dresden Helmholtzstr. 10, 01062 Dresden, Germany
Web page: http://www.tu-dresden.de

Key words: Two-phase flow, Horton runoff, Leakance, Soil gas release, Sequential iterative coupling, OpenGeoSys

Abstract. This paper presents an inter-compartment boundary condition for the simulation of surface runoff, soil moisture, and soil air as a coupled system of partial differential equations. The boundary condition is based on a classic leakance approach to balance water between differently mobile regions such as the land surface and subsurface. Present work applies leakances to transfer water and air simultaneously through the land surface for soils, which are connected by an air flux with a steady atmosphere. Shallow flow and two phase flow in a porous medium are sequential calculated in an iteration loop. General criteria are stated to guarantee numerical stability in the coupling loop and for leakances to control inter-compartment fluid fluxes. Using the leakance approach, a numerical model captures typical feedbacks between surface runoff and soil air in near-stream areas. Specifically, displacement of water and air in soils is hampered at full-water saturation over the land surface resulting in enhanced surface runoff in the test cases. Leakance parameters permit the simulation of air out-breaks with reference to air pressures, which fluctuate in the shallow subsurface between two thresholds.
1 BACKGROUND

Fluid mass flows between partial differential equations of surface and subsurface flow via inter-compartment boundary conditions at the land surface or riverbeds. Highly detailed models\cite{6} combine free-flow of the Stokes or Navier Stokes equations with flow in a porous medium through a transition zone, where a slip condition by Beavers and Joseph \cite{1} balances momentum.

Modeling runoff of water over the land surface usually involves application of some form of hydrostatic shallow water equation, namely the Saint-Venant, diffusive or kinematic wave equation. Hydrostatic shallow flow has been coupled to flow in the subsurface in form of Darcy and Richards’ flow\cite{7,8} and recently also to two-phase flow \cite{2}. Some coupled runoff models enforce continuity between the hydrostatic pressure of surface water and the matric pressure of variably saturated soils at the land surface\cite{7}. Others assume the existence of a small interface (transition zone) and control the mutual mass exchange between flow compartments with an additional leakance parameter\cite{2,4,8}.

2 COUPLING FLOW COMPARTMENTS

Flow of a liquid in the overland compartment is represented by the diffusive wave equation in our numerical model. The shallow flow equation assumes hydrostatic pressure $p^H$ in a depth-integrated 1D or 2D flow field. 1D, 2D or 3D flow of a liquid (superscript $l$) and gas (superscript $g$) in the soil compartment is simultaneously calculated with a two-phase flow equation. In contrast, diffusive wave overland flow and two-phase flow are sequentially calculated in an iteration loop. Thus, the numerical model iterates between the following algebraic equation systems\cite{2}

$$A^H h = b^H, \quad A^p \begin{pmatrix} p^c \\ p^g \end{pmatrix} = \begin{pmatrix} b^l \\ b^g \end{pmatrix},$$

where $A$ are system matrices for overland flow (superscript $H$) and two-phase flow. Right hand side vectors $b$ account for source / sink terms, which originate from inter-compartment fluxes (Sect. 2.1), precipitation, outlets, etc.. Primary variables are the hydraulic head $h$ in the surface water (Fig. 1(a)), the capillary pressure $p^c$, and the soil gas pressure $p^g$.

2.1 Fluid fluxes through the land surface

A new inter-compartment boundary condition provides liquid and gas fluxes through a transition zone at the surface of a porous medium, which is a soil in this work (Fig. 1(a)). The transition zone is constructed with a continuum approach and consists of:

1. **A porosity in overland flow** for homogenization of surface structure (rills, etc.). Overland flow starts, if liquid depth exceeds certain thickness of surface structure ($H > a$ in Fig. 1(a))\cite{8}.
2. **An interface layer** $a^l$ for a mutual coupling flux between the liquid in the overland compartment and the liquid phase of the porous medium in the soil compartment[8]

$$ q^l_c = -\lambda^l (p^H + p^c - p^g_e), \quad \lambda^l = k^l_c k^l \mu^l a^l, $$

where $\lambda^l$ is the leakance for liquid, $\mu^l$ the dynamic liquid viscosity, $k^l_c$ the relative liquid interface permeability (Sect. 2.2), $k$ the intrinsic soil permeability, $p^g_e = p^g - p^g_{atm}$ the atmospheric excess gas pressure, and $p^g_{atm}$ the atmospheric pressure.

3. **An interface layer** $a^g$ to connect gas of the soil and atmosphere compartments[2]

$$ q^g_c = \lambda^g p^g_e, \quad \lambda^g = \frac{k^g_c k^g}{\mu^g a^g}, $$

where $\lambda^g$ is the leakance for gas, $\mu^g$ the dynamic gas viscosity, and $k^g_c$ the relative gas interface permeability (Sect. 2.2).

The fluxes $q^l_c$ and $q^g_c$ are implemented as implicit source terms in the algebraic equation systems (1) for iterative coupling. Concerning stability, it is important (Sect. 4.3) that the hydrostatic surface water pressure $p^H$ is often negligible. The atmospheric pressure $p^g_{atm}$ can be set as zero in many practical cases.
2.2 Functions of leakance

The leakance approach enables the modeler to control inter-compartment fluid fluxes (2) and (3) with reference to state variables (pressures). To achieve this, leakances $\lambda_l$, $\lambda_g$ are modified by relative interface permeabilities $k^l_r$, $k^g_r$, respectively. It holds $0 \leq k^l_r(p^H, p^g) \leq 1 - k^g_r$, where $k^g_r$ is the residual value for the relative gas interface permeability $k^g_r \approx 1 - k^l_c$. This accomplishes:

1. To limit the liquid exchange flux (2) by the available liquid with depth $H$ in the overland compartment when liquid infiltrates into dry soil with high capillary forces. $k^l_c$ is zero for if the liquid depth $H$ exceeds certain threshold $a^s$.

2. To disconnect soil gas from the atmosphere compartment during liquid ponding and runoff in the overland compartment. The relative gas interface permeability $k^g_c$ becomes the residual value $k^g_{cr}$ for $H \geq a^s$.[2]

3. To act as a valve for compressed soil gas by following a hysteresis curve (Fig. 1(b)). Gas pressure $p^g$ in the porous medium fluctuates between two thresholds $p_{break}$, $p_{close}$ where gas breaks out of a liquid-covered surface and the surface closes for gas, respectively.[9] $k^l_c$ is reduced by the minimum flow control factor $f_{break}$ during air pressure counterflow through the land surface.[2]

3 TEST CASES

3.1 Classic Smith and Woolhiser (1971) benchmark on Horton runoff

A light oil was applied[7] for 15 minutes with a rate of 4.2 cm per minute on a sand flume with a length of 12.2 meter and a slope of 1% (Fig.2(a)). Liquid infiltration was tracked inside the flume and Horton runoff recorded at an outlet. Sand properties varied only slightly. Thus, flow is simulated in 1D with vertical and homogeneous soil columns.[2]
Soil air can escape into the atmosphere at the top of columns through the gas exchange flux (3). Air pressure is enforced to equal atmospheric pressure $p_{atm}$ at 30 cm soil depth.

### 3.2 Synthetic floodplain

An inclined plain has a length of 10 meter and a slope of 2% (Fig. 2(b)). The plane is initially ponded at 1 meter length. At this point, a source term of $q = 3$ meter per second is assigned for 10 seconds. The water depth quickly rises and a flood wave flushes around 85% of the plane length. Inter-compartment fluxes (2) and (3) permit water and air exchange through the land surface. The remaining boundaries are assumed as impervious for water and air. Soil parameters are chosen as in the Smith and Woolhiser (1971) benchmark to represent a homogeneous sand.

### 4 RESULTS AND DISCUSSION

#### 4.1 Model verification

The coupled overland / two-phase flow model is verified with liquid data of a Horton flow experiment (Sect. 3.1). Soil air flow was not measured, although air flow effects were noticed by the experimentators[7]. Results presented in Fig. 3(a) reveal that soil gas pressures impede the calculated infiltration of water and, consequently, amplify Horton runoff. A distinctive air out-break event (Sect. 2.2) was simulated for the early part of the hydrograph. The increase in surface runoff corresponds well with the experimental data regarding the amount. However, by using 1D soil columns (Fig. 2(b)), the coupled model was not able to capture the late rise in the experimental runoff. Obviously, a broader experimental data basis is needed to test the coupled model throughout.

#### 4.2 Interface thicknesses

Interfaces (transition zones) are introduced in a leakance concept (Sect. 2.1). The extra parameters require careful examination[5]. A parameter sensitivity study revealed[2] that interfaces impede liquid and gas exchange between compartments if their thicknesses exceed certain thresholds

$$a_l > a_{c_l}, \quad a_g > a_{c_g}.$$  \hspace{1cm} (4)

Thresholds are $a_{c_l} = 0.1$ mm for the liquid (Fig. 3(b)) and $a_{c_g} = 1$ mm for the gas in the test case on Horton flow (Sect. 3.1), and independent of each other. The threshold $a_{c_g}$ for the gas interface thickness linearly increases with the length of the soil column. Leakances at thresholds $a_{c_l}$ and $a_{c_g}$ are independent of other parameters than intrinsic permeability $k$ and the geometry (e.g. soil column length) of the subsurface porous medium system.

#### 4.3 Stability in the coupling loop

The inter-compartment fluxes (2) and (3) are implicitly calculated in a sequential iterative coupling algorithm. One part of a flux $q_{c_l}$, $q_{c_g}$ is implemented in the matrix
and the other part in the right hand side vector of the algebraic equation system (1) if atmospheric pressure $p_{\text{atm}}^g$ and hydrostatic pressure in surface flow $p^H$ have finite values. As a consequence, the capillary pressure $p^c$ and gas pressure $p^g$ exhibit discontinuities at the compartment interface. Fig. 4(a) shows discontinuities in atmospheric excess gas pressure $p_{\text{e}}^g$. They have a constant step size of

$$\Delta p_{\text{e}}^g = 10^{-12} p_{\text{atm}}^g.$$  

Thus, the gas flux (3) has discontinuities with a step size of $\Delta q_{\text{c}}^g = \lambda^g \Delta p_{\text{e}}^g = 10^{-12} \lambda^g p_{\text{atm}}^g$. In the use of inter-compartment coupling fluxes $q_{\text{c}}^g$ and $q_{\text{c}}^l$ with surface pressure terms $p^H$, $p_{\text{atm}}^g$, it is important to ensure stability by selecting low leakances $\lambda^l$ and $\lambda^g$ without impeding fluid exchange. The criteria (4) and (5) guide the model applier to proper ranges for leakances $\lambda^l$ and $\lambda^g$ in practice, e.g. $a^l > 10^{-6} \text{ m}$ and $a^g > 10^{-10} \text{ m}$ for stability in the test case on Horton flow (Sect. 3.1).

### 4.4 Soil air entrapment

Flooded soils exhibit a reduced infiltration capacity, if a near-surface water table prevents compressed soil air from disappearing into deeper subsurface regions. Thus, Fig. 4(b) illustrates the capabiliy of the coupled model to produce soil air entrapment in a 2D floodplain test case (Sect. 3.2). Infiltration of flood water is strongly retarded and occurs mainly adjacent to the water-free part of the pervious plain ($x \approx 17 \text{ m}$). As a consequence, the infiltrating water isolates an air cell of high soil air pressures.
Figure 4: (a) Discontinuities in iterative coupling ($p_{atm}^g = 101325$ Pa in the Horton flow benchmark, source: Delfs et al. (2013)[2]); (b) Snapshots of a synthetic floodplain for 2 h simulation time (vertical axis exaggerated).

5 SUMMARY AND CONCLUSIONS

In the numerical study, leakances controlled the mutual fluid mass transfer through an inter-compartment transition zone (Fig. 1) in a unique hydrostatic shallow / two-phase flow model. The numerical stability is ensured when the boundary condition is used in combination with a sequential iterative coupling algorithm (Fig. 4(a)). Soil gas pressures impeded surface runoff from infiltrating in a liquid-covered soil flume (Fig. 3(a)) and a synthetic floodplain (Fig. 4(b)). Thus, our results suggest that air entrapment below pervious land surfaces amplifies stream flow considerably during high precipitation besides other principal stream flow generation mechanisms by Horton, Dunne, etc.. Insitu measurements of surface runoff and soil air pressures are needed for further testing of the air counterflow algorithm (Fig. 1(b)).

Various chemical and physical clogging mechanisms at land surfaces can be simulated by modifying leakances with state variables in future work (e.g. biofilms, unsaturated zones below riverbeds). The sequential coupling algorithm enables developers to couple the presented model with a dynamic atmosphere model to include soil air / atmosphere interactions[3]. The implementation is available via the open source scientific software project OpenGeoSys (www.opengeosys.net) and can be combined with various model
components, e.g. for non-isothermal flow, deformation processes, and transport of multiple reactive components.

ACKNOWLEDGMENTS

This work was partly funded by the German Ministry of Education and Research (BMBF) project IWAS International Water Research Alliance Saxony (Project No. 02WM1027). WESS is supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg (AZ Zu 33-721.3-2) and the Helmholtz Centre for Environmental Research, Leipzig (UFZ).

REFERENCES


Towards interaction of elastic structures with turbulence induced acoustics at low Mach numbers

Frank Flitz*† and Michael Schäfer*

*Institute of Numerical Methods in Mechanical Engineering
Technische Universität Darmstadt
Dolivostraße 15, 64293 Darmstadt, Germany
e-mail: \{flitz,schaefer\}@fnb.tu-darmstadt.de, http://www.fnb.tu-darmstadt.de

†Graduate School of Computational Engineering
Technische Universität Darmstadt
Dolivostraße 15, 64293 Darmstadt, Germany
e-mail: flitz@gsc.tu-darmstadt.de, http://www.graduate-school-ce.de

Key words: Acoustics, Fluid-structure interaction (FSI), Splitting approach, Fluid-structure-acoustics interaction

Abstract. We present an approach for fluid-structure-acoustics interaction that aims at the low Mach number range and combines a splitting approach for the fluid, similar to the linearized perturbed compressible equations of Seo and Moon, with a strongly coupled fluid-structure interaction (FSI) concept (split FSI). We compare this approach to an acoustic FSI via a numerical simulation of a vibrational acoustics test case.

1 INTRODUCTION

1.1 Motivation

Combining aeroacoustics and FSI provides a field of interesting applications. One such application, for instance, is the impact of turbulence induced acoustic waves on an elastic structure. For this, an interaction of fluid, structure and acoustics has to be considered.

1.2 State of the art

In the area of fluid-structure-acoustics interaction, which includes the simultaneous treatment of flow induced and vibration induces acoustics, the approaches have different underlying models. The approach with the least modeling uses the compressible Navier-Stokes equations for the fluid. Although there are applications of this approach [1],[2], it is not generally applicable because of its inherent multiscale character [3]. Therefore,
another approach is to incorporate the acoustic analogy of Lighthill [4],[5] together with a one-way coupling from the structure to the acoustics [6],[7]. The approach of Zheng et al. [8] applies a splitting approach for the fluid together with a one-way coupling from the structure to the acoustics as well. No approach with a joint feedback of the incompressible and the acoustic part to the structure is found.

1.3 Procedure

We present an approach for a split FSI (a interaction between a structure and a fluid following a splitting approach) and show its feasibility to reproduce vibrational acoustics via a comparison to an acoustic FSI. In this approach of a split FSI the ability to model aeroacoustics is not harmed.

1.4 Outline

In section 2 the used methods are presented, followed by section 3, where a test case is introduced that illustrates the capabilities of the split FSI. In section 4 the conclusions are drawn.

2 METHODS

The equations for the split fluid as well as for the acoustic fluid are solved in the in-house finite volume solver FASTEST. The structural equations are solved with the finite element solver FEAP. The coupling between the fluid and the structure is implemented in an implicit, strongly coupled way and boundary conformal grids are used.

2.1 Split FSI

The split FSI is realized via a consistent combination of the splitting approach with the FSI.

2.1.1 Splitting approach

The splitting approach is the basis for our fluid model. It originates from Hardin and Pope [9] as acoustic/viscous splitting and was improved by Shen and Sørensen [10],[11], Ewert and Schröder [12], Seo and Moon [13] and Munz et al. [14] proposed further formulations. Here, we use the formulation of Kornhaas [15], which is mainly based on Seo and Moon’s linearized perturbed compressible equations [13]. Its starting point is the following splitting.

\[ \rho = \rho^f + \rho^a \]  \hspace{1cm} (1)

\[ \mathbf{v} = \mathbf{v}^f + \mathbf{v}^a \]  \hspace{1cm} (2)

\[ p = p^f + p^a \]  \hspace{1cm} (3)
Here \( \rho, v \) and \( p \) denote density, velocity and pressure, respectively. The indices \((\cdot)^f\) and \((\cdot)^a\) mark the incompressible and the acoustic (or perturbation) variables. The incompressible variables are determined by the incompressible Navier-Stokes equations

\[
\nabla \cdot v^f = 0 , \tag{4} 
\]

\[
\rho^f \frac{\partial v^f}{\partial t} \bigg|_\chi + \rho^f (v^f - v^s) \cdot \nabla v^f + \mu \nabla^2 v^f - \nabla p^f = 0 , \tag{5} 
\]

where \( t \) denotes the time and \( \mu \) the dynamic viscosity. We consider the arbitrary Lagrangian Eulerian (ALE) framework (for Details see for instance [16]) for the fluid, with \( \chi \) being the referential coordinate, together with the grid velocity

\[
v^s = \frac{\partial x}{\partial t} \bigg|_\chi , \tag{6} 
\]

with \( x \) denoting the spatial coordinate.

Inserting \( \rho, v \) and \( p \) into the compressible Navier-Stokes equations yields after simplifications [15]:

\[
\frac{\partial \rho^a}{\partial t} \bigg|_\chi + (v^f - v^s) \cdot \nabla \rho^a + \rho^f \nabla \cdot v^a = 0 , \tag{7} 
\]

\[
\rho^f \frac{\partial v^a}{\partial t} \bigg|_\chi + \rho^f (v^f - v^s) \cdot \nabla v^a + \nabla p^a = 0 , \tag{8} 
\]

\[
\frac{\partial p^a}{\partial t} \bigg|_\chi + \rho^a c^2 \nabla \cdot v^a + (v^f - v^s) \cdot \nabla p^a = - \frac{\partial p^f}{\partial t} \bigg|_\chi . \tag{9} 
\]

We utilize this quite rough variant, since we are not interested in the model of the splitting approach itself, but in the consistent application to the FSI.

### 2.1.2 FSI coupling

As governing equations in the structure we use

\[
\rho^s \frac{d^2 u}{dt^2} - (\lambda + G) \nabla (\nabla \cdot u) - G \nabla^2 u = 0 \tag{10}
\]

where \( \rho^s \) is the density, \( u \) the displacement, \( \lambda \) Lamé’s first parameter and \( G \) Lamé’s second parameter, or shear modulus.

The core of the splitting approach (1)–(3) has implications for the boundary conditions and the FSI. On the fluid-structure interface the fluid velocity has to be equal to the structural velocity. Since we consider low Mach number flow, this has to be modeled via the incompressible flow velocity \( v^f \).

\[
v^f - v^s = 0 \tag{11}
\]
At a wall and at the fluid-structure interface only the acoustic density and the acoustic pressure have to follow the zero gradient condition, while the acoustic velocity has to be set to zero, since the incompressible velocity is already equal to the wall or structural velocity, respectively.

\[
\nabla p^a \cdot n = 0 \quad (12)
\]

\[
v^a \cdot n = 0 \quad (13)
\]

\[
\nabla \rho^a \cdot n = 0 \quad (14)
\]

Note, that (13) is one key point in the consistent application of the splitting approach to FSI.

At the fluid-structure interface additionally to the incompressible flow stress the acoustic pressure has to be considered for the momentum balance.

\[
\sigma^s \cdot n = \mu \left( \nabla v^f + \left( \nabla v^f \right)^T \right) \cdot n - p^f I \cdot n - p^a I \cdot n \quad (15)
\]

Here, \( \sigma^s \) is the Cauchy stress in the structure. Note, that via (15) the feedback of the acoustics to the structure is implemented, which is another key point in the consistent application of the splitting approach to FSI.

### 2.2 Acoustic FSI

In the applied test case the classical acoustics act as the reference for the splitting approach. To distinguish its variables from the acoustic variables introduced in (7)–(9), we denote the classical acoustics ones with \((\cdot)^r\). In an ALE framework and without a background flow, the governing equations read as

\[
\begin{align*}
\left. \frac{\partial v^r}{\partial t} \right|_x - v^s \cdot \nabla v^r + \frac{1}{\rho^f} \nabla p^r &= 0, \quad (16) \\
\left. \frac{\partial p^r}{\partial t} \right|_x + \rho^f c^2 \nabla \cdot v^r - v^s \cdot \nabla p^r &= 0. \quad (17)
\end{align*}
\]

For the classical acoustics the boundary conditions at a wall and at the fluid-structure interface are set to

\[
\nabla p^r \cdot n = 0, \quad (18)
\]

\[
v^r \cdot n = v^s \cdot n, \quad (19)
\]

and the feedback to the structure is described via

\[
\sigma^s \cdot n = -p^r I \cdot n. \quad (20)
\]

### 3 Numerical Investigations

In this section we illustrate, that the splitting approach, consistently applied to FSI, is capable of simulation of vibrational acoustics.
3.1 Setup

We adopt the well known FSI benchmark setup from Turek and Hron [17] (figure 1, table 1). We regard the lower and the upper boundary as a wall, the left boundary as a flow wall and acoustic outlet and the right boundary as an outlet. Note, that there is no inflow. We use the material parameters in table 2 for the fluid and the gray indicated flag, while the circle is treated as a rigid body. Initially all variables are set to zero.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value/m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length</td>
<td>$L$ 2.5</td>
</tr>
<tr>
<td>Channel width</td>
<td>$H$ 0.41</td>
</tr>
<tr>
<td>Cylinder radius</td>
<td>$r$ 0.05</td>
</tr>
<tr>
<td>Flag thickness</td>
<td>$h$ 0.02</td>
</tr>
<tr>
<td>Flag tip (at $t = 0$)</td>
<td>$A_x$ 0.6</td>
</tr>
<tr>
<td></td>
<td>$A_y$ 0.2</td>
</tr>
<tr>
<td>Circle center</td>
<td>$C_x$ 0.2</td>
</tr>
<tr>
<td></td>
<td>$C_y$ 0.2</td>
</tr>
<tr>
<td>Reference point</td>
<td>$D_x$ 0.41</td>
</tr>
<tr>
<td></td>
<td>$D_y$ 0</td>
</tr>
<tr>
<td>Reference point</td>
<td>$E_x$ 0.6</td>
</tr>
<tr>
<td></td>
<td>$E_y$ 0</td>
</tr>
</tbody>
</table>
Table 2: Material parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young’s modulus $E$</td>
<td>0.14 MPa</td>
</tr>
<tr>
<td>Poisson’s ratio $\nu$</td>
<td>0.4</td>
</tr>
<tr>
<td>Structural density $\rho^s$</td>
<td>$1 \times 10^2$ kg m$^{-3}$</td>
</tr>
<tr>
<td>Fluid density $\rho^f$</td>
<td>$1 \times 10^3$ kg m$^{-3}$</td>
</tr>
<tr>
<td>Dynamic viscosity $\mu$</td>
<td>$1 \times 10^{-3}$ kg m$^{-1}$ s$^{-1}$</td>
</tr>
<tr>
<td>Speed of sound $c$</td>
<td>$5 \times 10^{-2}$ m s$^{-1}$</td>
</tr>
</tbody>
</table>

To generate vibrational acoustics we prescribe a boundary movement at the lower boundary between $D$ and $E$ via

$$y = \begin{cases} 
  y_a \sin \left(\pi \frac{x - D}{E_x - D_x}\right)^2 \left(\frac{t}{T} - \frac{1}{2\pi} \sin \left(2\pi \frac{t}{T}\right)\right) & \text{if } t < T \\
  y_a \sin \left(\pi \frac{x - D}{E_x - D_x}\right)^2 & \text{if } t \geq T 
\end{cases}$$

(21)

with $y_a = 0.001$ m, $T = 2$ s. With a maximal boundary velocity of $\frac{2y_a}{T}$ at $t = \frac{T}{2}$ this results in a Mach number of $Ma = 0.02$ and, with the diameter of the circle, a Reynolds number of $Re = 100$.

3.2 Results

The object of interest is the displacement of the tip of the flag (figure 2). The minor differences between the split FSI and the acoustic FSI are presumably due to the very simplified version of the splitting approach (7)–(9) used here. Note that if we had, in contrast to (15), neglected the feedback from the acoustics part to the structure, we would have gained a very different displacement, due to the immediate response of the flag, caused by the incompressible part.

The contour plots of the split pressure $p = p^f + p^a$ and the acoustic pressure $p^f$ in figure 3 underline the generally good agreement. This illustrates that the consistent application of the splitting approach to fluid-structure interaction is capable of simulating vibrational acoustics.

4 CONCLUSIONS

We presented the consistent application of the splitting approach to fluid-structure interaction, enabling the simulation of vibrational acoustics, without restrictions on the aeroacoustics. This makes the split FSI a decent approach for fluid-structure-acoustics interaction.
Frank Flitz and Michael Schäfer

Figure 2: Displacement of the flag tip $\Delta A_y(t) = A_y(t) - A_y(0)$. Until $t = 4$ s additionally to the split FSI and the acoustic FSI the results for the split FSI with absence of feedback from the acoustics to the structure are displayed.

5 ACKNOWLEDGEMENTS

This work was supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

REFERENCES


Figure 3: Contour plots of the split fluid pressure $p = p_f + p_a$ and the acoustic pressure $p_r$


RECENT ADVANCES IN THE PARTICLE FINITE ELEMENT METHOD. TOWARDS MORE COMPLEX FLUID FLOW APPLICATIONS.

NORBERTO M. NIGRO A, PABLO NOVARA B, JUAN M. GIMENEZ A, MARTA B. BERGALLO C, NESTOR A. CALVO B, PEDRO MORIN C AND SERGIO R. IDELSOHN D

A Centro de Investigación en Métodos Computacionales (CIMEC)
Consejo Nacional de Investigaciones Científicas y Técnicas y la Universidad Nacional del Litoral
Parque Tecnológico Litoral Centro, Colectora Ruta Nacional 168, 3000 Santa Fe, Argentina
e-mail: nnigro@intec.unl.edu.ar, www.cimec.gov.ar

B Grupo de Geometría Computacional
Universidad Nacional del Litoral
Ciudad Universitaria, Colectora Ruta Nacional 168, 3000 Santa Fe, Argentina
e-mail: nestor.calvo@gmail.com

C Departamento de Matemáticas e Instituto de Matemáticas del Litoral
Consejo Nacional de Investigaciones Científicas y Técnicas y la Universidad Nacional del Litoral
Parque Tecnológico Litoral Centro, Colectora Ruta Nacional 168, 3000 Santa Fe, Argentina
e-mail: pmorin@intec.unl.edu.ar

D ICREA Research Professor at the International Center for Numerical Methods in Engineering (CIMNE)
Barcelona, Spain.
e-mail: sergio@cimne.upc.edu

Key words: Particle Finite Element method, Lagrangian, Preconditioner, Real Time, High Performance Computing.

Summary. One of the main drawbacks of the explicit integration using Eulerian formulations is the restricted stability of the solution with the time steps and with the spatial discretization. For the case of the Navier-Stokes equations, it is well known that the time step to be used in the solution is stable only for time step smaller than two critical values: the Courant-Friedrichs-Lewy (CFL) number and the Fourier number. The first one is concerning with the convective terms and the second one with the diffusive ones. Both numbers must be less than one to have stable algorithms. For convection dominant problems like high Reynolds number flows, the condition CFL<1 becomes crucial and limit the use of explicit method or outdistance it to be efficient. On the other hand, implicit solutions using Eulerian formulations is restricted in the time step size due to the lack of convergence of the convective non-linear terms. Both time integrations, explicit or implicit are, in most cases, limited to CFL no much larger than one. The possibility to perform parallel processing and the recent upcoming of new processors like GPU and GPGPU increase the possibilities of the explicit
integration in time due to the facility to parallelize explicit methods having results with speed-up closed to one. Although the incompressible condition cannot be solved explicitly, the solution of the momentum conservation equations with an explicit integration of the convective terms together with a parallel processing reduces considerably the computing time to solve the whole problem provided that a large time-step may be preserved independently to the discretization in space. Only to remember the new Particle Finite Element Method, called PFEM 2nd generation (PFEM-2) uses a Lagrangian formulation with an explicit time integrator without the CFL<1 restriction for the convective terms. This allows large time-steps, independent of the spatial discretization, having equal or better precision that an implicit integration. Moreover, PFEM-2 has two versions, one for moving mesh with permanent remeshing and one for fixed mesh [1]. In this lecture we will present some recent advances in the Particle Finite Element Method (PFEM) to solve the incompressible Navier-Stokes equations coupled with another fields like in multiphysics exploiting some nice features found in the fixed version. On the other hand we will also present the moving mesh version applied to multifluids using a parallel remeshing that makes this efficient in terms of cpu time. This updated proposal will be tested numerically and compared in terms of accuracy as in computing cpu time with other more standard Eulerian formulations.

1 INTRODUCTION

During the last years a huge amount of work have been done in order to reduce the computational cost of engineering simulations. Particle based meshless methods like Smooth Particle Hydrodynamics (SPH), explicit cellular automata like Lattice Boltzmann (in particular BGK), enhanced implementations of standard finite element, like edge based FEM, or standard finite volume schemes in graphical processor units (GPU) and particle finite element method arise to be the most predominant alternatives towards this target. These methods and reduced order models also plan to reach real time performance, therefore a big effort should be put in order to fulfill these hard requirements. All these methods have their own advantages and disadvantages. Without enter into details about the reason of selecting the particle finite element method in this paper we try to present the new features of the second generation of this method, called PFEM-2. In [1] the main characteristics of PFEM-2 were introduced where the conclusion was the ability of this novel method to manage very large time steps in a robust way. Here a brief overview of this method is presented, specially the X-IVAS time integration scheme and both, moving and fixed mesh versions are revisited showing their advantages and disadvantages. In [2] some implementation improvements were presented in order to produce a high performance computing. Parallel issues in shared memory architectures with OpenMP were introduced. In [3] an extension to distributed memory architectures is presented. From the analysis carried out in [1] and [2] it is obvious that moving mesh version has the advantage of reducing the numerical diffusion introduced by the projection between mesh and particles and also employs fewer particles for the same level of accuracy. However, mainly the permanent remeshing and also the implicit solution of Poisson equation for the pressure-velocity coupling demand higher cpu time with the drawback of their scalability. Being parallel remeshing currently a big challenge, the development of PFEM-2 was moved towards its fixed mesh version where the remeshing is not needed at all and due to the fact that the mesh remains fixed the Poisson equation matrix is constant for single fluids. A priori factorization used for constant parameter in scalar and NS equations allows to reduce the cost significantly as it is shown in [2,3]. This feature allow to factorize the matrices involved in the computation once at the beginning keeping in memory for future usage reducing the time of assembling and mostly for solving the linear systems. With this nice feature PFEM-2 not only solves the Poisson equation in more efficient way also allows to enlarge the time step drastically solving implicitly the diffusive part of all the equations involved. Up to this point only scalar advection-diffusion problems and Navier-Stokes equations were solved. In [4] a thermally coupled flow solver was presented where the
emphasis was put on the accuracy without searching a finer improvement in the performance. However, the idea is to show how PFEM-2 behaves in problems where several degrees of freedom are solved, as in multi-species flow typical in reactive (combustion) problems. The inclusion of more real industrial applications push the method on a challenge. Turbulent flow in reactive or multi-species problems put a limitation in the usage of a priori factorization because of the changes in the diffusivity of all the fields involved in the mathematical model. Also in multifluids the Poisson equation has its own limitation because the density changes when the different fluids with different physical properties moves over the fixed mesh in background. In the next sections some details about how to solve linear systems with matrices that changes mainly for their physical parameters keeping the mesh fixed are going to be discuss. On the other hand the moving mesh version is also under development. This version has a hard limitation in the lack of scalable parallel remeshing algorithms. Without enter into details in this paper a divide and conquer algorithm is being used for this task with the target in becomes this version competitive.

2 A REVIEW OF PFEM-2

The particle finite element method (PFEM) proposed originally in [8] inspired in the finite point method, a conceptually meshless method uses a lagrangian formulation combined with a fractional step method to solve the pressure with a background mesh that serves to integrate the corresponding equations using a finite element discretization. This method had been applied to solve fluid-structure interaction problems with very success. However in the opinion of their authors for several problems the time step involved in the computations becomes so small that demands a lot of cpu time losing its attractive feature of being almost a meshless method.

Among the main drawbacks detected that may be responsible for this efficiency decreasing the rough time integration and the moving mesh dynamics may be cited. The former produces a bad integration of the nonlinearities and the later restricts severely the time step. In order to circumvent these drawbacks the second generation of PFEM was introduced.

Let $\mathbf{x}_p$ be the vector defining the position of a particle in a 3D space, function of the time $t$ that we will write for simplicity $\mathbf{x}_p^t$. At time $t = t^n$ we will write $\mathbf{x}_p^n$, at time $t = t^n + \Delta t = t^{n+1}$ we will write $\mathbf{x}_p^{n+1}$ and in general, in any time between $t = t^n$ and $t = t^{n+1}$ we will write $\mathbf{x}_p^{n+t}$.

Let $\mathbf{V}^{n+t}(\mathbf{x}_p^{n+t})$ and $\mathbf{A}^{n+t}(\mathbf{x}_p^{n+t})$ vectors defining the velocity and the acceleration respectively of a particle $\mathbf{x}_p$ at any time $t^{n+t}$.

\[
\begin{align*}
\mathbf{V}^{n+t}(\mathbf{x}_p^{n+t}) &= \frac{D\mathbf{x}_p^{n+t}}{Dt} \\
\mathbf{A}^{n+t}(\mathbf{x}_p^{n+t}) &= \frac{D\mathbf{V}_p^{n+t}}{Dt}
\end{align*}
\]  

(2.1) (2.2)

where $\frac{D\phi}{Dt}$ represents the material (Lagrangian) derivative in time of any function $\phi$. The material derivative is connected with the spatial derivative by the convective terms:
In all initial value problems like the transient Navier-Stokes equations, the time solution of a problem consists in: knowing all the variables at time $t = t^n$, find the same variables at time $t = t^{n+1}$. In other words, to integrate in time equations (2.1) and (2.2):

\[
\begin{align*}
\frac{D\phi}{Dt} &= \frac{\partial \phi}{\partial t} + V \frac{\partial \phi}{\partial x}, \\
\phi &= \phi + V^T \nabla \phi
\end{align*}
\]

The accuracy of the results will depend to a great extent in the accuracy of the discretization of the velocity and acceleration in the space, but also in the approximation introduced in the integration of (2.3) and (2.4). In this paper we will be concern with the time integration only, being possible to use any space discretization to achieve analogous results.

### 2.1 Time integration of the velocity

Equation (2.3) may be approximated in different ways. The simplest one is the constant velocity explicit integration in which the velocity is considered constant in the whole time interval with the value of the velocity at the initial position:

\[
\begin{align*}
x^{n+1}_p &= x^n + V^n(x^n)_p \Delta t \\
V^{n+1}(x^{n+1}_p) &= V^n(x^n)_p + \int_{n}^{n+1} A^T(x^n)_p d\tau
\end{align*}
\]

The question is: are there other explicit formulations better than the constant velocity approximation used in (2.5)? Of course there are. We can use previous time steps, like $t^{n-1}$, $t^{n-2}$, etc. to approximate high order time curves. We propose here a different way to improve the explicit integration. The idea is to use the velocity streamlines obtained at time step $t^n$ to approximate the final position of a particle $x^{n+1}_p$.

Let then

\[
x^{n+1}_p \approx x^n + \int_{n}^{n+1} V^n(x^n)_p d\tau
\]

Equation (2.7) is explicit because we are using only information at time step $t^n$. In this case we are not using a constant none a linear, approximation of the velocity field. We are using the same high order approximation the velocity field has at time $t^n$. The only difference with the exact integration (2.3) is that here we are performing the integral (inside each time step) following a pseudo trajectory of the particles calculated with the velocity streamline, instead of following the true trajectory (Fig. 2.1). It must be noted that in the stationary case,
the particle position evaluated with the velocity streamlines and the trajectory are coincident. This time integration will be named Explicit Integration following the Velocity Streamlines (X-IVS).

![Integration following the Velocity Streamlines](image)

**Fig. 1 Integration following the Velocity Streamlines**

### 2.2 Time integration acceleration

In classical explicit integration, equation (2.4) is replaced by:

$$V^{n+1}(x_{p}^{n+1}) = V^n(x_p^n) + A^n(x_p^n) \Delta t$$  \hspace{1cm} (2.8)

The value of $x_p^{n+1}$ may be evaluated with any of the possibility described before:

- **a)** the fully explicit case, that is using (2.5), reduce to:

$$\begin{cases}
  x_{p}^{n+1} = x_{p}^{n} + V^n(x_p^n) \Delta t \\
  V^{n+1}(x_{p}^{n+1}) = V^n(x_p^n) + A^n(x_p^n) \Delta t
\end{cases}$$  \hspace{1cm} (2.9)

- **b)** the X-IVS case, that is using (2.7), remains:

$$\begin{cases}
  x_{p}^{n+1} = x_{p}^{n} + \int_{n}^{n+1} V^n(x_p') dt \\
  V^{n+1}(x_{p}^{n+1}) = V^n(x_p^n) + A^n(x_p^n) \Delta t
\end{cases}$$  \hspace{1cm} (2.10)

In implicit (linear) integration equation (2.4) is replaced by

$$V^{n+1}(x_{p}^{n+1}) = V^n(x_p^n) + A^n(x_p^n)(1-\theta)\Delta t + A^{n+1}(x_{p}^{n+1})\theta \Delta t$$  \hspace{1cm} (2.11)

which also may be used with any of the previous time integrations for the particle position.

However, we will propose something new for evaluating the velocity: the idea proposed in (2.7) may be also used for the acceleration. That is, for improving the time integration of the acceleration while remaining explicit in time. This means to approximate equation (2.4) by:

$$V^{n+\tau}(x_{p}^{n+\tau}) = V^n(x_p^n) + \int_{n}^{n+\tau} A^n(x_p') d\tau$$  \hspace{1cm} (2.12)

Equation (2.12) represents an integration following the acceleration streamlines (See Fig. 2.2) obtained at time $t^n$. This may be solved using any of the particle position integrations described before. For consistence, we will use the X-IVS method described in (2.7):
We must note that equations (2.13) are still explicit because they are using the velocity and acceleration at time \( t^n \) (Fig. 2.2). Nevertheless, they do not assume any constant or linear variation in time neither the velocity nor the acceleration, as is the standard assumption in classical explicit or implicit integration scheme. This approach will be named Explicit Integration following the Velocity and Acceleration Streamlines (X-IVAS).

The position of the particles \( x_p^{n+1} \) is used for remeshing in the moving mesh version or for the projection between particles and mesh in the fixed mesh version. As it was mentioned in the introduction both versions present some advantages and disadvantages. However the fixed version seems to be much more efficient than the first one. In this version a fixed mesh is used with particles that move over the mesh in background. One of the nice features of this version is the removing of the remeshing. Another one is the efficiency in the resolution of the linear system based on the fact that the matrix involved in the linear system that being constant may be factorized once at the beginning and used along the simulation in an efficient way. This feature has shown this advantage in [2] for constant physical parameters. In the next sections this feature is revisited in cases where the mesh remains fixed but the physical coefficients changes with space and time, typical in real engineering situations.

3 A PRIORI FACTORIZATION METHOD TO SOLVE POISSON EQUATION FOR VELOCITY-PRESSURE COUPLING

As it was shown in the above section the velocity-pressure coupling in PFEM-2 is carried out via a Poisson like pressure equation. For single phase problems, where the density is constant, this equation may be written as:

\[
\Delta t \cdot \nabla \left( \frac{1}{\rho} \nabla (\theta * p^{n+1} + (1-\theta) * p^n) \right) = \nabla \cdot \mathbf{v}
\]  

(3.1)

In multifluids or multiphase problems the density becomes a function of another variable to be solved, the void fraction, i.e:

\[
\rho = \rho(vof(x,t))
\]  

(3.2)

\[
\frac{D(vof)}{Dt} = 0
\]  

(3.3)

Therefore, the density is not more constant, neither in space nor in time avoiding the usage of the a priori factorization. Moreover in mutiphase flows where the resolution scale produces a strong separation of the fluids through a very thin interface the density jumps. This feature has an impact on the stiffness of the linear system to be solved.

In order to keep the advantage of an a priori factorization instead of solving the linear system with a direct method we change for an iterative one using a preconditioner based on an a priori factorization of a matrix that remains constant and another simple diagonal matrix that change at each time step. The former is built from the topological mesh with a constant diffusivity (density), remember that the left side of the Poisson equation behaves like a diffusion, and the later contains the inverse of the square root of the density and apply both at left and right of the former, i.e:
This preconditioning matrix $P$ is used inside a preconditioned conjugate gradient iterative method to solve the Poisson equation and it is based on the incomplete Cholesky factorization of the topological part of the matrix that remaining fixed during the computation is only computed once at the beginning.

4 A PRIORI FACTORIZATION METHOD TO SOLVE THE DIFFUSION OF TRANSPORT EQUATIONS

This section extends the above ideas of preconditioning the iterative solver using a preconditioner that contains at one side the topology of the mesh solved as good as possible via an incomplete Cholesky factorization corrected by the variable in space and time physical diffusivity via a very simple diagonal matrix. In this case it is applied to the diffusive part of the transport equation that for efficiency reasons is solved implicitly. Explicit schemes for this kind of problems seem to be expensive in terms of cpu time. Even though enhanced schemes may be written to enlarge the time step for heat equation [5], it may be proved that this strategy does not produce advantages in terms of cpu time spent in the computation.

Using a lagrangian formulation and a fractional step methodology for the velocity pressure coupling PFEM-2 may solve the momentum predictor in an explicit way or implicitly. In order to enlarge the time step drastically the last case is preferred but only if the linear system may take advantage of the a priori factorization of the matrix involved. First the momentum equation is split in two steps, first a lagrangian X-IVAS integration is carried out without the diffusive terms and afterwards the diffusive part is solved implicitly. For this second step the following discrete equation is found:

\[
\left( M + \Delta t * K_\mu \right) * U = F \\
\left( M + \Delta t * K_\mu \right) \approx P = Q^T * Q \\
Q = \sqrt{M} + \sqrt{\Delta t} * R * D \\
D = diag\left( \sqrt{\mu} \right) \\
R = IC\left( K_{p=1} \right)
\]  

in an iterative way using again a preconditioned conjugate gradient (PCG) due to the fact that the linear system is symmetric and positive definite.

In this way it is possible to compute the a priori factorization only once at the beginning and afterwards using it with some matrix-vector products for getting a preconditioner for the PCG that behaves in a very good manner for large time steps. If the time step is short, i.e. Fourier number is low, this preconditioner is worse than the standard solution of the linear system but only by a constant. Normally PFEM-2 looks for enlarge the time steps as much as possible in order to accelerate the computation.
5 SOME RESULTS

In this section some results obtained with the preconditioners introduced in the last two sections are presented. In all the cases the domain is a unitary square discretized with different number of degree of freedom with a fixed and constant right hand side.

5.1 STEADY CASES

Three cases were solved, the first one with a constant diffusivity, the second one with the diffusivity parameter varying smoothly in space and the last one with a sharp diffusivity variation through the interface. The first two cases are not included here for brevity reasons. These results show a very good behavior of the preconditioner when compared with the standard diagonal preconditioner, both in iteration numbers and in cpu time. For the last case a parameter P is used for the sharpness. Different ways of factorization were assessed. In terms of cpu time it may be concluded that the incomplete Cholesky (LU) factorization with some tolerance (4 levels) for the filling results a very good choice in general. The case B presents a big ratio between the maximum and the minimum value of the diffusivity but with the minimum value not so close to zero while the case C is similar to case B but with the minimum value close to zero.

![Case B](image1)

Fig. 2: Case B – Steady – P=100 – CPU time vs #of dofs

![Case C](image2)
5.2 UNSTEADY CASES

For the unsteady case we have proved the same diffusivity field as before but now in a transient context. As in this case a new parameter is added, the Fourier number, then we plot the cpu time vs Fourier number for a fixed number of dofs.

The results show more or less the same conclusion where it is shown that as large the time step better the new preconditioner and as short the time step better the standard diagonal preconditioner.
Fig. 6: Case B – Unsteady – P=100 – CPU time vs #of dofs

Fig. 7: Case B – Unsteady – P=1000 – CPU time vs #of dofs

Fig. 8: Case C – Unsteady – P=100 – CPU time vs #of dofs
6 CONCLUSIONS

- PFEM-2 has shown to be a competitive method to reduce drastically the cpu time involved in the computations due to its inherent stability.
- Also it has an accuracy level acceptable for engineering applications even though the time-steps used are high enough. A better accuracy may be reach reducing the time step arriving at the same level of that reported in scientific papers. The advantage here is the possibility of enlarging or reducing the time-steps at demands in an stable way.
- A good preconditioner for discrete steady and unsteady elliptic equations with variable in space and time physical coefficients is presented in this paper that may be extended for more general applications. This preconditioner allows PFEM-2 to include turbulent effects, reactive systems where multi-species are produced and also multifluids with sharp interfaces.

REFERENCES


NOVEL KINETIC CONSISTENT 3D MHD ALGORITHM FOR HIGH PERFORMANCE PARALLEL COMPUTING SYSTEMS

B. CHETVERUSHKIN*, N. D’ASCEndo† AND V. SAVELIEV*†

* Keldysh Institute of Applied Mathematics (KIAM)
Russian Academy of Science
Miuuskaya sq. 4, 125047 Moscow, Russia
e-mail: chetver@imamod.ru, saveliev@mail.desy.de, www.kiam.ru

† Deutsches Elektronen Synchrotron (DESY)
Notkestrasse 85, 22607, Hamburg, Germany
e-mail: ndasc@mail.desy.de, www.desy.de

Key words: Kinetic Consistent Schemes, MHD, Applications, Computing Methods.

Abstract. The impressive progress of the kinetic consistent schemes in the solution of the gas dynamics problems and the development of the effective parallel algorithms for the modern high performance parallel computing systems lead to the development of advanced methods for the solution of the magnetohydrodynamics problems for plasma physics. The novel feature of the method is the formulation of the complex Boltzmann-like distribution function of the kinetic method with the implementation of the electromagnetic interaction term. The numerical method is based on the explicit schemes, due to the logical simplicity and high efficiency of the algorithm and the easy adaptation to the modern high performance parallel computing systems.

1 INTRODUCTION

The tremendous progress in that the development of high performance computing systems, especially expecting drastically new exascale computing systems gives new opportunities for mathematical modelling of most important physical phenomena in present and future.

A feature of the present is the development of technologies and computer systems are well ahead of the software development. The software problems are primarily associated with the complexity of adaptation of the algorithms to the high performance computing systems architecture. In particular they refer to one of the important requirements as the accuracy in combination with the correctness of initial mathematical models. Another requirement for the methods is their logical simplicity and high efficiency at the same time. Numerical algorithms should be simple and transparent from a logical point of view.

One of the important directions to overcome these problems is the development of non-traditional approach to initial mathematical models and computational algorithms. In the present study for the solution of the multidimensional magnetogasdynamics problems kinetic difference schemes are proposed. They are convenient from the physics point of view,
because the gas dynamics and magnetogasdynamics quantities are defined from close relations between the kinetic and gas dynamics description of physics processes [1, 2].

Another aspect is the study of the explicit finite difference schemes, which seem to be preferable for the future high performance parallel computing, especially in term of their simplicity and well adaptation for parallel program realization, including hybrid high performance parallel computing systems. The weakness of explicit schemes is a strictly limited time step that ensures the computational stability. The advanced explicit kinetic finite difference schemes have soft stability condition giving the opportunity to enhance the stability and to use very fine meshes [3].

The mentioned aspects are used for the development of the framework for the study of dynamics of the hot conducting gas media in strong magnetic fields at high performance parallel computing systems.

2 THEORETICAL ISSUES

2.1 Gas Dynamics Processes

The kinetic theory described the gas dynamics by the Boltzmann differential equation through the evolution of the distribution function \( f(x, \xi, t) \) [4]:

\[
\frac{\partial}{\partial t} f(x, \xi, t) + \xi \cdot \nabla f(x, \xi, t) = C(f)
\]

where \( C(f) \) is a nonlinear integral operator which describes the collisions between gas molecules.

This evolution equation is following naturally from relations between the kinetic and gas dynamics description of continuous media. The macroscopic observables such as density, momentum, energy flux as a functions of \( x \) and \( t \) are obtained from the moments of the distribution function with respect to the macroscopic velocity. The evolution equations for these hydrodynamics quantities are obtained by integrating Eq.1 over molecular velocities \( \xi \) with summational invariants \( (m, m\xi, 1/2m\xi^2) \). The computational interest in kinetic formulations of the gas dynamics is high due to the linearity of the differential operator on the left side of Eq.1. Nonlinearity is confined by the collision term, which is generally local in \( x \) and \( t \).

An important feature is that the collision integral vanishes in the equilibrium state, when the local Boltzmann distribution function \( f \) is Maxwellian:

\[
f_0(x, \xi, t) = \frac{\rho(x,t)m^{1/2}}{(2\pi kT(x,t))^{3/2}} \exp\left\{ -\frac{m}{2kT(x,t)}(\xi - u(x,t))^2 \right\}
\]

This leads to use this model for numerical methods and for possible generalizations in order to provide a natural kinetic description of systems of conservation laws. This approximation is sufficient for the hydrodynamics processes and called kinetic approach [1].

2.2 Electromagnetic Processes

In [5] was shown that the electromagnetic field does not destroy the validity of the Boltzmann equation and this opened the way to the implementation of the electromagnetic processes term in the Boltzmann distribution function. From the vector nature of the
electromagnetic interaction, the distribution function should have also vector behaviour and
provide correct kinetic formulation for the evolution of the magnetic field, i.e. the magnetic
field should be generally defined as the momentum of Boltzmann-like distribution function.

Few useful attempts to formulate the vector Boltzmann-like distribution function could be
found in [7, 8, 9].

We propose an evaluation of the electromagnetic processes in context of complex
distribution function including electromagnetic processes. The electromagnetic field is treated
as a complex vector field taking to account the axial nature of the magnetic field, following
[6]:

\[ F = E + iB \]  \hspace{1cm} (3)

For the purposes of magneto hydrodynamics, the effect, which a magnetic field exerts on a
certain volume, is obtained by integrating the electromagnetic stress tensor over the surface of
that volume and the correspondent propagation velocity can be defined as the complex vector
of velocity:

\[ \nu_{em} = u_{em} + iw_{em} \]  \hspace{1cm} (4)

For first approximation the term defined of electric forces could be neglected and the
magnetic term defined through the tension of the magnetic field line showing a similarity to
the Alfvén wave mechanism:

\[ w_{em} = \frac{B}{\sqrt{\rho}} \]  \hspace{1cm} (5)

2.3 Proposed Distribution Function for MHD

Using the above definitions we define the local complex Boltzmann-Maxwell distribution
function of magnetohydrodynamics:

\[ f_N(x, \xi, t) = \frac{\rho(x,t)m^{1/2}}{(2\pi kT(x,t))^3/2} \exp\left\{ \frac{-m}{2kT(x,t)}[(\xi - u(x,t)) - iw_{em}]^2 \right\} \]  \hspace{1cm} (6)

The first term on the right-hand side of (6) includes the internal energy of the media and
the second term is the electromagnetic field energy. The gas dynamics observables are real
scalars and vectors. The complex component includes the dynamics of the macroscopic
observables introduced by the evolution of the magnetic field, keeping their specific pseudo-
vectorial nature.

The magnetogasdynamics observables are obtained as integrals of the distribution function
(6) with the summational invariants \( \left( m, m\xi, l/2m\xi^2, m\xi^* \right) \). The integration is performed
respect to the molecular velocities \( \xi \) in the complex plane.

The proposed complex Boltzmann-Maxwell-like distribution function contains the
hydrodynamics terms and the electromagnetic terms. By using the proposed distribution
function to calculate mass, momentum and energy fluxes most of the electromagnetic
contributions are calculated directly, i.e. one does not have to solve the hydrodynamics and
magnetic force components separately.
3 IDEAL MHD SYSTEM OF EQUATIONS

To provide the first step of the formulation of the MHD conservation laws equation, consider the equilibrium gas state with the proposed distribution function. The MHD system of equations is obtained by the integration of (1) with vanishing collision integral with the summational invariants following the definition in (6):

\[ \int m \frac{\partial f}{\partial t} + \int m \xi \cdot \nabla f d^3 \xi = 0 \]
\[ \int m \xi \frac{\partial f}{\partial t} + \int m \xi \xi \cdot \nabla f d^3 \xi = 0 \]
\[ \int \frac{1}{2} m \xi^2 \xi \frac{\partial f}{\partial t} + \int \frac{1}{2} m \xi^2 \xi \cdot \nabla f d^3 \xi = 0 \]
\[ \int m \xi^2 \xi \frac{\partial f}{\partial t} + \int m \xi^2 \xi \cdot \nabla f d^3 \xi = 0 \]

The result is the set of Eq.8 which is the ideal magnetohydrodynamics system of equations:

\[ \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j} \rho u_j = 0 \]
\[ \frac{\partial}{\partial t} \rho u_i + \frac{\partial}{\partial x_k} \left( p + \frac{B^2}{2} \right) \delta_{ik} + \rho u_i u_k - B_i B_k = 0 \]
\[ \frac{\partial E}{\partial t} + \frac{\partial}{\partial x_i} \left[ u_i \left( E + p + \frac{B^2}{2} \right) - B_i u_k B_k \right] = 0 \]
\[ \frac{\partial B_i}{\partial t} + \frac{\partial}{\partial x_k} \left[ u_k B_i - u_i B_k \right] = 0 \]

4 KINETIC CONSISTENT MHD FINITE DIFFERENCE SCHEME

The model of the kinetic consistent difference schemes is based on the discrete model of evolution of the distribution function and is formulated directly from the Boltzmann kinetic equation.

Let's consider the local volume of the gas with the distribution function at the time \( t \). The evolution of the distribution function by first order difference scheme for the kinetic Boltzmann equation can be written as:

\[ \frac{f^{i+1} - f^{i}}{\Delta t} = \frac{f^{i+1} - f^{i-1}}{2\Delta x} - \frac{|\xi| f_i^{i+1} - 2f_i^{i} + f_i^{i-1}}{2} + C(f^i) \]

As mentioned before collisions of particles lead to the establishment of the equilibrium state which is adequately described by the single-particle Maxwell distribution function with vanishing of the collision integral in the right part of the balance relations.

Using this feature, let's consider the conditions when the gas is characterized by the equilibrium states, as time or space is comparable with free mean path. The time evolution of the distribution function can be represented as the time evolution of the local Maxwellian
distribution function in discrete moments:

- At time \( t^i \), on each cell, the locally constant one-particle Maxwellian distribution function is defined:

\[
f_M = \frac{\rho m_0}{\sqrt{2\pi kT}} \exp \left( \frac{m}{kT} \left( \xi - u \right) - \frac{\beta^2 u^2}{2} \right)
\]  \hspace{1cm} (10)

where the gas dynamics parameters \( \rho, u, T, B \) are not varied on the cell.

- During the time interval \( \Delta t = t^{i+1} - t^i \) a collisionless processes of the gas dynamics occurs

- At time \( t^{i+1} \) the distribution function is instantaneously maxwellised

- For the time \( t^{i+2} \) these processes are repeated

The kinetic consistent differential scheme in this case can be written:

\[
\frac{f^{i+1} - f^i}{\Delta t} + \frac{1}{\Delta V} \frac{\partial f^i}{\partial \xi} \Delta \sigma_i = \frac{\Delta x^2}{2} \frac{\Delta x_i}{\Delta \xi} \frac{\partial f^i}{\partial \xi} \Delta \sigma_i
\]  \hspace{1cm} (12)

where:

- \( \Delta \sigma_i \) is the surface element \( \Delta x_k \Delta x_m \) perpendicular to the direction \( \xi \)

- \( f^i \) is the value of the distribution function at the surface \( \sigma \) between the two volume elements \( I_i \) and \( I_{i+1} \).

The kinetic consistent scheme of the conservation laws of the macroscopic observables for 3D magnetohydrodynamics processes could be obtained by integrating the balance relation (12) with the summational invariants \((m, m\mathbf{\xi}/2m\mathbf{\xi}, m\mathbf{\xi}^*)\), using the same integration rules as in Eq.8:

\[
\frac{\rho^{i+1} - \rho^i}{\Delta t} + (\rho u_i) \frac{\Delta x_i}{2} \left[ \rho u_i \text{Erf} (\beta u_i) + \frac{\rho}{\sqrt{\pi} \beta} e^{-\beta^2 u_i^2} \right]_{\xi \xi_i} = \left( \left( p + B^2 \right) \delta_{ik} + \rho u_i u_k - B_i B_k \right)_{\xi \xi_k}
\]  \hspace{1cm} (13)
Where \( \beta = \sqrt{\frac{\rho}{2p+B^2}}, i,k = 1,\ldots,3, x_k = (x,y,z) \).

5 KINETIC CONSISTENT QUASI MHD EQUATIONS

The kinetic quasi magnetogasdynamics system of equations is closely related to the kinetic consistent scheme and represents a differential form notation for the numerical algorithms.

The balance relation Eq.12 can be rewritten as:

\[
\frac{\partial f}{\partial t} + \frac{1}{\Delta V} \int_{\sigma} \xi_{f\sigma} \, d\sigma = \frac{1}{2\Delta V} \int_{\sigma} |\xi_i| \Delta x_i \frac{\partial f}{\partial x_i} \, d\sigma = \frac{1}{2\Delta V} \int_{\sigma} \tau_i^2 \frac{\partial f}{\partial x_i} \, d\sigma
\]

Using the Gauss-Ostrogradsky formula it is possible to transform Eq.14 to the differential form:

\[
\frac{\partial f}{\partial t} + \nabla \cdot (\xi f_{\xi}) = \frac{\tau_i}{2\alpha_x \alpha_{x_k}} \frac{\partial}{\partial x_i} \xi_j^k f_{f_{\xi}}^j
\]

The quasi magneto gas dynamics system of equations involves explicitly two \( \tau \) parameters. Gas dynamics processes are introduced by the quantity \( \tau \) that corresponds to the time of free distance flight of particles, or the characteristic time of particle collisions. By analogy the quantity \( \tau_m \) is introduced as the characteristic time of propagation of gas dynamics by electromagnetic processes. The characteristic time values \( \tau \) and \( \tau_m \) are defined respectively for hydrodynamics and electromagnetic processes equation:

\[
\tau = \alpha \frac{\Delta x_i}{c_h} \quad \tau_m = \alpha_m \frac{\Delta x_i}{c_m}
\]

where:

\( \Delta x_i \) is the size of the computational cell,

\( c_h, c_m \) are the sound and Alfven speeds in the computational cell.

The introduction of the physical meaning of the characteristic time values \( \tau \) and \( \tau_m \) provides an important contribution into the understanding of the processes and the simplification of the numerical scheme in Eq.15.

The evolution equations of the gas dynamics parameters and the magnetic field are obtained from Eq.15 by integration with the summation invariants \( (m, \frac{m \xi_1}{\alpha}, \frac{1}{2}m \xi^2, m\xi^* \) over the molecular velocities under the assumption:

\[
\int f^{i+1} \phi(\xi) d\xi = \int f_{\xi}^{i+1} \phi(\xi) d\xi
\]

The integration is performed as in Eq.8 and Eq.13. The hydrodynamics and magnetic field quantities are obtained respectively as the real and imaginary parts of the path integral in the complex plane.

The resulting 3D kinetic consistent quasi MHD equations are:

\[
\frac{\rho^{i+1} - \rho^i}{\Delta t} + \frac{\partial}{\partial x_i} \rho u_i = \frac{\tau_i}{2 \alpha_x} \frac{\partial}{\partial x_k} \left[ \left( p + \frac{B^2}{2} \right) \delta_{ik} + \rho u_i u_k - B_i B_k \right]
\]

\[
\frac{\rho^{i+1} u_i^{i+1} - \rho^i u_i^i}{\Delta t} + \frac{\partial}{\partial x_k} \left[ \left( p + \frac{B^2}{2} \right) \delta_{ik} + \rho u_i u_k - B_i B_k \right] = \]

983
The dissipative terms appear because the construction of the quasi magnetogas dynamics system is based on the assumption that the distribution function slightly changes over the distance between neighbor cells, which is related to the characteristic times $\tau$ and $\tau_m$. It was shown in [2] that the dissipative terms of the quasi gas dynamics system, on the right side, are small in comparison with the convective terms. With the condition of cell size equivalent of the free path they converge to the viscous terms of the corresponding Navier-Stokes equations, i.e. the correspondent dissipative terms are associated with real physics processes. An important remark is that in this case the gas dynamics parameters as viscosity and heat conductivity converge from the kinetic theory.

The Navier-Stokes thermal flux vector is identified within the dissipative terms of the energy equation:

$$
\frac{\tau}{2} \frac{\partial}{\partial x_i} \left[ \frac{B_i}{\rho} \left( p + \frac{B^2}{2} \right) \delta_{im} - \frac{B_k}{\rho} \left( p + \frac{B^2}{2} \right) \delta_{ik} \right] - \left( B_i u_k - B_k u_i \right) =
$$

$$
\frac{\tau}{2} \frac{\partial}{\partial x_i} \left[ \frac{B_i}{\rho} \left( p + \frac{B^2}{2} \right) \delta_{ik} + \frac{B_k}{\rho} \left( p + \frac{B^2}{2} \right) \delta_{im} - \frac{B_m}{\rho} \left( p + \frac{B^2}{2} \right) \delta_{ik} \right] - u_i u_k B_m - u_k u_m B_i - \frac{B_i B_k B_m}{\rho} =
$$

The similar analysis of dissipative terms of the electromagnetic processes gives the estimation of the smallness of values of dissipative terms of the electromagnetic processes and with the correct condition for the size of cells the equation converges to the correct representation of the magnetic viscosity. The resistivity is identified within the dissipative term of the magnetic field evolution equation:
\[ \Pi_{ik}^{\tau} = \frac{\tau m}{2} \left[ \left( p + \frac{B^2}{2} \right) \frac{\partial B_i}{\partial x_k} - \frac{\partial B_k}{\partial x_i} \right] = \eta \left[ \frac{\partial B_i}{\partial x_k} - \frac{\partial B_k}{\partial x_i} \right] \] 

6 COMPUTATIONAL ALGORITHM

The numerical algorithm uses a Cartesian, staggered, divergence free mesh configuration detailed description presented in [10]. The hydrodynamics observables: mass density, momentum and energy density are defined at the cell centre. The components of the magnetic field are defined at the face centres of cells. A duality is established between the fluxes and electric field at the edges. The electric field is then utilized to make an update of the magnetic fields that preserves the solenoidal nature of the magnetic fields and ensures that the magnetic fields in an magneto hydrodynamics modelling remain strictly solenoidal up to discretization errors.

Generally we are using the explicit numerical scheme, believing that it is perspective for the modern high performance computing systems due to the logical simplicity and efficiency of the algorithms. The finite volume method is used to update the conserved observables: mass, momentum and energy by calculating the fluxes of these observables across the cell face. The update of the magnetic field is more complicated procedure and performed via electric field integration along the edge of the cells. For the calculation of the observables is used the proposed distribution function method described above. For the time evolution is used explicit scheme of the integration of the quasi magnetohydrodynamics system of equations. The code uses a variable time steps, the time step in an explicit scheme is controlled by a Courant type condition on the time step estimation [2].

7 RESULTS OF NUMERICAL MODELING

The simulation framework is created on the base of Fortran 90 and c++, parallel implementation on MPI. The demonstration of the performance of the present method is performed on the base of the solution of the spherical expansion problem of ionised gas and the solution of the expansion problem of ionised gas in strong magnetic field. The simulations are performed for the Cartesian rectangular mesh 100 × 100 × 100 in the physics domain [0,1].

The first solution is the three dimensional thermal expansion of the ionised gas in the physical domain. The initial conditions are represented as the unit cell in the centre of the physical region with pressure of 100 in comparison to the overall area with pressure 1. Fig. 1,2,3 present the state of the simulation in time 0.03. On the 3D picture the arrows represent the velocities of the ionised gas and the colour represent the density of gas.

The second solution is a three dimensional thermal expansion of ionised gas in a strong uniform magnetic field along the z coordinate. The initial conditions are the same, in addition the magnetic field is \( 5/\sqrt{\pi} \). On the Fig. 4,5,6 the view is shown of the process of expansion of the hot ionized gas in a strong uniform magnetic field. The picture shows the time when the conductive gas is confined in the cylindrical area of space along z.
Figure 1: 3D gas expansion view and 2D gas density projection

Figure 2: 2D pressure and 2D kinetic energy projection

Figure 3: 1D density profile
Figure 4: 3D gas expansion in magnetic field and 2D gas density projection

Figure 5: 2D gas pressure and 2D magnetic pressure projections

Figure 6: 2D kinetic energy projection and 1D gas density profile
8 CONCLUSIONS

A new 3D kinetic consistent algorithm has been developed for the solution of magnetohydrodynamics problems. The novel feature of the method is that the local complex Boltzmann-like distribution function incorporated most of the electromagnetic processes terms. The fluxes of mass, momentum and energy across cell interface as well as the magnetic field are calculated by integrating a local complex Boltzmann-like distribution function over the velocity space.

Results of the numerical simulations demonstrate that the proposed method can achieve high numerical accuracy and resolves strong shock waves of the magneto gas dynamics problems.

REFERENCES

ON THE DIFFICULTIES OF REAL-TIME CO-SIMULATION

G. STETTINGER*, M. BENEDIKT†, N. THEK† AND J. ZEHETNER†

*Alpen-Adria Universität Klagenfurt
Institute of Smart System Technologies
Universitätsstrasse 65-67, 9020 Klagenfurt, Austria
e-mail: georg.stettinger@aau.at
web page: http://www.uni-klu.ac.at/tewi/ict/sst/me/index.html

†Virtual Vehicle Research Center
Area E - Electrics/Electronics and Software
Inffeldgasse 21/A/2, 8010 Graz, Austria
e-mail: martin.benedikt@v2c2.at, norbert.thek@v2c2.at, josef.zehetner@v2c2.at
web page: http://www.v2c2.at

Key words: Co-Simulation, Real-Time, Extrapolation, Coupling Types, Dead-Time

Abstract. In a co-simulation, subsystems are coupled via their in- and outputs to simulate the overall system behaviour. The subsystems are modelled in their domain specific simulation tools. The task changes if one coupled subsystem represents a real-time system. A real-time system which has to guarantees hard-real-time conditions influences the co-simulation concept: now the co-simulation also has to fulfill hard-real-time conditions. This type of co-simulation is called real-time co-simulation. The most important difference to a non-real-time co-simulation is the time correct overall simulation speed with respect to the involved real-time systems. To achieve this, all subsystems in form of non-real-time systems have to be synchronised to the involved real-time systems. The focus of this work lies on the problems that occur in a real-time co-simulation environment compared to a classical one. A concept to handle the additional problems is outlined and tested on an example real-time co-simulation.

1 INTRODUCTION

The co-simulation concept is often used in a modern vehicle development process in the field of the automotive industry [3, 5, 6]. Different subsystems are modelled in their domain specific simulation tools. To get an overall simulation result all subsystems have to be combined via their specific in- and outputs. The correct coupling of these signals is the main task of a so called co-simulation platform. The co-simulation concept has the big
advantage that all subsystems can use their domain specific solvers and no adaptations in the existing simulations are necessary [1]. Often the classical co-simulation concept is not enough to solve all occurring problems in the vehicle development process. One extension is the integration of hard-real-time systems in the co-simulation. This is the case when at least one modelled subsystem gets replaced by real hardware, e.g. HiL (Hardware in the Loop) systems or test benches. In this case the co-simulation is called real-time co-simulation and now the typical hard real-time conditions must hold during the simulation [3, 8]. The main difference compared to a classical co-simulation is the correct coupling with respect to the wall-clock-time. The involved real-time systems which fulfill the hard-real-time conditions run with the so called wall-clock-time. The paper is organised as follows: Chapter 2 discusses the difficulties of a classical co-simulation. Chapter 3 describes the additional problems occurring in a real-time co-simulation. Chapter 4 shortly illustrates the structure of a real-time co-simulation. In chapter 5 an example real-time co-simulation is presented. The difficulties are shown and possible approaches to solve these problems are outlined.

The presented work is part of the ongoing research project ACoRTA (Advanced Co-Simulation Methods for Real-Time Applications) at the Virtual Vehicle Competence Center in Graz, Austria. All techniques to handle the real-time co-simulation problem are implemented in the ICOS Real-Time Framework (ICOS RT).

2 DIFFICULTIES OF A CO-SIMULATION

Typically co-simulation problems are handled with the help of so called co-simulation platforms, which offer specific interfacing capabilities of the simulation tools and perform the data exchange between involved subsystems. The main requirements of such classical co-simulation problems are [7, 9, 10]:

- Proper choice of the coupling time instants
- Selection of an extrapolation method
- Definition of the subsystem scheduling

2.1 Choice of the coupling time instants (macro-time-step)

To distinguish between the step sizes of domain specific numerical solvers of the involved subsystems and the step sizes between two coupling instants, the micro- ($\delta T$) and macro-time-steps ($\Delta T$) are introduced (see Fig. 1). Every domain specific simulation tool typically has its own micro-time-step which doesn’t get influenced during the co-simulation. For a co-simulation only the macro-time-step has to be defined by the user. The involved subsystems get synchronised at every macro-time-step (which may be fixed or adaptive) [2, 4]. The coupling data exchange is done with the help of a co-simulation
platform (e.g. the ICOS\textsuperscript{2}, Independent Co-Simulation Framework). The choice of the macro-time-step is very important for the quality of the overall simulation result because often a trade-off between the overall simulation time and the desired accuracy of the simulation results. In conclusion, smaller macro-time-steps lead to more accurate simulation results but with the drawback of longer simulation times [7, 10].

2.2 Signal-based extrapolation techniques

In the case of two interdependent subsystems, which are connected via their in- and outputs, at least one input has to be extrapolated to solve the coupled system [7]. In a classical co-simulation typically polynomial extrapolation techniques of low order are used to solve this problem [4, 5]. Fig. 1 shows the most commonly used polynomial extrapolation techniques such as zero-order extrapolation (zero-order hold, ZOH), first-order extrapolation (first-order hold, FOH) and second-order extrapolation (second-order hold, SOH). Due to the fact that every extrapolation technique is a prediction of the future coupling signal an error gets introduced [7]. This coupling error influences the overall simulation result and thus, this error must be as small as possible. Another important problem is the introduction of dead-times due to the used polynomial extrapolation. This dead-times can lead to instability in closed loop systems. The introduced dead-times mainly depend on the chosen macro-time-step (see Fig. 2) [7, 9].

2.3 Simulation tool scheduling

In principle, there are two different ways of subsystem scheduling: parallel and sequential. In the parallel case all inputs of the coupled subsystems has to be extrapolated
which lead to shorter simulation times, with the drawback of more extrapolation errors. Sequential scheduling requires less extrapolation effort with the the disadvantageous fact of longer simulation times. Additionally the execution sequence influences the entire system behavior [6, 7].

3 COMMENTS ON REAL-TIME CO-SIMULATION

For real-time co-simulation basically the same problems as for classical non-real-time co-simulation occur. One important aspect for a real-time co-simulation are hard real-time conditions which have to be fulfilled during a real-time co-simulation which is responsible for the following problems [9]:

1. Requirements for coupleable subsystems
2. Synchronisation of the involved subsystems
3. Handling of dead-times
4. Extrapolation of noisy coupling signals

These additional difficulties are discussed in detail in the following sections.

3.1 Requirements for coupleable subsystems

To fulfill the required hard-real-time conditions a time correct exchange of coupling data with respect to the \textit{wall-clock-time} is necessary. This is only possible if the involved non-real-time systems, which do not guarantee hard-real-time conditions, have a simulation time which is faster than the \textit{wall-clock-time}. Otherwise no synchronisation between the involved real-time and non-real-time systems is possible [9].

3.2 Synchronisation of the involved subsystems

The main task of the real-time co-simulation problem is the time correct exchange of coupling data with respect to the \textit{wall-clock-time}. The synchronisation is typically realised
via a slow down mechanism for the non-real-time systems. The simulation speed of these systems is faster than the wall-clock time and so these systems have to slow down until both system types run with the wall-clock time to perform a time correct synchronisation. So the involved non-real-time systems get paused to realise this slow down effect [9].

3.3 Handling of dead-times

To guarantee hard-real-time conditions a time correct coupling is mandatory as discussed in 3.2. So the communication delays due to the communication medium between coupled systems play an important role. These dead-times influence the stability of the closed loop system and can violate the hard-real-time conditions when the synchronisation fails. The coupling mechanism between the coupled systems split the communication dead-times into sending and receiving dead-times (see Fig. 3). To guarantee the required hard-real-time conditions an approximate compensation of the sending and receiving dead-times is required [9].

![Figure 3: Sending and receiving dead-time due to the data transfer [9]](image)

3.4 Extrapolation of noisy coupling signals

As discussed before, extrapolation is necessary to solve the co-simulation problem [9]. If real-time systems enter the co-simulation several coupling signals are typically measured with sensors and are often corrupted by noise. Signal-based extrapolation techniques, which are typically used for extrapolation, amplify the noisy measurements and so they cannot provide reliable coupling signals. So extrapolation methods which are robust against noisy measurements are required. To satisfy this requirement e.g. model-based extrapolation methods can be used [9].

4 STRUCTURE OF A REAL-TIME CO-SIMULATION

The focus of this work lies on the coupling of real-time and non-real-time systems. The synchronisation of the involved real-time and non-real-time systems is done via ICOS RT (see Fig. 4). ICOS RT is responsible for the time correct communication between the involved subsystems. So the slow down mechanism is the most important task of ICOS
RT to perform a correct synchronisation between the real-time and non-real-time systems. Two additional coupling types which are not part of this work are handled by ICOS RT:

![Diagram of coupling real-time and non-real-time systems](image)

**Figure 4:** Coupling of real-time and non-real-time systems [9]

The coupling of real-time systems and the coupling of tasks on one real-time system [9].

## 5 EXAMPLE: MILD HYBRID

The following example demonstrates the synchronisation mechanism of a real-time co-simulation.

### 5.1 Model description

The used example represents a hybrid vehicle which consists of seven subsystems (see Fig. 5). These involved subsystems are modelled in different simulations tools (*MATLAB/SIMULINK*[^3], *KULI*[^4] and *AVL CRUISE*[^5]):

- **Vehicle (**AVL CRUISE**):**
  The vehicle model includes the power train of a parallel hybrid vehicle. This subsystem interacts with 17 inputs and 8 outputs with the other involved subsystems. The micro- and macro-time-steps are set to $\delta T = 0.001s$ and $\Delta T = 0.05s$, respectively.

- **Thermal Network (**KULI**):**
  The thermal network model includes the cooling circuit of the vehicle.
Inputs: 9; Outputs: 12; $\delta T = 1s$; $\Delta T = 1s$.

- **Cockpit (**MATLAB/SIMULINK**):**
  The cockpit model describes the driver of the hybrid car. So this system defines the velocity profile and the gear selection.
Inputs: 1; Outputs: 7; $\delta T = 0.01s$; $\Delta T = 0.05s$.

- **Hybrid Module (**MATLAB/SIMULINK**):**
  The hybrid module includes the parallel hybrid structure of the vehicle. So the

[^3]: http://www.mathworks.com
[^4]: http://www.kuli.at
[^5]: http://wwwavl.com/cruise1
recuperation of the brake energy and the partition of the desired driving torque to the combustion and electric engine are the main tasks of this subsystem. Inputs: 7; Outputs: 5; δT = 0.01s; ΔT = 0.05s.

- **Energy Management System (MATLAB/SIMULINK):**
  This subsystem includes the energy management of the hybrid vehicle. Especially the energy exchange between the super cap, the electric motor and the Li-Ion battery is represented herein.
  Inputs: 8; Outputs: 6; δT = 0.01s; ΔT = 0.05s.

- **Super Cap (MATLAB/SIMULINK):**
  The super cap subsystem models the behaviour of the super cap.
  Inputs: 1; Outputs: 4; δT = 0.01s; ΔT = 0.05s.

- **Li-Ion Battery (MATLAB/SIMULINK):**
  The Li-Ion battery subsystem models the behaviour of the Li-Ion battery.
  Inputs: 2; Outputs: 5; δT = 0.01s; ΔT = 0.05s.
5.2 Real-time co-simulation configuration

For this example the four involved MATLAB/SIMULINK models are compiled and run on an ETAS real-time system (see Fig. 6). The non-real-time part includes the KULI, one MATLAB/SIMULINK and the AVL CRUISE model. The subsystems of the non-real-time part are coupled via the ICOS Co-Simulation framework to achieve a correct interaction of these subsystems. The essential part of the real-time co-simulation is the synchronisation of the real-time and non-real-time part. This synchronisation mechanism slows down the non-real-time part such that a time correct interaction between the real-time- and non-real-time part is performed. The macro-time-step, where coupling data between the real-time- and non-real-time system gets exchanged, is set to $\Delta T = 200 ms$. The overall real-time co-simulation, consisting of the real-time and non-real-time part, use a parallel scheduling while the subsystems in the non-real-time part are sequential scheduled.

![Figure 6: Configuration of the real-time co-simulation](image)

5.3 Simulation results

Fig. 7 shows the simulation results of the real-time co-simulation compared to a non-real-time realisation of the same system. In the case of the non-real-time co-simulation all subsystems are simulated using the co-simulation platform ICOS. The two simulation results are nearly the same over the simulation time. This means that a subsystem of the non-real-time co-simulation, which runs faster than the wall-clock time, can be replaced by a real-time system with the same functionality and the simulation result is still reliable. The small deviations between the simulation results stem from the communication medium between the real-time- and non-real-time part of the real-time co-simulation. Especially the introduced dead-times influence the overall simulation result. Fig. 8 shows the synchronisation effect of the time correct coupling: All involved subsystems have a simulation speed which is close to the wall-clock time due to the slow down mechanism.
So a correct interaction between the real-time and non-real-time systems is possible. Fig. 9 shows the slow down effect of the real-time co-simulation compared to the non-real-time
realisation. In the non-real-time case the whole simulation is done via ICOS where all involved subsystems have a simulation speed which is faster than the wall-clock time. With the help of ICOS RT the non-real-time systems slow down and a synchronisation with the real-time system is possible.

![Graph](image)

**Figure 9:** Slow down effect of the coupling element

### 6 CONCLUSIONS & OUTLOOK

The integration of hard real-time systems into co-simulation frameworks faces several challenges. The presented approach shows that a real-time co-simulation is handable if the non-real-time systems run faster than the wall-clock time. The required synchronisation with respect to the wall-clock time is realised via the discussed slow down mechanism. Future activities will concentrate on dead-time handling due to the involved communication medium and on extrapolation techniques which are robust against noisy coupling signals. These difficulties are focused in the ongoing research project ACoRTA at ViF, together with Porsche, AVL and Uni Klagenfurt. By solving these problems co-simulation under hard-real-time conditions will be possible.

### REFERENCES


COMPARISON BETWEEN TWO DIFFERENT DECOMPOSITIONS FOR THE SOLUTION OF FLUID-STRUCTURE INTERACTION PROBLEMS

JORIS DEGROOTE*, JORIS BOLS* AND LIESBETH TAEELMAN†

*Department of Flow, Heat and Combustion Mechanics
Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
e-mail: Joris.Degroote@UGent.be, web page: http://www.FSI.UGent.be/

†IBiTech-biOMMeda
Ghent University
De Pintelaan 185, B-9000 Ghent, Belgium

Key words: Fluid-Structure Interaction, Gauss-Seidel iteration, Robin Boundary Condition, Artificial Compressibility

Abstract. Several types of decompositions and coupling algorithms can be used when solving a fluid-structure interaction problem in a partitioned way. In the case of Dirichlet-Neumann (DN) decomposition, the flow equations are solved with a Dirichlet boundary condition on the fluid-structure interface, while the structural equations are solved with a Neumann boundary condition on the interface. Robin-Neumann (RN) decomposition denotes a Robin boundary condition on the fluid side of the interface.

It is well-known that Gauss-Seidel iteration is often unstable for strongly coupled problems with DN decomposition. Conversely, this coupling algorithm can have good convergence properties in combination with RN decomposition. The Interface Artificial Compressibility (IAC) method is one of the techniques to improve the convergence of the Gauss-Seidel iterations for cases with DN decomposition.

In this paper, it is demonstrated that there is a common idea behind Gauss-Seidel iterations with RN decomposition and with DN decomposition plus IAC. Both approaches include a local, linear approximation for the structural equations into the flow equations. The numerical examples demonstrate that this approach is very suitable for the flow in flexible tubes, but that the application to other cases is not always straightforward.

1 INTRODUCTION

Different decompositions can be used for partitioned fluid-structure interaction simulations. They are characterized by the boundary conditions at the common boundary of the
fluid and the structure. In the case of Dirichlet-Neumann (DN) decomposition, the flow
equations and the structural equations are solved with respectively a Dirichlet and a Neu-
mann boundary condition on the fluid-structure interface. By contrast, Robin-Neumann
(RN) decomposition implies a Robin boundary condition (i.e. a linear combination of a
Dirichlet and a Neumann boundary condition) on the fluid side of the interface and a
Neumann boundary condition on the structure side [1].

It is well-known that Gauss-Seidel (GS) or fixed-point iteration is often unstable for
strongly coupled problems with DN decomposition [2]. Conversely, this coupling algorithm
can have good convergence properties in combination with RN decomposition. To improve
the convergence of the GS iterations for problems with DN decomposition, the Interface
Artificial Compressibility (IAC) method has been proposed [3]. Despite the name, the
IAC method uses a constant fluid density and the artificial source term disappears at
convergence of the coupling iterations.

The remainder of this paper is organized as follows. Section 2 will explain that a local,
linear approximation for the structural equations is included into the flow equations by
both GS iterations with RN decomposition and with DN decomposition and IAC [4].
Subsequently, two numerical models are described in Section 3. Finally, the numerical
examples in Section 4 will demonstrate that this approach is very suitable for the flow in
flexible tubes, but that the application to other cases is not always straightforward.

2 COMPARISON

2.1 Interface conditions

On the fluid-structure interface $\Gamma_{i}^{n+1}$, the kinematic equilibrium condition

$$\bar{v}^{n+1} = \delta_{t} \bar{u}^{n+1}$$

and the dynamic equilibrium condition

$$\bar{\sigma}_{f}^{n+1} \cdot \bar{n}^{n+1} = \bar{\sigma}_{s}^{n+1} \cdot \bar{n}^{n+1}$$

need to be satisfied. $\bar{u}$ is the structural displacement, $\bar{v}$ the fluid velocity, $\bar{w}$ the grid
velocity and $\bar{\sigma}_{f}$ the Cauchy stress tensor. The vector $\bar{n}$ is the normal pointing outwards
of the fluid subdomain. Backward Euler time discretization is used for simplicity, so the
notation $\delta_{t}$ refers to

$$\delta_{t} \bar{w}^{n+1} = \frac{\bar{u}^{n+1} - \bar{u}^{n}}{\Delta t},$$

with the superscript $n$ denoting the time step and $\Delta t$ the time step size.

In addition, the normal velocity of the fluid grid has to match the normal structural
velocity on $\Gamma_{i}^{n+1}$.

$$\bar{w}^{n+1} \cdot \bar{n}^{n+1} = \delta_{t} \bar{w}^{n+1} \cdot \bar{n}^{n+1}$$

Within each time step, GS coupling iterations are performed between the flow solver
and the structural solver until some convergence criteria for the equilibrium conditions
on the fluid-structure interface are satisfied. The superscript \( k + 1 \) indicates the current coupling iteration in the current time step \((n + 1)\).

2.2 GS iterations with DN decomposition and IAC (GS-DN-IAC)

GS-DN uses a Dirichlet boundary condition on \( \Gamma_k^{k+1} \) for the flow equations, given by

\[
\vec{v}^{k+1} = \delta_t \vec{u}^k
\]  

(2a)

For the grid velocity, a Dirichlet boundary condition

\[
\vec{w}^k \cdot \vec{n}^k = \delta_t \vec{u}^k \cdot \vec{n}^k
\]  

(2b)

is applied. The values of \( \vec{u} \) and \( \vec{n} \) are determined by the structural calculation at the end of coupling iteration \( k \).

While solving the structural equations, the most recent flow values are used in the Neumann boundary condition

\[
\bar{\sigma}_s^{k+1} \cdot \vec{n}^{k+1} = \bar{\sigma}_f^{k+1} \cdot \vec{n}^{k+1}.
\]  

(2c)

When applying a finite volume discretization in arbitrary Lagrangian-Eulerian (ALE) formulation, the conservation of mass for an incompressible fluid is given by

\[
\frac{V_i^k - V_i^n}{\Delta t} + \sum_j \left( \vec{v}_{i,j}^{k+1} - \vec{w}_{i,j}^k \right) \cdot \vec{n}_{i,j}^k S_{i,j}^k = 0,
\]  

(3)

with \( V_i \) the volume of cell \( i \), \( S_{i,j} \) the area of face \( j \) and \( \vec{n}_{i,j} \) the normal pointing outwards. All geometrical values (including \( V_i \), \( S_{i,j} \) and \( \vec{n}_{i,j} \)) correspond with the structural calculation in the previous coupling iteration.

The IAC then adds the source term

\[
- \frac{p_{i,m}^{k+1} - p_{i,m}^k}{\Delta t} \frac{d(\vec{u}_{i,m} \cdot \vec{n}_{i,m})}{dp_{i,m}} S_{i,m}^k
\]  

(4)

to the right-hand side of Eq. (3) but only in cells adjacent to the fluid-structure interface. In [3], the coefficient \( d(\vec{u}_{i,m} \cdot \vec{n}_{i,m})/dp_{i,m} \) is calculated by finite differencing between two structural calculations for different pressures. Together with the first term of Eq. (3), the source term becomes a linear approximation for \((V_i^{k+1} - V_i^n)/\Delta t\).

Combination of the interface conditions for the fluid velocity (Eq. (2b)) and the grid velocity (Eq. (2a)) yields

\[
\vec{w}^k \cdot \vec{n}^k = \delta_t \vec{u}^k \cdot \vec{n}^k = \vec{v}^{k+1} \cdot \vec{n}^k.
\]  

(5)

As a result, the second term in the continuity equation vanishes on the face \( j = m \) that lies on the fluid-structure interface, giving

\[
\frac{V_i^k - V_i^n}{\Delta t} + \sum_{j \neq m} \left( \vec{v}_{i,j}^{k+1} - \vec{w}_{i,j}^k \right) \cdot \vec{n}_{i,j}^k S_{i,j}^k = - \frac{p_{i,m}^{k+1} - p_{i,m}^k}{\Delta t} \frac{d(\vec{u}_{i,m} \cdot \vec{n}_{i,m})}{dp_{i,m}} S_{i,m}^k
\]  

(6)
2.3 GS iterations with RN decomposition (GS-RN)

The GS-RN technique uses a Robin boundary condition on $\Gamma_{k+1}^i$ for the fluid, given by

$$\vec{v}_{k+1}^i + \alpha \vec{u}_f^k \cdot \vec{n}_f^k = \delta_t \vec{u}_s^k + \alpha \vec{u}_s^k \cdot \vec{n}_s^k.$$  (7)

with $\alpha$ a coefficient. In [1], an analytical expression for this coefficient $\alpha$ is obtained by considering a membrane so that the structural equations can be written in the same form as the Robin boundary condition. Moreover, an optimal value for $\alpha$ derived from a Fourier analysis has been proposed in [5]. The boundary conditions for the grid velocity and for the structural equations are identical to those in Section 2.2.

By including the Robin condition, the factor for face $j = m$ on $\Gamma_{k+1}^i$ becomes

$$\left(\vec{v}_{i,j}^{k+1} - \vec{w}_{i,j}^k\right) \cdot \vec{n}_{i,j}^k.$$  (8a)

Substitution of Eq. (2b) and Eq. (2c) leads to

$$\left(\vec{v}_{i,m}^{k+1} - \vec{w}_{i,m}^k\right) \cdot \vec{n}_{i,m}^k = \alpha_{i,m} \left(\vec{u}_{i,m}^k \cdot \vec{n}_{i,m}^k S_{k,i,j}^k - \vec{w}_{i,m}^k \cdot \vec{n}_{i,m}^k S_{k,i,m}^k\right).$$  (9)

2.4 Equivalence

For many applications, $\vec{u}_f$ can be simplified to $-p \vec{I}$. Consequently, Eq. (9) becomes

$$\frac{V_i^k - V_i^n}{\Delta t} + \sum_{j \neq m} \left(\vec{v}_{i,j}^{k+1} - \vec{w}_{i,j}^k\right) \cdot \vec{n}_{i,j}^k S_{i,j}^k = -\alpha_{i,m} \vec{n}_{i,m}^k \cdot \left(\vec{u}_{i,m}^k \cdot \vec{n}_{i,m}^k \frac{p_{i,m}^{k+1} - p_{i,m}^k}{\Delta t}\right) \cdot \vec{n}_{i,m}^k S_{i,m}^k$$  (10)

So, if the viscous tractions on the interface are small, Eq. (6) and Eq. (10) are equal if

$$\alpha_{i,m} = \frac{1}{\Delta t} \left\frac{d(\vec{u}_{i,m} \cdot \vec{n}_{i,m})}{dp_{i,m}}\right.$$  (11)

The Navier-Stokes equations lead to the same result as this analysis based on the conservation of mass.

As could already be seen from Eq. (7), the coefficient $\alpha_{i,m}$ relates a change in velocity of the interface to a change in traction on the interface. If the parameter $\alpha_{i,m}$ (respectively $d(\vec{u}_{i,m} \cdot \vec{n}_{i,m})/dp_{i,m}$) is set so that it approximates the actual velocity/traction (respectively displacement/pressure) relation of the structural model, then an approximation for the structural model is included into the flow calculation. This implies that both GS-RN and GS-DN-IAC take the fluid-structure interaction into account while solving the flow equations, as opposed to GS-DN.
3 MODELS

3.1 Tube

The first case is the propagation of a pressure wave in a two-dimensional axisymmetric flexible tube, as described in [6, 7]. This tube is a simplified model for a large artery. It has an inner radius of 0.005 m, a length of 0.05 m and a thickness of 0.001 m. The fluid is incompressible and has a density of 1000 kg/m^3 and a viscosity of 0.003 Pas. The tube’s wall consists of a linear elastic material with density 1200 kg/m^3, Young’s modulus $3 \times 10^5$ N/m^2 and Poisson’s ratio 0.3. The structure is clamped in all directions at the inlet and outlet.

The finite volume flow solver Fluent 14.5 (Ansys Inc., Lebanon, NH, USA) uses second-order pressure interpolation, second-order upwind for the momentum and backward Euler for the time discretization. It solves the Navier-Stokes equations in arbitrary Lagrangian-Eulerian (ALE) formulation with the PISO scheme. The grid of the fluid domain is adapted to the displacement of the fluid-structure interface with a spring analogy. The finite element structural solver Abaqus 6.12 (Simulia Inc., Providence, RI, USA) uses implicit time integration of continuum elements with 8 nodes and reduced integration. The fluid grid consists of 100x10 cells, the structural grid of 50x5 elements.

Both the fluid and the structure are initially at rest. During the first $3 \times 10^{-3}$ s, an overpressure of $1333.2$ N/m^2 is applied at the inlet. The wave propagates through the tube during $10^{-2}$ s, simulated with time steps of $10^{-4}$ s.

3.2 Membrane pump

The second case is a two-dimensional axisymmetric model for a membrane pump [8]. This pump consists of a casing and a membrane with a hole at its middle to allow the fluid below the membrane to reach the outlet. The complete geometry is depicted in Figure 1. The radius and width of the casing are 46 mm and 11 mm, respectively. The membrane has an inner radius of 8 mm and an outer radius of 38 mm and a thickness of 1 mm. The outer edge of the membrane is attached to an electromagnetic actuator which moves it up and down.

The fluid is water with a density of $998.2$ kg/m^3 and a viscosity of $0.001003$ kg/ms. The hyperelastic material of the structure is described with the Van der Waals model. It has a density of $1160$ kg/m^3 and a stiffness between $1.4$ and $2.3$ N/mm^2.

The same flow solver and structural solver and the same discretization schemes as for the previous case have been applied. For this case with large deformation, cells in the fluid grid which are too large or too small are split or merged, respectively. The fluid grid initially consists of 8830 triangular cells, the structural grid of 40x4 elements.

Both the fluid and the structure are initially at rest. The actuator has an amplitude of 1 mm and a frequency of 50 Hz. The time step size is 0.5 ms.
Figure 1: The geometry of the membrane pump.

Table 1: The average number of coupling iterations per time step for the propagation of the pressure wave in the flexible tube.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-DN-IAC</td>
<td>4.0</td>
</tr>
<tr>
<td>IQN-ILS(5)</td>
<td>4.5</td>
</tr>
<tr>
<td>IQN-ILS(10)</td>
<td>3.6</td>
</tr>
</tbody>
</table>

4 RESULTS

The propagation of the pressure wave in the flexible tube has been calculated using the GS-DN-IAC algorithm [3] and the IQN-ILS algorithm [9]. Pressure contours at three instants are shown in Figure 2.

For each of these coupling algorithms, the number of coupling iterations in a time step is listed in Table 1, averaged over all time steps. The number between brackets behind IQN-ILS indicates from how many time steps data is reused by this algorithm. In each time step, the residual is reduced by three orders of magnitude with respect to its initial value in that time step. For this first case, the GS-DN-IAC algorithm performs similarly compared to the IQN-ILS algorithm. In addition to the 4 coupling iterations per time step, GS-DN-IAC requires two structural calculations to obtain the finite difference calculation of the compressibility factor.

Velocity vectors in the membrane pump are depicted in Figure 3. Despite various different attempts to calculate the compressibility coefficients for GS-DN-IAC, this scheme did not converge for this case. Using the IQN-ILS algorithm, on average 16 coupling iterations per time step are required if no data from previous time steps is reused.

For the flexible tube, the local structural model constructed by GS-DN-IAC and GS-RN is a good approximation of the structural behaviour. By contrast, the relation between
Figure 2: The pressure contours for the propagation of the pressure wave in the flexible tube.

Figure 3: The velocity vectors in the membrane pump after 0.125 s.
the pressure and the displacement in the membrane pump will not be local. For example, the displacement of the membrane depends on the pressure difference between both sides of the membrane and not on the pressure on one side. As the structural response to a local change in pressure is not local, it can be expected that it cannot be approximated by a local model.

5 CONCLUSIONS

In this paper, it has been demonstrated that the Robin-Neumann decomposition and the Dirichlet-Neumann decomposition combined with Interface Artificial Compressibility both apply the same principle. They include a local, linear approximation for the structural behaviour into the flow solver. These techniques accelerate the convergence of Gauss-Seidel coupling iterations for the propagation of a pressure wave in a flexible tube. Yet, their application to a membrane pump is not straightforward.

REFERENCES


MACRO-STEP-SIZE SELECTION AND MONITORING OF THE COUPLING ERROR FOR WEAK COUPLED SUBSYSTEMS IN THE FREQUENCY-DOMAIN

M. BENEDIKT*, D. WATZENIG*, J. ZEHEHNER* AND A. HOFER†

*Virtual Vehicle Research Center (ViF)
Area E - Electrics/Electronics and Software
Inffeldgasse 21A/II, 8010 Graz, Austria
e-mail: martin.benedikt@v2c2.at, web page: http://www.v2c2.at

†Institute of Automation and Control (IRT)
Graz University of Technology
Kopernikusgasse 24/II, 8010 Graz, Austria
e-mail: anton.hofer@tugraz.at, web page: http://www.irt.tugraz.at

Key words: Modular Simulation, Co-Simulation, Macro-Step-Size, Quality, ICOS

Abstract. A rather general approach to establish a multiphysic simulation is referred to as non-iterative co-simulation or weak coupled simulation. The involved subsystems are coupled in a weak sense and thus stepwise extrapolation of the coupling signals is required. Extrapolation is associated with an error which may influence the dynamical behavior of the coupled system. This coupling error depends significantly on the coupling step-size, i.e. the macro-step-size, and is one of the most critical parameter of a non-iterative co-simulation. In practice, appropriate macro-step-sizes are determined by some numerical tests or chosen according to the experience of domain-specific engineers. But who assesses the results of a non-iterative co-simulation? Comparison between reference and co-simulation results in the time-domain is often practiced in case studies using complex subsystems to validate co-simulation performance and accuracy of the achieved results. But this approach is counter-productive and mostly not applicable in practice. In this work we consider the coupling process as single source of distortion and analyse it in the frequency-domain. As a consequence, a relation between the macro-step-size and the coupling signals is available which leads to a ‘rule-of-thumb’ for an adequate macro-step-size selection. In addition, the gained insight into the coupling process itself enables new possibilities to monitor the coupling error leading to the ability to assess the results of a weak coupled simulation. The proposed methodologies are examined using a complex mechatronic system describing a vehicle (multi-body system, MBS), which is controlled via an anti-lock braking system (ABS) during different scenarios.
1 INTRODUCTION

Engineers are highly interested in multiphysic simulations because of the improved possibilities to design and/or to analyse a complex system. Typically, separated components of the overall system, i.e. subsystems, are treated by domain-specific teams of engineers to handle the resulting system complexity. For such an approach distributed modelling of the subsystems using specific modelling languages and distributed simulation using tailored solvers is mandatory. Advantageously, this approach leads to maximum flexibility for virtual system design. But nowadays, the performance of the overall system becomes more and more important and thus, the entire (dynamical) behavior of the overall system has to be considered [7]. During the development process performing of holistic simulations of the entire coupled system is required.

Since about two decades an increasing demand for holistic system simulation is recognisable and significant efforts in devising standards were done, for instance see MODELISAR [3]. On the one hand a monolithical simulation approach exists and on the other hand, coupled modular simulation of the subsystems is possible [2]. The idea behind a monolithical simulation approach is to import all subsystems - eventually including the accompanying numerical solver - into a single simulation tool, e.g. MATLAB/SIMULINK\(^1\) or DYMOLA\(^2\).

The main drawback is obvious: Not every used simulation tool exhibits model-export functionalities and thus, an additional (huge) effort for porting the submodel is required. Even though mandatory, specific numerical solvers might not be available in the used (master) simulator. By the latter mentioned approach the subsystems are modelled and solved by tailored simulation tools and the required interactions are established via connection of the corresponding inputs and outputs. This approach is also referred to as co-simulation.

Interfacing and external control of the involved domain-specific simulation tools is necessary and coupling data, i.e. relevant inputs and outputs, are exchanged at predefined points in time for synchronisation purposes [1, 4, 7]. As disadvantageous fact, limited interfacing capabilities of the used simulation tools avoid the application of highly accurate iterative (implicit) coupling schemes [2]. Mostly, the application of non-iterative (explicit) coupling schemes is required. By the non-iterative coupling approach the accuracy significantly depends on the chosen coupling step-size, the so called macro-step-size. Thus, the variable or fixed macro-step-size is the most critical parameter of a non-iterative co-simulation. In practice, for large systems appropriate macro-step-sizes are determined by numerical tests (‘trial & error’ approach) or chosen according to the experience of domain-specific engineers. Additionally, estimates on the numerical error are very pessimistic and these measures are not adequate for the practical assessment of simulation results [1]. Furthermore, the result of a non-iterative co-simulation is typically evaluated by comparison to the result of a standalone simulation, i.e. a simulation without coupling

\(^1\)http://www.mathworks.com
\(^2\)http://www.dymola.com
errors. But this approach is only applicable for simple coupled problems and mostly used in scientific co-simulation studies where the possibility to compute a solution for reference is given. In contrast, for complex (real) co-simulations it is not possible to perform a standalone simulation and thus, there is no possibility to assess results of a non-iterative co-simulation.

The error behavior and bounds on accuracy are important aspects for virtual-prototyping using non-iterative co-simulation. In this work we consider the non-iterative coupling scheme as single source of distortion. Instead of analysing the overall complex system, we focus on the coupling process itself. From considerations in the frequency-domain it is possible to determine the influence of the couplings on the coupling signals. Meaning, small distortions of the coupling quantities intuitively result in small numerical errors. Keeping this idea in mind, a reduction of coupling errors will lead to more accurate co-simulation results [4].

This article is devoted to show the capabilities of a frequency-based description of the coupling process and is organised as follows: In Section 2 a short outline of the basic principle of the non-iterative coupling scheme is given and the required coupling process is modelled by a so called coupling element. In Section 3, based on considerations of the coupling element in the frequency-domain, a 'rule-of-thumb' for a suitable selection of the macro-step-size is proposed. In Section 4, to assess the quality of the results of a non-iterative co-simulation, this approach is enhanced. Finally, in Section 5, the proposed methodologies are verified by an example describing a multi-body system (MBS), which is controlled via an anti-lock braking system (ABS) during different scenarios.

2 THE COUPLING SCHEME MODELLED AS COUPLING ELEMENT

Using co-simulation, subsystems are solved independently by their implemented domain-specific numerical solvers with problem-specific fixed or variable step-sizes, i.e. the so called micro-step-sizes $\delta T$. For synchronisation purposes the individual subsystems are paused at discrete time instants to perform a data exchange (weakly coupled). The time intervals between consecutive coupling time instants are referred to as macro-time-steps $\Delta T$. In particular, at coupling time instants input quantities of the subsystems are updated according to simulation results of the corresponding outputs of the connected subsystems [1, 7].

In the general case of bidirectional dependencies in between subsystems several coupling quantities are unknown which have to be estimated by extrapolation based on past coupling data. Thereby, an estimation error, i.e. a coupling error, is introduced. Typically, polynomial extrapolation techniques are applied which are based on past coupling quantities. Commonly used strategies are the zero-order-hold (ZOH), the first-order-hold (FOH) and the second-order-hold (SOH) extrapolation approaches [1, 5]. Because of the non-iterative coupling scheme, no iterations are performed at each macro-time-step which results in an explicit character of the coupling scheme. Assuming that the subsystems are adequately (accurate) solved by the implemented numerical solvers, the required cou-
Sequential coupling of two subsystems and artificially introduced coupling element yields as single source of distortion. From this abstract point of view, the progress of subsequent extrapolations of coupling quantities may be considered as an additional artificial introduced subsystem [4]. In this work, this subsystem is denoted as coupling element. Given a coupling signal \( y(t) \), a piece-wise extrapolated function \( \hat{y}(t) \) is generated by the coupling element. Figure 1 depicts this abstraction. Two bidirectional coupled subsystems are illustrated. Because of a predefined sequential scheduling of the subsystems - Subsystem 2 is solved previous to Subsystem 1 - piece-wise extrapolation of the unknown coupling quantities \( y(t) \) by \( \hat{y}(t) \) over the time interval of each macro-time-step \( \Delta T \) is mandatory. The unintentional effect of the required extrapolation are represented by the artificially introduced coupling element. For co-simulation commonly used polynomial extrapolation techniques are modelled and analysed in [5]. In this context, the main coupling problems are possible aliasing effects due to sampling of the coupling signal, an unintentional introduced time-shift and discontinuities at coupling time instants. However, a coupling element represents the introduced coupling error and influences the entire behavior of the overall coupled system. Especially for linear and time invariant systems the analysis of the system behavior in the frequency-domain is possible. As an advantageous fact, the commonly used polynomial extrapolation techniques (ZOH, FOH and SOH) represent LTI-systems and may be analysed in the frequency domain by a transfer function \( H(s) \):

\[
H(s) = \frac{\hat{y}(s)}{y(s)}
\]

For instance, the transfer function describing the behavior of the zero-order-hold extrapolation (ZOH) is written as

\[
H_{zoh}(s) = \frac{1 - e^{-s\Delta T}}{s\Delta T},
\]

where \( s \) denotes the LAPLACE variable. By this approach many different extrapolation schemes may by modelled and analysed in the frequency-domain. For the commonly used extrapolation techniques the corresponding transfer functions are outlined in [5]. The magnitude and the phase of the resulting frequency responses are illustrated in Figure 2 in dependence on the normed frequency \( \omega \Delta T \). Obviously, over the illustrated interval of the normed frequency the zero-order-hold extrapolation technique leads to a significant frequency dependent phase-shift. This effect may be intuitively interpreted as a
unintentional time-delay. Both higher order polynomial extrapolation techniques (FOH, SOH) exhibit a nearly ideal transfer behavior in the lower frequency range. In addition, an increasing normed frequency results in a significant modification of the amplitudes of the (harmonic) signal components of the coupling signal $y(t)$ and a frequency dependent phase-shift.

3 MACRO-STEP-SIZE SELECTION

Based on general considerations of the coupling elements in the frequency-domain a gained insight into the process of coupling is established. Concerning the magnitude and the phase characteristics of the modelled coupling elements, different coupling schemes may be assessed [5]. In fact, for ideal transfer behavior of the coupling element the following equation must be satisfied:

$$H(j\omega) = 1 + j0 \quad \forall \omega$$

This ideal characteristic is impossible because of time-delays in the transfer functions, for example see (2). To compensate the introduced coupling error a non-causal system would be required which is not realisable [6]. But, by permitting small coupling errors, efficient normed bandwidths for the modelled coupling elements may be defined. As
Table 1: Heuristically defined maximum normed efficient bandwidth

<table>
<thead>
<tr>
<th>Extrapolation</th>
<th>$\omega\Delta T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZOH</td>
<td>0.05</td>
</tr>
<tr>
<td>FOH</td>
<td>0.3</td>
</tr>
<tr>
<td>SOH</td>
<td>0.5</td>
</tr>
</tbody>
</table>

illustrated in Figure 2, the maximum normed bandwidths denoted by $\omega_{zoh}$, $\omega_{foh}$ and $\omega_{soh}$ are depicted for the three frequency responses of the ZOH, FOH and the SOH extrapolation scheme, respectively. The maximum normed bandwidths $\omega$ are chosen according to the characteristics of the coupling elements with the focus on a negligible distortion of the coupling signal. For instance, a maximum amplification of the coupling signal of $|H(j\omega\Delta T)| < 1.03$ and a maximum phase-shift of $|\angle H(j\omega\Delta T)| < 3^\circ$ seems to be adequate. Table 1 summarises the relations between the feasible efficient normed bandwidth of the coupling signal regarding the extrapolation techniques [5]. This table yields as ‘rule-of-thumb’ for macro-step-size selection and is determined heuristically based on considerations of the coupling elements in the frequency domain.

Of course, the definition of the efficient bandwidths also depends on the coupled system. Meaning, discontinuities at coupling time instants represent high frequency components in the coupling signal and may excite fast dynamics of the subsequent subsystem. Thus, the proposed ‘rule-of-thumb’ is valid for nonstiff coupled problems, if the coupling signals which correspond to the slow system dynamics are extrapolated.

4 VALIDATION OF THE COUPLING PERFORMANCE

Besides the selection of a suitable macro-step-size, also the quality of the coupling itself may be assessed [4]. Once again, this is enabled due to investigations of the modelled coupling elements in the frequency-domain. Based on the knowledge of the time varying maximum bandwidth of the coupling signals $\omega(t)$ a relation for the evaluation of the couplings of a non-iterative co-simulation can be defined:

$$\omega(t) < \frac{\omega\Delta T}{\Delta T(t)}$$

In this inequality the macro-step-size depends on time ($\Delta T(t)$). This is necessary in general due to the application of (adaptive) macro-step-size control algorithms. Obviously, to proof relation (4) the extraction of information on frequency components of the coupling signal is required. This means, that co-simulation results can be qualitatively evaluated if the actual maximum bandwidth of the coupling signal may be extracted and an usable bandwidth of the corresponding coupling element is defined. Furthermore, a topological (location of the coupling element in the network of subsystems) as well as a temporal assignment of eventually occurring coupling errors according to (4) is enabled.
5 AN ANTILOCK BRAKING SYSTEM EXAMPLE

To evaluate the proposed approaches a rather complex multibody system designed for vehicle dynamic simulations is examined. Typically, simulations are carried out for different real world scenarios and predefined driving cycles, to analyse the dynamic vehicle behavior or to optimise the setup of the system. To enhance the engineering and analysis possibilities, components out of various domains are embedded into the multibody vehicle simulation leading to a multidisciplinary application [1]. Often, additional subsystems are modelled in other domain specific simulation tools. A huge effort is necessary to translate the subsystem into an appropriate modelling language or to embed the resulting code (using a export-functionality) into the whole vehicle simulation, using tailored interfacing capabilities. During the development process a separate translation has to be performed for each adaptation of the additional systems, which leads to a time consuming procedure and rapidly increasing development costs. In contrast, co-simulation approaches enable simultaneous modifications of the involved subsystems besides establishing the connection between different simulation tools.

Figure 3: Co-simulation representation of the coupled subsystems
5.1 Description of the coupled system

For the chosen evaluation example co-simulation is used to enhance a multibody vehicle model implemented in ADAMS/Car\(^3\) by an antilock braking system (ABS), which is realised using MATLAB/SIMULINK. Additionally, a virtual driver and the predefined driving manoeuvre are also configured in ADAMS/Car. In particular, the virtual driver actuates the brake during a left-turn according to the lateral acceleration of the vehicle [4]. The co-simulation representation of the coupled system is sketched in Figure 3. The interactions in between the subsystems are established by the vehicle velocity \(v_{\text{vehicle}}(t)\), the wheel speeds \(\omega_{\text{wheel},r}(t)\), \(\omega_{\text{wheel},f}(t)\), \(\omega_{\text{wheel},r}(t)\), \(\omega_{\text{wheel},l}(t)\) and the control signals \(u_{\text{abs},r}(t)\), \(u_{\text{abs},f}(t)\), \(u_{\text{abs},r}(t)\), \(u_{\text{abs},l}(t)\). Thereby, the subscripted characters \((l, r, f, r)\) denote the topological position of the wheels mounted on the vehicle as \textit{left}, \textit{right}, \textit{front} and \textit{rear}, respectively. Both domain-specific subsystems (ABS-Controller, MBS-System) are coupled by the co-simulation platform ICOS\(^4\) (Independent Co-Simulation), which satisfies efficient interfacing of the used simulation tools and handles the exchange of coupling data during co-simulation [8].

Beside the co-simulations a standalone simulation (a monolithical simulation) was carried out using a single ADAMS/Car model for evaluation purposes. In this case, the modelled controller subsystem was embedded into the ADAMS/Car subsystem using special model-export (real-time workshop from MATLAB) and model-import (general state equation in ADAMS/Car) functionalities. Note, without the co-simulation approach, with every change of the ABS-Controller subsystem a model-export and model-import is required which results in a significant additional effort.

5.2 Non-iterative co-simulation

For this evaluation example three differently configured co-simulations are carried out. In all cases the subsystems are scheduled in sequential order. The multibody vehicle subsystem modelled in ADAMS/Car is solved previous to the ABS-Controller subsystem for each macro-time-step \(\Delta T\). This configuration requires some kind of piece-wise extrapolation of the inputs, i.e. the control signals, of the MBS-System, compare to Figure 1. For the three performed co-simulations the commonly used polynomial extrapolation techniques ZOH, FOH and SOH are used. Because of less information on the dynamics of the subsystems itself different constant macro-step-sizes are chosen for the extrapolation schemes. Further, according to the selected scenario (left-turn braking) only the control actions of the rear left wheel are illustrated in Figure 4.

Each plot compares the result of the monolithical simulation and the performed co-simulations using a specific extrapolation technique. The upper plot shows the application of the zero-order-hold (ZOH) extrapolation scheme. The results of the co-simulation

\(^3\)http://www.mscsoftware.com  
\(^4\)http://www.V2C2.at/icos
Figure 4: Co-simulation results according to different extrapolation schemes. Zero-order-hold (ZOH) with $\Delta T = 0.0075$ s, first-order-hold (FOH) with $\Delta T = 0.001$ s and second-order-hold (SOH) with $\Delta T = 0.01$ s.
Figure 5: Maximum occurring instantaneous frequency $\omega_{abs,hl}(t)$ of the coupling signal and depicted maximum bandwidths (horizontal lines) of the applied coupling elements

is computed using a constant macro-step-size of $\Delta T = 0.0075\,s$. Obviously, there are strong deviations between the depicted simulation results. Besides increased oscillations of the control signal also a time-shift is introduced. In contrast, the middle plot illustrates the results concerning first-order-hold (FOH) extrapolation. The macro-step-size is fixed to $\Delta T = 0.001\,s$. In this case, the results of the performed co-simulation are visually very similar to the result of the monolithical simulation. In the case of second-order-extrapolation (SOH, lower plot) with a macro-step-size of $\Delta T = 0.01\,s$, significant oscillations are eliminated in the time interval $t \in [1.5, 1.9]$. Eventually, these omitted oscillations may be important for the overall system design. Different configurations of a co-simulation may lead to significantly different simulation results, as shown in Figure 4. One may expect that the SOH extrapolation scheme intuitively may lead to the best result of the performed co-simulations for this example. But in any case both, the applied macro-step-size and the used extrapolation scheme have to be accounted for coupling performance aspects. As an important remark: up to now, for the engineer it is impossible to assess co-simulation results without knowing a reference (monolithical) solution. However, to mitigate this significant drawback, the proposed approaches based on considerations in the frequency-domain are applied.
5.3 Performance of the coupling elements

The proposed ‘rule-of-thumb’ (Tab. 1) provide relations between the macro-step-size and the usable bandwidth of the coupling signal for commonly used extrapolation schemes. Knowing the macro-step-size $\Delta T$ the efficient bandwidths of the coupling elements can be determined using (4). The resulting efficient bandwidths of the three coupling elements are illustrated in Figure 5. Because of the chosen constant macro-step-sizes all efficient bandwidths of the applied couplings are constant over the simulation time. Additionally, in Figure 5 the maximum instantaneous frequency $\omega_{\text{abs,rt}}(t)$ of the control signal $u_{\text{abs,rt}}$ is depicted. Obviously, the efficient bandwidths of the three coupling elements are significantly different. In particular, the ZOH scheme provides the smallest efficient bandwidth and many frequency components of the coupling signal are strongly modified. For the SOH scheme the efficient bandwidth is slightly enlarged, but also frequency components are distorted by the coupling process. Thus, by the application of ZOH ($\Delta T = 0.0075 \text{s}$) and SOH ($\Delta T = 0.01 \text{s}$) extrapolation significant distortions are introduced due to the required couplings. In contrast, the application of the first-order-hold extrapolation technique with a macro-step-size of $\Delta T = 0.001 \text{s}$ leads to the largest efficient bandwidth. Only marginal frequency components are influenced and thus, accurate results of the co-simulation can be expected. As shown in Figure 4, only this coupling approach leads to meaningful simulation results. Thus, without knowing a reference solution, based on considerations in the frequency-domain the performance of the coupling elements according to the corresponding coupling signals are qualitatively assessable. The other way around, if the maximum occurring bandwidth of the coupling signal is a-priori known, a suitable macro-step-size can be defined based on Table 1. This macro-step-size ensures a sufficient bandwidth of the coupling element.

6 CONCLUSIONS

The non-iterative co-simulation approach requires marginal interfacing capabilities of the involved simulation tools. Thus, this coupling scheme is most often applicable. As a significant drawback, the required couplings introduce coupling errors. Coupling errors strongly depend on the macro-step-size and the used scheme for extrapolation of the coupling quantities. Concerning accuracy, bounds on the coupling error based on the approximation order of the used extrapolation technique are very pessimistic and therefore not applicable.

In our approach, we are focusing on the couplings itself instead of analysing the overall coupled system. From the abstract point of view the coupling scheme is considered as an additional subsystem, i.e. a coupling element. The analysis of the coupling element in the frequency-domain leads to a ‘rule-of-thumb’ to select the most critical parameter of a non-iterative co-simulation, the macro-step-size. Based on the resulting relations the assessment of the performance of the coupling elements is possible. In fact, results of a non-iterative co-simulation can be qualitatively assessed without a reference solution.
7 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the "COMET K2 - Competence Centres for Excellent Technologies Programme" of the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT), the Austrian Federal Ministry of Economy, Family and Youth (BMWFJ), the Austrian Research Promotion Agency (FFG), the Province of Styria and the Styrian Business Promotion Agency (SFG).
We would furthermore like to express our thanks to our scientific project partner, namely to the Institute of Automation and Control (IRT) at the Graz University of Technology.

REFERENCES


NEPCE - A NEARLY ENERGY-PRESERVING COUPLING ELEMENT FOR WEAK-COUPLED PROBLEMS AND CO-SIMULATIONS

M. BENEDIKT∗†, D. WATZENIG∗, J. ZEHETNER∗ AND A. HOFER†

∗Virtual Vehicle Research Center (ViF)
Area E - Electrics/Electronics and Software
Inffeldgasse 21A/II, 8010 Graz, Austria
e-mail: martin.benedikt@v2c2.at, web page: http://www.v2c2.at

†Institute of Automation and Control (IRT)
Graz University of Technology
Kopernikusgasse 24/II, 8010 Graz, Austria
e-mail: anton.hofer@tugraz.at, web page: http://www.irt.tugraz.at

Key words: Modular Simulation, Co-Simulation, Weak Coupling, NEPCE, ICOS

Abstract. Application of the weak-coupled or non-iterative co-simulation approach is very popular in industry because of its simple synchronisation scheme of the involved subsystems, even though there are some significant drawbacks. The main problems are obvious: On the one hand the explicit character of this coupling strategy may lead to instability of the overall system solution and on the other hand the selection of an appropriate coupling step-size is challenging. Investigation of the weak-coupling effects is necessary to improve the performance as well as the usability of this coupling approach. Concerning non-iterative co-simulation, subsystems are solved once over coupling time steps and in the general case of bidirectional dependencies between subsystems, several coupling quantities are unknown. This requires some kind of extrapolation, leading to an explicit character of the weak coupling approach. If small coupling step sizes have to be applied, this strategy leads to a significant increase of overall simulation time.

In this work we analyse the coupling process as a standalone subsystem and on this basis we present a novel coupling strategy. From an energy flow point of view this strategy considers the transmission of generalised energy between coupled subsystems and ensures (nearly) energy conservation in the coupled overall system. Thus, this modified coupling-process is referred to as nearly energy-preserving coupling element (NEPCE). Besides the fundamental theory some examples demonstrate the improved performance of the weak coupling scheme leading to more accurate simulation results and to a reduced amount of simulation time as well.
1 INTRODUCTION

In the engineering and development process of large-scaled dynamic systems virtual prototyping is mandatory to avoid expensive re-designs in late design phases. In addition, often cross-domain (multiphysic) considerations are necessary to improve the system performance and to satisfy increasing customer needs. To handle the resulting complexity the system under investigation is typically splitted into subsystems [1, 5]. Those domain-specific system parts are analysed by domain specific engineers. Therefore, well-established simulation tools targeting specific requirements are used to cover the physical behavior of the separated subsystems. Each subsystem is modelled within a specific domain by an individual model description language and solved using a tailored solver. In consequence, to estimate the entire overall system behavior, a holistic system analysis is required. Besides the independent simulation of the separated subsystems also the interaction between the individual subsystems have to be considered [2].

In fact, there are two possibilities: Integration of all subsystem models into an overall system model using a single simulation tool or integration of the distributed modelled as well as simulated subsystems via a co-simulation approach [1]. The first mentioned approach may not be applicable, due to the lack of available model-export or -import functionalities, the lack of domain-specific modelling language characteristics or required numerical solvers. In contrast, the mentioned co-simulation approach seems to be suitable for a modular holistic system simulation. The subsystems are modelled using domain-specific modelling languages and are solved by tailored numerical solvers (multi-method) with individual problem-specific step-sizes (multi-rate). But as a disadvantageous fact, the couplings in between interacting subsystems have to be established without influencing the solution of the overall system behavior.

Concerning co-simulation, subsystems are solved independently over pre-defined time-intervals, the so called macro-time-steps, by the implemented numerical solvers. In the general case of bidirectional interacting subsystems at least two subsystems depend on each other. To solve co-simulation network internal loops iterative or non-iterative coupling schemes exist [5]. As shown in [1, 5], well-established iterative (implicit) coupling methodologies can solve network internal loops in an efficient way. Unfortunately they are not applicable to all problems, because of limited interfacing capabilities of the used simulation tools. For example, at each iteration, resetting of the involved subsystems as well as a reconfiguration of the simulation tools is mandatory. Furthermore, for fast simulations and for (soft) real-time applications iterative methodologies are not well suited because of the non-deterministic timing behavior. In contrast, by using the non-iterative (explicit) strategy the involved simulation tools are paused to perform a data exchange. Thus, this coupling methodology is applicable in general for co-simulation applications. As a disadvantageous fact, the non-iterative strategy requires extrapolation of unknown coupling quantities and thereby some kind of distortion, i.e. a coupling error, is introduced [2].
This coupling error depends strongly on the macro-step-size, smaller macro-step-sizes imply smaller errors, but they are not applicable in general [1, 5]. Thus, to mitigate the unintentional introduced distortion, a modification of the chosen extrapolation technique is necessary. Besides overall energy consistency further requirements on a compensation scheme are: Minor computational effort for fast applications and flexible adaptivity for on-line adjustments due to an adaptive control of the macro-step-size, resulting in an efficient easy-to-use coupling strategy for non-iterative heterogeneous co-simulation.

This article is organized as follows: In the Section 2 the non-iterative principle is explained. Based on a commonly used coupling scheme the occurring main problems are extracted. Therefore, a so called nearly energy-preserving coupling element (NEPCE) is introduced in Section 3. From an energy flow point of view a strategy to mitigate inherent coupling errors is proposed. Section 4 shows how the developed compensation scheme works. A realistic co-simulation example describing a hybrid electric vehicle is used to demonstrate the efficiency of the improved coupling element.

2 AN ARTIFICIAL SUBSYSTEM - THE COUPLING ELEMENT

In general, the subsystems of a co-simulation are solved independently for each macro-time-step. The subsystems are solved using the chosen domain-specific numerical solvers with problem-specific step-sizes. In the context of co-simulation, the steps of the numerical solvers are referred to as micro-time-steps $\delta T$. Therefore, in the general case of variable step size solvers, the $L$-th numerical evaluation of a subsystem is performed at the time instant $t^{<L>}$

$$t^{<L>} = \sum_{l=1}^{L} \delta T^{<l>},$$

where $\delta T^{<l>}$ denotes the size of the $l$-th micro-time-step. Simultaneously, the same notation can be introduced to express the $M$-th coupling time instant $t^{<M>}$:

$$t^{<M>} = \sum_{m=1}^{M} \Delta T^{<m>}$$

The $m$-th macro-time-step is denoted by $\Delta T^{<m>}$ [2]. Subsystems are independently solved for each macro-time-step and at each coupling time instant coupling data is exchanged. If subsystems are interconnected in a closed loop (internal loop) this approach is not directly applicable, because of unknown coupling quantities. In fact, if bidirectional dependencies in between subsystems exist several inputs of the subsystems for the computation of the actual macro-time-step are unknown. For the non-iterative coupling scheme the unknown coupling quantities are extrapolated based on past coupling information [3, 5]. Synchronisation of the subsystems are only performed at the coupling time instants which leads to a weak coupling of subsystems. Thus, this coupling approach is referred to as weak coupling scheme.
To demonstrate the fundamental problem two interacting subsystems are sketched in Figure 1. Both subsystems directly depend on each other and thus, extrapolation of the required inputs (coupling quantities) is necessary. If Subsystem 2 is solved previous to Subsystem 1 - sequential scheduled - the input signal \( y(t) \) of Subsystem 2 has to be estimated by \( \hat{y}(t) \) through an extrapolation for the time interval of each macro-time-step \( t \in (t^{<M-1>}_\Delta, t^{<M>}\Delta] \), see Figure 2. The required extrapolation process influences the entire system behavior and can be considered as an additional artificially introduced subsystem.

In this work the extrapolation process is referred to as coupling element.

As shown in Figure 2, based on past coupling data \( y(t^{<L>}_\Delta) \) unknown coupling signals \( \hat{y}(t) \) are extrapolated. Therefore, interpretation of the coupling data \( y(t^{<L>}_\Delta) \) as discrete time function values of a continuous coupling signal \( y(t) \) enables flexible handling of different step-sizes. Macro-step-sizes \( \Delta T^{<m>} \) are completely decoupled according to fixed or variable micro-step-sizes \( \delta T^{<l>} \) of the underlying numerical solvers. For extrapolation purposes sampling of the continuous sampling signal at coupling time instants is required and thus aliasing-effects may be introduced [3].
2.1 Sample & Hold

Typically, in co-simulation applications extrapolation is performed signal-based via polynomial extrapolation techniques of low order [1]. The most commonly used extrapolation approach is the zero-order extrapolation scheme: Unknown inputs are treated as constant quantities over the interval of the actual constant macro-time-step $\Delta T$ equal the value of the coupling signal $y(t_{\Delta}^{<M>})$ at the actual coupling time instant $t_{\Delta}^{<M>}$: 

$$\hat{y}(t) = y(t_{\Delta}^{<M>})$$

with $t_{\Delta}^{<M>} \leq t < t_{\Delta}^{<M+1>}$ 

(3)

This extrapolation scheme results in a continuous time signal $\hat{y}(t)$ describing a piece-wise constant function. Utilising the LAPLACE transformation and appropriate scaling regarding required sampling of the coupling signal leads to a transfer function $H(s)$

$$H(s) = \frac{\hat{y}(s)}{y(s)} = \frac{1 - e^{-s\Delta T}}{s\Delta T},$$

(4)

which describes the behavior of the coupling element in the special case of zero-order extrapolation [3]. After substitution of the LAPLACE variable $s = j\omega$ and simple trigonometric transformations the frequency response of the coupling element may be written as follows:

$$H(j\omega) = \frac{2\pi \sin (\omega\Delta T/2)}{\pi\omega} e^{-j\omega \frac{\Delta T}{2}} = \overbrace{\text{I} \left(\omega\right)}^{\text{real}} e^{-j\omega \frac{\Delta T}{2}}$$

(5)

As expected and represented by Equation (5), due to the zero-order extrapolation scheme a time-delay is introduced. In Figure 3, the effect of the simple coupling scheme is sketched. Given a continuous time coupling signal $y(t)$ a piece-wise constant function $\hat{y}(t)$ is generated by the coupling element. Obviously, discontinuities at coupling time instants occur and they represent high frequency components of the coupling signal. These

---

1For the consecutive part of this article exclusively constant macro-step-sizes $\Delta T = \Delta T^{<m>} \forall m$ are considered for analysis.
discontinuities may influence the accuracy of the numerical solvers and may excite existing fast dynamics of the subsequent subsystem [3]. Assuming, that discontinuities do not influence the numerical accuracy and high frequency components in the coupling signal are suppressed by the subsystem, only a time-delayed coupling signal is relevant, see Figure 3.

This sample & hold extrapolation scheme limits the macro-step-size significantly when subsystems with high dynamics have to be considered and high accuracy is required. Especially in closed loop configurations, e. q. control systems, the artificially introduced time-delay influences significantly the dynamics of the coupled system and may lead to an unstable system behavior. As a consequence, small step-sizes have to be chosen resulting in a rapidly increasing overall simulation time.

3 A NEARLY ENERGY-PRESERVING COUPLING ELEMENT

The so called weak coupling scheme introduces an inherent coupling error due to the required extrapolation of unknown coupling quantities. Extrapolation is directly associated with an estimation error. From an energy flow consideration point of view, with each macro-time-step a generalised energy is produced or dissipated during modular simulation [2]. In this work generalised energy is considered, because in co-simulation applications two coupling signals \( u(t) \) and \( y(t) \) may not represent power in the physical sense. Without coupling - performing a monolithical simulation - in the time interval of the \( m \)-th macro-time-step the generalised energy \( E_{g,y}^{<m>} \) is transmitted:

\[
E_{g,y}^{<m>} = \int_{\Delta}^{t_{\Delta}^{M-1}} u(t)y(t)dt.
\]  

As mentioned in Section 2, if bidirectional dependencies in between subsystems exist (Fig. 1), extrapolation of coupling signals is necessary. In this case the energy

\[
E_{g,\hat{y}}^{<m>} = \int_{\Delta}^{t_{\Delta}^{M-1}} u(t)\hat{y}(t)dt
\]

is transmitted, which leads to a deviation in the balance of transferred generalised energy:

\[
E_{g,y}^{<m>} \neq E_{g,\hat{y}}^{<m>}
\]

In the terminology of numerical simulation, the introduced coupling error represents a local discretisation error and is defined as

\[
\epsilon(t) = y(t) - \hat{y}(t).
\]

On the basis of (6), (7) and (9) the produced or dissipated amount of generalised energy \( E_{g,\epsilon}^{<m>} \) in the time interval of the \( m \)-th macro-time-step can be determined:

\[
E_{g,\epsilon}^{<m>} = E_{g,y}^{<m>} - E_{g,\hat{y}}^{<m>}
\]
In contrast to the classical approach of adapting the step size of the numerical solver to meet predefined bounds on accuracy, this approach is not possible for non-iterative co-simulation. For such an approach resetting of the involved simulation tools is mandatory and not possible in general. Therefore, to compensate the coupling error, the extrapolated coupling signal \( \hat{y}(t) \) could be modified by a correction signal \( c(t) \) based on past coupling errors:

\[
\hat{y}_c(t) = \hat{y}(t) + c(t)
\]

Furthermore, to establish an energy-preserving (ideal) coupling the correction signal \( c(t) \) has to satisfy the following relation:

\[
E_{g,c}^{<m>} = \int_{t_{\Delta - 1}}^{t_{\Delta}} u(t)c(t)dt \overset{!}{=} E_{g,c}^{<m>}
\]

However, due to the non-iterative coupling scheme an exact compensation of the coupling error is not possible. As shown in Section 2, the coupling element introduces some kind of time-delay and for an ideal compensation a non-causal system - i.e. future coupling errors have to be known a priori - would be required. Because of this aspect it is not possible to realise an ideal correction signal. But the undisputed goal is to approximate an ideal compensation (12) and therefore a so called compensation system is introduced, see Figure 4. In the general case, this compensation system implements a control system with the aim of tracking the coupling error:

\[
c(t) \overset{!}{=} c(t)
\]

Based on past coupling errors (9) the compensation system determines a correction signal and the extrapolated coupling signal is modified (11). In fact, the compensation scheme leads to a coupling-error-based modification of the used extrapolation technique. This approach is rather general and applicable for different extrapolation schemes. Thus, for
the class of nearly energy-preserving coupling elements on the basis of the proposed and fundamental considerations regarding generalised energy transmission, the abbreviation NEPCE is introduced [2].

3.1 Realisation of the correction signal

The proposed, nearly energy-preserving coupling scheme (NEPCE) modifies the extrapolated coupling signal by the use of a correction signal, which is determined based on past coupling data. Therefore, perhaps the most simplest choice for correction is the time-delayed coupling error signal (9):

$$c(t) = \epsilon(t - \Delta T^m)$$

Under the assumption of slowly varying coupling signals the coupling error of the previous step approximates the actual coupling error and therefore, this compensation approach is applicable to mitigate unintentional introduced distortions. But there are some problems concerning the realisation of the correction signal. First, in the case of time varying macro-step-sizes $\Delta T^m \neq \Delta T^{m+1}$ a variable time-delay has to be realised. In this case, it is difficult to interpret the resulting behavior of the compensation scheme. As a second aspect, the continuous time correction signal must be represented in form of discrete function values, i.e. a quasi-continuous approximation of the continuous time signal is necessary. In both cases many samples are required to represent the correction signal. Simulation tools may not be able to handle the increased amount of coupling data at coupling time instants because of limited interfacing capabilities. A possible simplification is the consideration of the integrated coupling error over the time interval of the last macro-time-step. Under the already mentioned assumption\(^2\) of slow varying coupling signals a piece-wise constant correction signal may be used to compensate the coupling error, which is defined by

$$c(t) := \frac{\alpha}{\Delta T^{m+1}} \int_{t^{m+1}}^{t^M} \epsilon(t)dt \quad \text{with} \quad t \in (t^M, t^{M+1}]$$

This compensation scheme is illustrated in Figure 5. The coupling error is integrated and used to realise a piece-wise constant correction signal. Instead of resetting the integrator at

---

\(^2\)By the use of sufficient small macro-step-sizes, this assumption is fulfilled for every continuous time coupling signal in general.
each coupling time instant a feedback path is introduced. Stability of the resulting control system as well as the performance of the proposed compensation system is adjusted by the tuning factor $\alpha \in [0, 2]$.

4 INDUSTRIAL EXAMPLE

For evaluation purposes a coupled system describing a hybrid electric vehicle (HEV) is considered. A hybrid electric vehicle simulation example is well suited for co-simulation because different engineering domains (electrical, mechanical, thermal) are involved. Typically, each subsystem is analysed using a specific simulation tool. For this example, a hybrid drive-train (AVL CRUISE, $\delta T = 0.001 \, s$) is coupled with a cooling subsystem (ECS KULI, $\delta T = 1.0 \, s$) and a third subsystem, which implements a virtual driver, a battery system for energy storage and a hybrid control unit (MATLAB, $\delta T = 0.01 \, s$). The co-simulation representation of the coupled system is sketched in Figure 6. Beside the use of problem-specific micro-step-sizes $\delta T$ coupling of the three subsystems is performed using the co-simulation platform ICOS [6]. In particular, to establish the interactions in between the subsystems unknown coupling quantities are estimated by the zero-order-hold (ZOH) extrapolation scheme.

Three different non-iterative co-simulations were carried out and the results are shown in Figure 7. Because of a predefined sequential scheduling of the involved subsystems (the vehicle is solved first) the inputs to the vehicle are estimated. This leads to piece-wise constant signals of the acceleration pedal (upper plot) and a required hybrid torque (middle plot). In the lower plot, as not estimated output signal the vehicle velocity is illustrated. For this example it is not possible to determine a monolithic (standalone) simulation and thus, the results of the co-simulation according to $\Delta T = 0.01 \, s$ are considered as a good numerical approximation of the exact solution. However, two co-simulations are performed with a macro-step-size of $\Delta T = 0.04 \, s$ with and without application of the proposed compensation scheme. As expected, a significant coupling error is introduced.
without compensation. Due to the bidirectional dependencies in between the subsystems a strong deviation of the entire system behavior is recognisable. Especially the deviation in the velocity of the vehicle lead to a different behavior of the implemented virtual driver (acceleration pedal and brake). In contrast, with the application of the proposed nearly energy-preserving coupling scheme (NEPCE) the simulation results are very close to the approximation of the exact solution ($\alpha = 1.5$).

Using the discussed coupling signals several definitions of generalised energy are possible. The results are demonstrated in Figure 8. The upper plot depicts the generalised energy $E_{g,\text{pedal}}(t)$ based on the vehicle velocity and the estimated acceleration pedal signal. In the lower plot, the generalised energy $E_{g,\text{torque}}(t)$ according to the vehicle velocity and the

![Figure 7: Co-simulation results for different zero-order-hold (ZOH) extrapolation schemes. Demonstration of the proposed correction scheme NEPCE on co-simulation results.](image_url)
Figure 8: Generalised energy according to specific coupling signals

estimated hybrid torque signal is illustrated. Obviously, for both interpretations, without a compensation strategy generalised energy is lost or added due to the coupling process. This leads to significant deviations and a strongly differing numerical solution, compare to Figure 7. However, the proposed coupling scheme (NEPCE) compensates coupling errors and leads to highly accurate simulation results, even though non-iterative co-simulation.

5 CONCLUSIONS

The modular treatment of subsystems in terms of modelling and simulation enables the engineering of complex multiphysic systems and yields flexibility during the product development process. As a drawback, bidirectional interaction between interconnected subsystems cause some problems and thus, an unintentional coupling error is introduced. Modelling of the coupling process by a coupling element leads to an additional (artificial) subsystem. Beside possible aliasing effects and discontinuities (high frequency components) at coupling time instants a time-shift of the coupling signal occurs intuitively. For exact compensation a non-causal system is mandatory, which is not realisable in general. Thus, in this work we present a novel strategy to compensate immanent coupling errors. From an energy flow consideration point of view between interacting subsystems a (nearly) energy-preserving coupling element (NEPCE) is proposed. The improvement
concerning the classical coupling element is demonstrated by an industrial co-simulation example. The accuracy of the co-simulation results is significantly increased.

Remark: The presented theory describes a novel approach of error compensation for non-iterative co-simulation and weak coupled problems. Protected by a pending European patent (EP 2 442 248 A1) since 2010 this approach is implemented within the co-simulation platform ICOS "Independent Co-Simulation" [6], which is developed at the Virtual Vehicle Research Center.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the "COMET K2 - Competence Centres for Excellent Technologies Programme" of the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT), the Austrian Federal Ministry of Economy, Family and Youth (BMWFJ), the Austrian Research Promotion Agency (FFG), the Province of Styria and the Styrian Business Promotion Agency (SFG).

We would furthermore like to express our thanks to our scientific project partner, namely to the Institute of Automation and Control (IRT) at the Graz University of Technology.

REFERENCES


2D VELOCITY-VORTICITY VISCOUS INCOMPRESSIBLE FLOWS

Raúl Téllez, Habersheel Acevedo, and Alfredo Nicolás
Departamento de Matemáticas, UAM–Iztapalapa, 09340 Iztapalapa, México City, México.
e-mail: anc@xanum.uam.mx

Key words: Velocity–vorticity formulation, direct method, moderate Reynolds numbers, steady state flows

Abstract. The 2D unsteady Navier-Stokes equations in its velocity-vorticity formulation, after time discretization, leads to a nonlinear elliptic system which may be solved by iterative methods; however, decoupling the nonlinear vorticity equation from the vorticity in the velocity equations by linear interpolation, a direct method is applied, allowing the vorticity equation to be linear. Steady state convergent flows from the un–regularized unit driven cavity problem, which causes recirculation because of the nonzero boundary condition on the top wall, are reported for Reynolds numbers $Re$, $400 \leq Re \leq 4000$. 

1033
1 INTRODUCTION

The main goal of this paper is to present numerical results at moderate Reynolds numbers in the range of $400 \leq Re \leq 4000$. The results are obtained using a simple numerical scheme for the unsteady Navier-Stokes equations in velocity and vorticity variables. The numerical scheme consists in a direct method to solve the nonlinear steady subproblem that results after a convenient time discretization is applied. This is a novelty contribution since usually that kind of problems are solved by iterative techniques: for instance, [1] for isothermal problems and [2] for thermal ones.

The flows are obtained from the well known un-regularized driven (or lid-driven) cavity problem which originates recirculation phenomena due to the nonzero velocity boundary condition on the top wall: the recirculation is originated by the fluid flow coming from the upstream top corner, and then hitting the downstream top corner.

At moderate Reynolds numbers, say for instance $Re \leq 7500$, the flow approaches to an asymptotic steady state as $t$ tends to $\infty$. The meshes in this work follow the size dictated by the thickness of the boundary layer (of order of $Re^{-\frac{1}{2}}$) and no refining on the mesh is required near the boundary. The results clearly show that as the Reynolds number increases the mesh has to be refined and this in turn leads to decrease the time step: numerically, by stability matters and physically, to capture the fast dynamics of the flow. At this stage the results that are shown are obtained with the contour values given by [3] which are easier to obtain than those given by [4].

2 THE CONTINUOUS PROBLEM AND THE NUMERICAL METHOD

Let $\Omega \subset R^N$ (N=2,3) be the region of the flow of a viscous incompressible fluid, and $\Gamma$ its boundary. It is well known that this kind of unsteady flow is governed by the non-dimensional Navier-Stokes equations; in $\Omega$, $t > 0$; given by

$$u_t - \frac{1}{Re} \Delta u + \nabla p + (u \cdot \nabla)u = f$$

$$\nabla \cdot u = 0$$

where $u$ and $p$ are the velocity and pressure of the flow, respectively. The parameter $Re = UL/\nu$, $\nu$=kinematic viscosity, is the Reynolds number. The momentum equation (1) must be supplemented with appropriate initial and boundary conditions, for instance $u(x,0) = u_0(x)$ in $\Omega$ ($\nabla \cdot u_0 = 0$) and $u = f_1$ on $\Gamma$, $t \geq 0$ ($\int f_1 \cdot n d\Gamma = 0$) respectively.

Taking twice the curl in the momentum equation in the primitive variables formulation of the Navier-Stokes equations given by (1), and restricting to the two-dimensional case, we obtain the velocity-vorticity formulation:

$$\omega_t - \frac{1}{Re} \nabla^2 \omega + u \cdot \nabla \omega = f,$$
for the vorticity $\omega$ and two Poisson equations for the velocity components

$$\nabla^2 u = -\frac{\partial \omega}{\partial y} \quad \text{and} \quad \nabla^2 v = \frac{\partial \omega}{\partial x},$$

(4)

which are related, with (1), by

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y};$$

(5)

where $\mathbf{u} = (u, v)$ is the velocity with components $u$ and $v$, see [1] for details for the general 3D case.

From equation (2) a function, called the streamfunction, is obtained, which using the relations

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x},$$

(6)

and, from (5), an elliptic equation for $\psi$ is obtained:

$$\nabla^2 \psi = -\omega.$$

(7)

A description of the numerical method follows. The time derivative that appears in the vorticity equation (3) is approximated with the following scheme, which is unconditionally stable when it is implicitly combined with the Laplacian operator, and it is well behaved when $t \to +\infty$, [5],

$$f_t(x, (n + 1) \Delta t) = \frac{3f^{n+1} - 4f^n + f^{n-1}}{2\Delta t} + \mathcal{O}(\Delta t^2), \quad n \geq 1,$$

(8)

where $f^n = f(x, n\Delta t)$; then, replacing that derivative in (3) by (8) we have the following totally implicit approximate problem

$$-\nabla^2 u^{n+1} = \frac{\partial \omega^{n+1}}{\partial y},$$

(9)

$$-\nabla^2 v^{n+1} = -\frac{\partial \omega^{n+1}}{\partial x}, \quad \mathbf{u}^{n+1}|_{\Gamma} = \mathbf{u}_{bc}^{n+1}$$

(10)

$$\alpha \omega^{n+1} - \nu \nabla^2 \omega^{n+1} + \mathbf{u}^{n+1} \cdot \nabla \omega^{n+1} = f_\omega, \quad \omega^{n+1}|_{\Gamma} = \omega_{bc}^{n+1}$$

(11)

where $\alpha = \frac{3}{2\Delta t}$, $f_\omega = \frac{4\omega^n - \omega^{n-1}}{2\Delta t}$ and $\frac{1}{Re}$ has been replaced by the kinematic viscosity coefficient $\nu$, considering $U = L = 1$. On the other hand, $\mathbf{u}_{bc}$ and $\omega_{bc}$ denote boundary condition for $\mathbf{u}$ and $\omega$.

We recall that the novelty here is to apply a direct method decoupling equations (9-10) from (11) with a linear interpolation from the values of $\omega$ known from the two previous time levels; then, in these conditions (9-10) are solved first, then with the values of $u$ and $v$, (11) is solved as a linear equation.
Renaming \( \{u^{n+1}, v^{n+1}, \omega^{n+1}\} \) as \( \{u, v, \omega\} \), and holding the linear interpolation of the right hand sides of \( u \) and \( v \) mentioned above, one has the following linear, uncoupled elliptic equations

\[
\begin{align*}
-\nabla^2 u &= 2 \frac{\partial \omega^n}{\partial y} - \frac{\partial \omega^{n-1}}{\partial y}, \\
-\nabla^2 v &= -2 \frac{\partial \omega^n}{\partial x} + \frac{\partial \omega^{n-1}}{\partial x}, \\
\alpha \omega - \nu \nabla^2 \omega + \mathbf{u} \cdot \nabla \omega &= f_\omega,
\end{align*}
\]

with the boundary condition mentioned in (9-11). To obtain \( (u^1, v^1, \omega^1) \) in (9-11) or (12-14), a first order approximation is applied for the time derivatives in (3), among other options, through a subsequence with smaller time step; a elliptic problem of the form (12-14) is also obtained.

For the spatial discretization to solve elliptic problems of the form (12-14) it may be applied either Finite Differences (FD) or Finite Element (FE) if the region \( \Omega \) is rectangular. For the (FD) case the following classic second order discretizations are used:

\[
\begin{align*}
\frac{d^2 f}{dx^2} &= \frac{f(x - h_x) - 2f(x) + f(x + h_x)}{h_x^2} + O(h_x^2), \\
\frac{d^2 f}{dy^2} &= \frac{f(y - h_y) - 2f(y) + f(y + h_y)}{h_y^2} + O(h_y^2), \\
\frac{df}{dx} &= \frac{f(x + h_x) - f(x - h_x)}{2h_x} + O(h_x^2), \\
\frac{df}{dy} &= \frac{f(y + h_y) - f(y - h_y)}{2h_y} + O(h_y^2).
\end{align*}
\]

**Remark 2.** We can also approximate \( \frac{df}{dx} \) and \( \frac{df}{dy} \) with (7) replacing \( \Delta t \) by \( h_x \) and \( h_y \).

Considering the rectangular flow region \( \Omega = (0, a) \times (0, b), \ a > 0, \ b > 0, \) which for the unitary cavity, \( a = b = 1. \) With \( M, N > 0 \) integers, we denote \( h_x = \frac{a}{M+1}, \ h_y = \frac{b}{N+1}, \) \( x_i = i \times h_x, \ i = 0, 1, \cdots, M, M + 1, \) and \( x_j = j \times h_y, \ j = 0, 1, \cdots, N, N + 1; \) and by, say \( \omega_{ij} = \omega(x_i, y_j). \) If \( M = N, \ h_x = h_y = h. \)

Then, with \( M = N, \)

\[
\frac{\partial \omega}{\partial y} = \frac{\omega_{ij+1} - \omega_{ij-1}}{2h}
\]

and
Raúl Téllez, Habersheel Acevedo, and Alfredo Nicolás

\[
\frac{\partial \omega}{\partial x} = \frac{\omega_{i+1} - \omega_{i-1}}{2h}; \quad (17)
\]

for \( \alpha \omega - \nu \nabla^2 \omega \) we have

\[
\alpha \omega_{ij}^{n+1} - \nu \omega_{i-1j}^{n+1} + \omega_{i+1j}^{n+1} - 4\omega_{ij}^{n+1} + \omega_{ij-1}^{n+1} + \omega_{ij+1}^{n+1} \quad \frac{2h^2}{h^2}; \quad (18)
\]

Similarly for \(-\nabla^2 u\) and \(-\nabla^2 v\).

Obtaining, from (18), a constant matrix \( m \) for \( \omega \). If we keep \( u \cdot \nabla \omega \) in the left hand side, discretizing it like

\[
u \omega_{ij}^{n+1} - \nabla^2 \omega = u \cdot (\nabla \omega) = u_{ij}^{n+1} \left[ \frac{\omega_{i+1j}^{n+1} - \omega_{i-1j}^{n+1}}{2h} \right] - v_{ij}^{n+1} \left[ \frac{\omega_{ij+1}^{n+1} - \omega_{ij-1}^{n+1}}{2h} \right] \quad (19)
\]

and adding it to \( \alpha \omega - \nu \nabla^2 \omega \) a matrix depending on time is obtained, i.e., \( m = m(t) \). The case \( M = N = 3 \) is shown next; it is split in two part for space reasons, the second part shows the last three columns.

\[
\begin{pmatrix}
4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{1,1}}{2} & 0 & -\nu + \frac{hv_{1,3}}{2} & 0 & 0 \\
-\nu - \frac{hv_{1,2}}{2} & 4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{1,2}}{2} & 0 & -\nu + \frac{hv_{1,2}}{2} & 0 \\
0 & -\nu - \frac{hv_{1,2}}{2} & 4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{1,2}}{2} & 0 & 0 \\
-\nu - \frac{hv_{2,1}}{2} & 0 & 0 & 4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{2,2}}{2} & 0 \\
0 & -\nu - \frac{hv_{2,2}}{2} & 0 & -\nu - \frac{hv_{2,3}}{2} & 0 & 0 \\
0 & 0 & -\nu - \frac{hv_{3,1}}{2} & 0 & -\nu - \frac{hv_{3,2}}{2} & 0 \\
0 & 0 & 0 & 0 & -\nu - \frac{hv_{3,3}}{2} & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 \\
-\nu + \frac{hv_{2,1}}{2} & 0 & 0 \\
0 & -\nu + \frac{hv_{2,2}}{2} & 0 \\
0 & 0 & -\nu + \frac{hv_{2,3}}{2} & 0 \\
4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{3,1}}{2} & 0 \\
-\nu - \frac{hv_{3,1}}{2} & 4\nu + \frac{h^2}{2\Delta T} & -\nu + \frac{hv_{3,2}}{2} \\
0 & -\nu - \frac{hv_{3,3}}{2} & 4\nu + \frac{h^2}{2\Delta T}
\end{pmatrix}
\]

(20)
All calculations are made with \( m = m(t) \), and so far it seems better than considering the constant matrix \( m \), mentioned before, as it is usually done. As is already known the matrix is symmetric and pentadiagonal of size \( N^2 \times N^2 \). For \( N \) (or \( M \)) large enough it gives rise a sparse matrix.

With \( \alpha = 0 \) and \( \nu = 1 \), it is clear that the above matrix is similar to \(-\nabla^2 u\) and \(-\nabla^2 v\), but constants; these latter matrices are diagonally dominant whereas the previous one is strictly diagonally dominant if \( \Delta t \), in \( \alpha \), is sufficiently small, \([8]\).

It should be noted that the above constant matrix \( m \) is the consequence of taking \( u \cdot \nabla \omega \), in (14), to the right hand side with the linear interpolation of the two previous time levels, i.e.,

\[
2(u \cdot \nabla \omega)^n - (u \cdot \nabla \omega)^{n-1},
\]

Therefore, in each time step, three algebraic linear equations systems, associated with the semi-discrete system for \( u, v \), and \( \omega \) in (12 – 14), will be solved, that is, algebraic systems of the form \( \mathcal{A} \, x = b \), with \( \mathcal{A} = m(t) \) (or \( m \)). Up to now these algebraic linear systems have been solved with LinearSolve, storing previously the matrix as a SparseArray, both Mathematica 8 Commands. The results are reported through the contour of \( \omega \) and \( \psi \); concerning \( \psi \), this is obtained, in the final time in which \( \omega \) has been already computed solving the equation for \( \psi \) given by (7).

3 NUMERICAL EXPERIMENTS AND DISCUSSION

The spatial discretization must be supplemented with boundary conditions, for each equation, on \( \Gamma \) for all \( t \geq 0 \), in this case for the driven cavity problem, we have \( (u, v) = (1,0) \) on the moving top wall \( y = 1 \) and \( (u, v) = (0,0) \) elsewhere; this problem, as mentioned above causes recirculation. Then in terms of this boundary condition for \( u \), the one given by \( u, v \) and \( \omega \) reads

\[
\begin{align*}
  u &= 0, v = 0; \quad \omega = \frac{\partial v}{\partial x} & \text{on } \Gamma_{x=0} \\
  u &= 0, v = 0; \quad \omega = \frac{\partial v}{\partial x} & \text{on } \Gamma_{x=1} \\
  u &= 0, v = 0; \quad \omega = -\frac{\partial u}{\partial y} & \text{on } \Gamma_{y=0} \\
  u &= 1, v = 0; \quad \omega = -\frac{\partial u}{\partial y} & \text{on } \Gamma_{y=1} \\
\end{align*}
\]

Besides, \( \omega(x,0) = \omega_0(x) \) denotes the vorticity initial condition which, by (5), has to satisfy \( \omega_0 = \frac{\partial v_0}{\partial x} - \frac{\partial u_0}{\partial y} \) if \( u_0 = (u_0, v_0) \) is the initial velocity.

For moderate Reynolds numbers \( Re \), the following values are reported: \( Re=400, 1000 \) and \( 4000 \); actually, \( Re = 400 \) is not shown but the calculation agrees perfectly with the one we are comparing with. The results, for each case of \( Re \) that is considered, correspond to steady state flows, the iso-contours for the vorticity \( \omega \) are shown first and the streamlines for the streamfunction \( \psi \) right below. The cases \( Re=400 \) and \( 1000 \) agree perfectly with the ones reported by [3] who solve the stationary problem using the streamfunction and vorticity formulation; they use the contour values \( vo = \{-5., -3, -1, 1, 3, 5.\} \) for vorticity and
\( v_z = \{ -0.11, -0.10, -0.08, -0.06, -0.04, -0.02, -0.01 \} \) for the streamfunction, which give rise the streamlines. Other kind of contour values are \( v1 = \{ -3., -2., -1., -0.5, 0., 0.5, 1., 2., 3., 4., 5. \} \) for vorticity given by [4] which are harder to obtain than those of [3], as has already already pointed out in earlier works of the third author. The case \( Re = 4000 \) does not agree at all with the one [3] but it does with the one in [1], where some explanation is given for this small disagreement.

The mesh size is denoted by \( h \) and the time step by \( \Delta t \) and they are specified in each case under study. Figures 1 and 2 pictures de iso-vorticity contours and streamlines for \( Re = 1000 \) respectively, obtained with \( h = 1/256 \) and \( \Delta t = 0.005 \). Figures 3 and 4 those for \( Re = 4000 \), obtained with \( h = 1/600 \) and \( \Delta t = 0.0025 \).

4 CONCLUSIONS

We have presented results for moderate Reynolds numbers that are obtained from a direct numerical method applied to the Navier-Stokes equations in its velocity-vorticity formulation, which is not a common formulation to use. Besides the novelty of the direct method the results become from a variable matrix depending on time and just using Mathematica 8. We are working to improve the scheme to be able to handle high Reynolds numbers as well as to extend it to thermal problems like those in [2] and in [7].

REFERENCES


Figure 1: $\omega$: $\text{Re}=1000$ (Keller); $h=1/256$, $dt=.005$
Figure 2: $\psi$: Re=1000 (Keller); $h=1/256$, $dt=.005$
Figure 3: $\omega$: Re=4000 (Keller); $h=1/600$, $dt=0.0025$
Figure 4: $\psi$: Re=4000 (Keller); h=1/600, dt=.0025
A FUNCTIONAL FOR THE MOMENTUM EQUATIONS OF INCOMPRESSIBLE VISCOUS FLOW

SHIHAI LI*, NINGNING LI* AND WENJIE DUAN*

* Institute of mechanics, Chinese academy of sciences
No.15 Beisihuanxi Road, Beijing, China
e-mail: shli@imech.ac.cn

Key words: Energy Functional, Incompressible Flow, Momentum Equations, Lagrangian Form, Eulerian Form.

Abstract. Vectorial mechanics and analytical mechanics are two time-honored forms of classical mechanics. Vectorial mechanics is mainly based on Newton’s laws in a clear and simple mathematical form. It has achieved a high degree of sophistication and success in solid mechanics. Analytical mechanics is based on the principle of virtual work and D’Alembert’s principle, which is highly universal. Often, the term vectorial mechanics is applied to the form based on Newton's work, to contrast it with analytical mechanics which uses two scalar properties of motion, the kinetic and potential energies, instead of vector forces, to analyze the motions. Analytical mechanics was primarily developed to extend the scope of classical mechanics in a systematic, generalized and efficient way to solve problems using the concept of constraints on systems and path integrals. In this paper, we give a functional of fluid in Lagrangian form. Then we demonstrate that the momentum equations of incompressible viscous flow can be achieved after several mathematical operations. At last, we show the Eulerian approximation of the energy functional under some assumptions. Our work lays a good foundation for our numerical methods.

1 INTRODUCTION

As well as in solid mechanics, there are also two time-honored approaches in hydrodynamics, namely, vectorial mechanics and analytical mechanics. Vectorial mechanics is mainly based on Newton’s laws in a clear and simple mathematical form. Due to Newton’s contribution to mechanics, vectorial mechanics has achieved an overwhelming advantage to analytical mechanics. The reference [1] gives several reasons why vectorial mechanics is preferred by engineers. In contrast to vectorial mechanics, analytical mechanics is based on the principle of virtual work and D'Alembert’s principle, which are highly universal. That is, if one more physical effect is to consider, just incorporate the corresponding energy into the total energy functional. With the development of FEM (finite element method), analytical mechanics has been widely used in solid mechanics. However, there is relatively less scientific research on variational principles of incompressible viscous flow. In hydrodynamics, the study of variational principles lags behind although several researchers have studied the variational principles in fluid mechanics [3, 4, 5, 6]. As Chien Wei-zang [2] stated, most of the work were focusing on inviscid flow and external flow. Based on the weak form of the momentum equations for incompressible flow, Chien Wei-zang[2] established an energy
functional in Eulerian form. Unfortunately, he gave an integral formulation for the terms corresponding the convective terms without the specific functional. Lu Wentang [1] developed a functional in Lagrangian form and showed the equivalence to incompressible viscous flow based the principle of least action. In his paper, the work done by the viscous force is evaluated along the path line of a fluid particle, which is relatively difficult to use in numerical methods. Due to the convective terms in the equations of incompressible viscous in Eulerian form, the symmetry would not exist and the equations can often become unstable if the Galerkin method is used. That is, when the incompressible flow problems are solved by the standard Galerkin method which employs the equal order basis for the velocity and pressure field, we will not be able to obtain simultaneously the satisfactory results for velocity and pressure [8]. Zienkienwicz [7] proposed the CBS (characteristic-based split) algorithm to circumvent the BB restrictions. So, it is essential to establish a more meaningful and practical functional. It is just what we are trying to do in the present paper.

The outline of the present paper is arranged as follows. First, we will develop a functional for incompressible viscous flow in Lagrangian form. Every term in the energy functional stands for a more physical meaning. We will also demonstrate its equivalence to the momentum equations of incompressible viscous flow using the Lagrangian equations. Once this is done, the Eulerian approximation to the Lagrangian energy functional will be given in section 3. In section 4, several conclusions will be drawn.

2 THE ENERGY FUNCTIONAL IN LAGRANGIAN FORM

Let

\[
\Pi \equiv \Pi_k + \Pi_p + \Pi_v + \Pi_\varepsilon + \Pi_f
\]

\[
\Pi_k \equiv \int_\Omega \frac{1}{2} \rho \dot{v}_i \dot{v}_i d\Omega \\
\Pi_p \equiv -\int_\Omega p u_i n_i dS + \int_\Omega \tau_{ij} u_j dS \\
\Pi_v \equiv \int_\Omega \tau_{ij} \varepsilon_{ij} d\Omega \\
\Pi_p \equiv \int_\Omega p u_i d\Omega \\
\Pi_f \equiv \int_\Omega f_i u_i d\Omega
\]

where \( \Omega \) is an arbitrary system full of material points, \( \partial \Omega \) is the boundary of the domain, \( \mathbf{X} \) are the curvilinear coordinates attached to the material while \( \mathbf{x} \) are the spatial coordinates, \( u_i = u_i(\mathbf{x}(\mathbf{X};t);t) \) and \( v_i = v_i(\mathbf{x}(\mathbf{X};t);t) \) are components of the displacement and velocity of material point \( \mathbf{X} \) at time \( t \) respectively, \( p \) is the pressure, \( \varepsilon_{ij} \) and \( \tau_{ij} \) are the strain and stress respectively, \( f_i \) is the body force (per unit volume), \( n_i \) is the unit outwards normal vector to the boundary. In incompressible flow, the density \( \rho \) is assumed constant and the velocity \( v_i \) is subjected to the following constraints

\[
v_{ij} = 0
\]

The boundary consists of two kinds of boundaries, namely velocity boundary \( S_1 \) and stress boundary \( S_2 \). The conditions for velocity \( v_i \) and traction \( t_i \) are introduced as follows:

\[
v_i = \hat{v}_i \ \text{on} \ S_1
\]

\[
t_i = (\rho \delta_{ij} + \mu (v_{ij} + v_{ji})) n_j = \hat{t}_i \ \text{on} \ S_2
\]
In addition, the initial conditions consist of specifying the values of velocity and pressure at the initial time:

\[ v_i(x_i, 0) = v_i^{(0)}(x_i) \]
\[ p(x_i, 0) = p^{(0)}(x_i) \]

The Lagrangian law states that the motion of a system subjected to complete constrains is governed by the following equations:

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial q_k} \right) - \frac{\partial T}{\partial q_k} = Q_k \quad (k = 1, 2, \cdots, N) \]

The above equations are called Lagrangian equations. Effectively, the Lagrangian law is also applied to a continuum. Now we will show how to get the momentum equation using the Lagrangian law.

### 2.1 Kinetic energy

Considering that the density \( \rho \) is constant, we know

\[ \frac{d}{dt} \left( \frac{\partial \Pi_k}{\partial v_i} \right) = \frac{d}{dt} \int \rho v_i d\Omega = \int \frac{\partial v_i}{\partial t} \rho d\Omega + \int \rho v_j v_{i,j} d\Omega \]

In the above equation, the Reynolds theorem and law of mass conservation are used.

### 2.2 Pressure energy

Given the Guass’s theorem, the energy \( \Pi_p \) can be written in the following form

\[ \Pi_p = \int p u_i d\Omega = \int (p u_i)_i d\Omega - \int p_u d\Omega = \int p u_i d\Omega - \int p_u d\Omega \]

### 2.3 Viscous energy

For Newtonian fluid, the work done by the viscous stress can be achieved by the following formulation

\[ \Pi_v = -\int \tau_{ij} \epsilon_{ij} dV = -\mu \int v_{i,j} u_{i,j} d\Omega - \mu \int v_{i,j} u_{i,j} d\Omega \]
\[ = -\mu \int v_{i,j} u_{i,j} d\Omega + \mu \int v_{i,j} u_{i,j} d\Omega + \mu \int v_{i,j} u_{i,j} d\Omega \]
\[ = -\int \tau_{ij} \epsilon_{ij} d\Omega + \int \tau_{ij} \epsilon_{ij} d\Omega \]

where \( \mu \) is the dynamics viscosity.

### 2.4 Summary of the equations

Substituting \( \Pi_k, \Pi_p, \Pi_v, \Pi_b \) and \( \Pi_f \) into the Lagrangian equations, we get

\[ \int \frac{\partial v_i}{\partial t} \rho d\Omega + \int \rho v_j v_{i,j} d\Omega + \int p_u d\Omega - \int \tau_{ij} \epsilon_{ij} d\Omega - \int f d\Omega = 0 \]
Considering that the $\Omega$ is arbitrary, the momentum equations of incompressible flow can be achieved as shown

$$
\rho \frac{\partial v_i}{\partial t} + \rho v_j v_{i,j} + p_j - \tau_{ij,j} - f_i = 0
$$

(10)

The above equation (10) or (11) is completed with the boundary conditions (2) and initial conditions (3).

3 AN EUCLERIAN APPROXIMATION TO THE ENERGY FUNCTIONAL

For any time interval $[\tau, T]$, there is a one-to-one transformation as shown

$$
x = x(X; t)
$$

$$
X = x(X; \tau)
$$

Therefore, the velocity of fluid particle $X$ is given by

$$
v(x(X; t); t) = \frac{\partial x(X; t)}{\partial t} [x]
$$

(12)

Based on the work of Duarte [9], we can build an approximation to $x$ as follows

$$
x = X + (t - \tau)v(x(X; \tau); \tau)
$$

(13)

It can be easily shown that this approximation will result in a first-order error of $O(t - \tau)$.

For simplicity, we introduce a general variable $\phi$. We denote $\Omega_{\tau}$ and $\Omega_{t}$, respectively, as the spatial domain occupied by a definitive system of fluid particles at time $\tau$ and $t$. Hence, we can obtain

$$
\int_{x \in \Omega_{\tau}} \phi(x(x(t); t)) d\Omega_{\tau} = \int_{x \in \Omega_{t}} \phi(x + (t - \tau)v(x(X; \tau); \tau); t) d\Omega_{\tau} + O[(t - \tau)^2]
$$

$$
= \int_{x \in \Omega_{\tau}} \phi|_{x = X, t = \tau} d\Omega_{\tau} + \int_{x \in \Omega_{t}} \frac{\partial \phi}{\partial t} |_{x = X, t = \tau} (t - \tau) d\Omega_{\tau}
$$

$$
+ \int_{x \in \Omega_{\tau}} \left[ v_k \left( \frac{\partial \phi}{\partial x_k} \right) \right]_{x = X, t = \tau} (t - \tau) d\Omega_{\tau} + O[(t - \tau)^2]
$$

(14)

Given the Guass’s theorem and the constraint of incompressibility, we can get the following approximation

$$
\int_{x \in \Omega_{t}} \phi(x(x(t); t)) d\Omega_{t} = \int_{x \in \Omega_{\tau}} \phi|_{x = X, t = \tau} d\Omega_{\tau} + \int_{x \in \Omega_{t}} \frac{\partial \phi}{\partial t} |_{x = X, t = \tau} (t - \tau) d\Omega_{\tau}
$$

$$
+ \int_{x \in \Omega_{\tau}} (\phi v_k) |_{x = X, t = \tau} (t - \tau) n_k dS_{\tau}
$$

(15)

From the viewpoint of numerical methods, the Equation (16) is always rewritten as

$$
\int_{x \in \Omega_{\tau}} \phi(x(x(t); t)) d\Omega_{\tau} = \int_{x \in \Omega_{\tau}} \left[ \phi + \frac{\partial \phi}{\partial t} (t - \tau) \right] |_{x = X, t = \tau} d\Omega_{\tau} + \int_{x \in \Omega_{\tau}} (\phi u_k) |_{x = X, t = \tau} n_k dS_{\tau}
$$

(16)
In principle, $\Pi_b$ should be expressed in Eulerian form. However, we still evaluate the integral over $\Omega$, instead of $\Omega$, since the time interval $\Delta t = T - \tau$ is always sufficiently small. For the same reason, the second term on the right-hand side of Equation (12) is also omitted. Hence, we obtain the Eulerian approximation to the energy functional:

$$
\Pi_{Euler} = \int_{\Omega_\tau} \left( \frac{1}{2} \rho \nu_i v_i + \rho u_{i,j} + \tau_{y,y} e_{ij} + f_i u_i \right) d\Omega
$$

$$
+ \int_{\partial \Omega_\tau} \left( \frac{1}{2} \rho \nu_i v_i + \rho u_{i,j} + \tau_{y,y} e_{ij} + f_i u_i \right) u_k n_k dS - \int_{\partial \Omega_\tau} p u_i n_i dS + \int_{\partial \Omega_\tau} \tau_{y,j} u_j n_j dS
$$

(17)

Using the Lagrangian equations on the Eulerian energy functional, we will get the momentum equations:

$$
\frac{d}{dt} \left( \frac{\partial \Pi_{Euler}}{\partial v_i} \right) = \int_{\Omega_\tau} \frac{\partial \nu_i}{\partial t} \rho d\Omega + \int_{\partial \Omega_\tau} \rho v_i v_{i,j} d\Omega + \int_{\partial \Omega_\tau} \frac{d}{dt} (\rho v_i) u_k n_k dS
$$

$$
- \int_{\partial \Omega_\tau} p_i d\Omega + \int_{\partial \Omega_\tau} \tau_{y,j} d\Omega + \int_{\partial \Omega_\tau} f_i d\Omega
$$

(18)

$$
\frac{\partial \Pi_{Euler}}{\partial u_i} = \int_{\Omega_\tau} -p_i d\Omega + \int_{\partial \Omega_\tau} \tau_{y,j} d\Omega + \int_{\partial \Omega_\tau} f_i d\Omega
$$

$$
+ \int_{\partial \Omega_\tau} \left( \frac{1}{2} \rho \nu_i v_i + \rho u_{i,j} + \tau_{y,y} e_{ij} + f_i u_i \right) u_k n_k dS
$$

$$
- \int_{\partial \Omega_\tau} p n_i dS + \int_{\partial \Omega_\tau} \tau_{y,j} n_j dS
$$

(19)

Inside the integration domain, the momentum equations can be obtained as follows

$$
\int_{\Omega_\tau} \frac{\partial \nu_i}{\partial t} \rho d\Omega + \int_{\partial \Omega_\tau} \rho v_i v_{i,j} d\Omega + \int_{\partial \Omega_\tau} p_i d\Omega - \int_{\partial \Omega_\tau} \tau_{y,j} d\Omega - \int_{\partial \Omega_\tau} f_i d\Omega = 0
$$

(20)

On the boundaries, the momentum equations satisfy:

$$
\int_{\partial \Omega_\tau} \frac{d}{dt} (\rho v_i) u_k n_k dS - \int_{\partial \Omega_\tau} \left( \frac{1}{2} \rho \nu_i v_i + \rho u_{i,j} + \tau_{y,y} e_{ij} + f_i u_i \right) u_k n_k dS
$$

$$
+ \int_{\partial \Omega_\tau} p u_i n_i dS - \int_{\partial \Omega_\tau} \tau_{y,j} u_j n_k dS - \int_{\partial \Omega_\tau} f u_k n_k dS + \int_{\partial \Omega_\tau} p n_i dS - \int_{\partial \Omega_\tau} \tau_{y,j} n_j dS = 0
$$

(21)

4 CONCLUSIONS

In this article, we give an energy functional of fluid in Lagrangian form and demonstrate the equivalence to the equations of incompressible viscous flow. In addition, under several assumptions, an approximation of the functional in Eulerian form is established. This functional lays a good foundation to develop a new numerical method of incompressible viscous flow.
ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of China National Program on Key Basic Research Project (973 Program, Grant No. 2010CB731500). The support from National Natural Science Foundation of China (NSFC) (NO:11002146) is fully acknowledged.

REFERENCES


MATHEMATICAL AND NUMERICAL APPROACH FOR A CRASHWORTHY PROBLEM

S. BORIA AND S. PETTINARI

School of Science and Technology
University of Camerino
Madonna delle carceri 9, 62032 Camerino, Italy
e-mail: simonetta.boria@unicam.it, web page: http://www.sst.unicam.it/SST/

Keywords: Theoretical Analysis, Crashworthy Problems, Energetic Approach, Composite Materials.

Abstract. Vehicle crashworthiness has been improving in recent years with attention mainly directed towards reducing the impact of the crash on the passengers. An optimal way to achieve this target is by exclusive use of specific impact attenuators, such as strategically placed tubular elements. Many of the mechanical devices are designed to absorb impact energy under axial crushing, bending and/or combined loading. An important requirement is that these structural members must be able to dissipate large amount of energy by controlled collapse in the event of a collision. Generally, the total energy dissipated depends on the governing deformation phenomena of all or part of the structural components of simple geometry, such as thin-walled tubes, cones, frames and sections. The energy absorbing capacity differs from one component to the next in a manner which depends on the mode of deformation involved and the material used.

During the last decades the attention given to crash energy management has been centred on composite structures. The use of fibre-reinforced plastic composite materials in automotive structures may result in many potential economic and functional benefits due to their improved properties respect to metal ones, ranging from weight reduction to increased strength and durability features.

Although significant experimental work on the collapse of fibre-reinforced composite shells has been carried out, studies on the theoretical modelling of the crushing process are quite limited since the complex and brittle fracture mechanisms of composite materials. Most of the studies have been directed towards the axial crush analysis, because it represents more or less the most efficient design.

In the present paper, a mathematical approach on the failure mechanisms, pertaining to the stable mode of collapse (Mode I) of thin-walled composite circular tubes subjected to axial loading, was investigated. The analysis was conducted from an energetic point of view; it is therefore necessary to identify the main energy contributions and then equate the total internal energy to the work done by the external load. The average crush load can be obtained minimizing the force contribution, function of several variables, on a domain using a numerical approach. Comparison between theory and experiments concerning crushing loads and total displacements was analysed, showing how the proposed analytical model is efficient for predicting the energy absorption capability of axially collapsing composite shells.
1 INTRODUCTION

The survivability of driver and passengers in an accident is achieved by a combination of the crash resistance of the car and its ability to absorb energy. This has been obtained by providing a survival cell, which is extremely resistant to damage, around which energy absorbing devices are placed at strategic points on the vehicle. From the numerous experimental studies [1-7], it is generally accepted that thin-walled tubes offer the most weight efficient solution for crashworthy aspects. The collapse behaviour of cylindrical and conical shaped shells of round cross sections has received attention in the recent years, because of their possible application to the design for crashworthiness of automotive vehicles. Moreover composite materials are now used in automobiles because they promise to be far more efficient than conventional materials. Carbon fibre reinforced plastic (CFRP) composites, in particular, provide significant functional and economic benefits such as enhanced strength, durability, weight reduction, lower fuel consumption and high level of structural vehicle crashworthiness. The challenge of design is to arrange the column of material such that the destructive zone can progress in a stable manner. On the contrary to the response of metals, progressive crushing of composite collapsible energy absorbers is dominated by extensive micro-fracture instead of plastic deformation. After the initial peak, the force displacement curve during axial crushing is much flatter than the plastically deforming metal tube and the amount of energy absorbed is, therefore, much greater. The mean load of crushing becomes an important factor to estimate while choosing a material and geometry for an impact energy absorbing application.

Recently the experimental campaign have been joined to the numerical analysis in order to predict the final deformation and validate the results obtained [8-10]. Contrary, very few attempts have been made to analyse the collapse mechanism of composite shells from the theoretical point of view [11-18]; this because the difficulty to model analytically the brittle behaviour and heterogeneity of these composite structures. Some additional work is rather on modelling of the plastic folding of round metallic tubes [19-21], always from the energetic point of view.

The present study addresses a mathematical and numerical approach on the failure mechanism, pertaining to the stable mode of collapse (Mode I), of thin-walled FRP composite circular tubes subjected to axial loading in order to predict the mean loads during collapse. The analysis is based on previous research [11-17] with the attempt to eliminate some simplifications dictated by experimental evidence and improve the modeling from the mathematical point of view. After the definition of the theoretical modeling using an energetic approach, different experimental tests were conducted in order to obtain some parametric values and crushing data useful for the model validation. Comparison between theory and experiments concerning loads and crushes was good enough, indicating therefore that the proposed theoretical model may be an efficient strategy for predicting the energy-absorbing capability of the axially collapsing composite shells, despite the complexity of the phenomenon.
2 THEORETICAL MODELING

During the crushing of a composite circular tube under axial impact, after the initial peak, the load oscillates around a mean load $P$. The first sharp drop in the load is due to the formation of a main circumferential intrawall crack of height $h$ at the top end parallel to the axis of the shell wall. As the deformation proceeds further, the externally formed fronds curl downwards with the simultaneous development, along the circumference of the shell, of a number of axial splits followed by splaying of the material strips. The post crushing regime is characterized by the formation of two lamina bundles bent inwards and outwards due to the flexural damage; they withstand the applied load and buckle when the load or the length of the lamina bundle reaches a critical value. At this stage, a triangular debris wedge of pulverized material starts to form; its formation may be attributed to the friction between the bent bundles and the platen of the drop mass of the hammer. Due to the constant size of the section, crush zone progresses through successive cycles that are repeated in the same manner. For this reason, the theoretical model investigated takes into account only the first cycle of deformation and the crush zone can be idealized as shown in Figure 1, where $R$ is the mean radius, $H$ is the height and $T$ is the thickness of the shell.

In order to simplify the deformation mechanism, the following assumptions have been adopted: the internal and external fronds maintain a constant length equal to $h$, thereby implying the same opening angle $\alpha$ the transition between straighten and bended zone is sudden, so the central crack is placed in A as shown in Figure 1; the elastic energy associated with the first impact phase was not considered, because very low respect to the other contributions. It has been observed in earlier studies [15] that energy is absorbed in four principal modes during the formation of crush zone in progressive crushing of cylindrical tubes: work required for bending of petals ($W_b$), work required for petal formation ($W_h$), work required for circumferential delamination ($W_c$) and energy dissipated due to friction between the debris wedge and fronds and between fronds and platen ($W_f$). Follow in detail the expressions used for the various energy contributions.
2.1 Bending energy

As the crushing process initiates, fibers bend both inside and outside the shell radius. Let \( t_1 \) and \( t_2 \) be the thickness of the fiber layers bending inside and outside the shell radius and \( \alpha \) the bending angles. By construction \( t_2 \) is equal to \( T-t_1 \). Assuming that the fiber layers are perfectly plastic during bending, the work required to bend the fiber inside and outside the shell radius can be expressed as

\[
W_b = W_{b1} + W_{b2} = (M_1 + M_2)\alpha = \frac{\pi}{2} R\sigma_0((T-t_1)^2 + t_1^2)
\]

where \( M_1 \) and \( M_2 \) are the bending moment capacity of the internal and external laminate, respectively, and \( \sigma_0 \) is the ultimate stress in uni-axial tension of the laminate.

2.2 Hoop strain energy

The expression for hoop strain energy in a single crush is

\[
W_h = \frac{1}{2} \sigma_0 \varepsilon dV = \sigma_0 \pi h^2 T \sin \alpha
\]

where \( \varepsilon = x \sin \alpha/R \) is the shear strain corresponding to the layers bending inside and outside the shell radius. For \( x \) it is indicated the distance from A to B, as shown in Figure 1, while \( dV \) is the differential volume for the inside (outside) layers.

2.3 Crack energy

The energy required for circumferentially delamination in a single stroke is

\[
W_c = 2\pi(R-T/2+t_1)hG
\]

where \( G \) is the critical strain energy release rate per unit interlaminar delaminated crack area and \( R-T/2+t_1 \) is the tube radius at the crack tip. \( G \) is determined experimentally through DCB test as discussed in Section 3.

2.4 Friction energy

After the formation of the internal and external fronds, normal stresses develop on the sides of the debris wedge followed by shear stresses along the same sides due to friction at the interface between the wedge and the fronds. Moreover additional normal and shear stresses develop at the interface between the steel platen and the deforming shell as the formed fronds slide along the interface. The energy dissipated in frictional resistance for a crush distance \( \delta \) is

\[
W_f = 4\pi R(\mu_1 P_1 + \mu_2 P_2)\delta
\]

where \( \mu_1 \) is the coefficient of friction between frond and platen and \( \mu_2 \) the coefficient of friction between the wedge and the fronds. Due to feasibility \( 0 < \mu_1 \leq \mu_2 \leq 1 \) and by
construction \( \delta = h(1 - \cos \alpha) \). \( P_1 \) is the normal force per unit length applied by the platen to the frond and \( P_2 \) is the normal force per unit length applied to the sides of the wedge, as shown in Figure 1. Static equilibrium at interface yields

\[
P = 2\pi R(P_w + 2P_1)
\]

where \( P_w \) is the normal force per unit length applied by the platen to the debris wedge given by

\[
P_w = 2(P_1 \sin \alpha + T_2 \cos \alpha)
\]

where \( T_2 = \mu_2 P_2 \) is the frictional force per unit length developed between wedge and fronds. According to [16]

\[
P_2 = \sigma h
\]

and the horizontal component of the force \( F_x \) applied by the wedge to the frond is

\[
F_x = P_2(\cos \alpha - \mu_2 \sin \alpha).
\]

For feasibility \( F_x \) has to be strictly positive, therefore \( \alpha < \tan^{-1}(1/\mu_2) \) that implies an angle \( \alpha \) not smaller than 45°. The frictional energy (4) can be rewritten as

\[
W_f = \mu_2 \delta P + 4\pi R \sigma h \delta (\mu \sin \alpha - \mu_2 \mu \cos \alpha + \mu_2).
\]

2.5 External work

The work done by the external load \( P \) on the crushing displacement \( \delta \) in a single progression is

\[
W_e = P \delta.
\]

Balance of energy yields

\[
W_e = W_i + W_t + W_r
\]

so the mean force \( P \) is given by

\[
P = P(h, t, \alpha) = \frac{1}{\delta(1 - \mu_t)}(W_i + W_t + W_r + 4\pi R \sigma h \delta (\mu_2 \sin \alpha - \mu_2 \mu_2 \cos \alpha + \mu_2)) = \frac{\pi}{2(1 - \mu_t) h(1 - \cos \alpha)} \cdot \left[2Gh(2R + 2t - T) + 4R \sigma R t \left(t^2 + (T - t)^2\right) + 2\sigma h^2 \left[T \sin \alpha + 4R(1 - \cos \alpha)(\mu_2 - \mu_2 \mu \cos \alpha - \mu_2 \sin \alpha)\right]\right].
\]

The expression for determining the crush zone length, the thickness of the plies bending inside the shell radius and the bending angle of the internal plies in a crush cycle is obtained by minimising the average crush load.

Note that this procedure allows to calculate the energy absorbed in a single crush cycle, as the product of the obtained load \( P \) for the vertical displacement \( \delta \). The total crush can be seen as the composition of equal single stages that repeat themselves. Therefore in order to evaluate the final deformation of the impact attenuator able to absorb the potential energy \( W_p \)
imposed by regulation, it is sufficient to divide \( W_p \) for the minimum load \( P \) obtained by previous calculation.

### 3 FABRICATION AND TESTING OF COMPOSITE SPECIMENS

A single kind of CFRP composite material was used for testing. Despite unidirectional composites being more efficient in energy absorption, fabric reinforced materials tend to be preferred in impact structures in order to ensure a stable crush failure. Therefore prepreg tapes of carbon fibres in epoxy resin with plain reinforcements were used to make the tubes manually. To obtain appropriate value for the \( \sigma_0 \) parameter, standard coupon tests were performed. The tests were performed in tension, according to ASTM Standard D3039, with the warp (weft) direction aligned with the test direction (0° and 90° tests, respectively) and also with fibres aligned at 45° to the test direction [9]. For the ultimate stress of the laminate was used that obtained from the 45° tensile test (Figure 2), in order to consider the worst condition of resistance given the manual realization of the tubes.

![Figure 2: Stress vs strain diagram for prepreg in tensile test](image)

The double cantilever beam (DCB) test was carried out as per ASTM D5528-94, in order to evaluate the mode I interlaminar fracture toughness in the same composite laminate used for the tubes. The test considers a composite beam with an initial delamination crack; the initial delamination is forced to open by applying a force that pull the two beams away from each other (Figure 3). From the test the fracture toughness \( G \) was calculated using the compliance calibration method, which is the most common in literature.

![Figure 3: DCB test](image)
In order to obtain experimental evidence about the governing failure mechanism during the axial compression of thin-walled CFRP composite circular tubes, dynamic tests were performed using a drop weight test machine with an impact mass of 294kg and an impact velocity of about 4m/s [9]. During the tests the acceleration of the mass were acquired, sampled and recorded. From the average deceleration, obtained from the beginning of the impact to the instant when the velocity vanishes, is then possible to determine the mean load and the real crushing.

4 RESULTS AND DISCUSSION

As discussed in Section 2, the central issue of the procedure is determining the minimum of the mean load $P$. In the present section, first the existence of a local minimum of $P$ inside the domain $D$ is proved, than since the roots of the gradient of $P$ cannot be analytically determined, the minimum is estimated using numerical methods.

The mean load $P$ is a function of the three variables $h$, $t_1$ and $\alpha$ belonging to the domain $D=]0,H]\times]0,T]\times]\pi/4,\pi/2]$. Consider a minimizing sequence for $P$, $(h_n,t_1,\alpha_n)_{n\in\mathbb{N}}$, in the set $D$. Equation (12) implies that $P$ tends to infinity when $h$ vanishes therefore $h_n$ is far from zero and by Weierstrass theorem there exists at least a minimum of $P$ in $D$. Figure 4 displays the direction of the gradient of $P$ on the boundary of the domain.

$$\nabla P_{\partial D}$$

In details, choosing $H>\sqrt{\pi TR}/\sqrt{2}$; $\partial_h P|_{h=H}>0$; setting $R$ and $T$ such that $RT>8GH/(\pi\sigma_0)$ $\partial_{\alpha} P|_{\alpha=0}<0$; $\partial_{\alpha} P|_{\alpha=T}>0$ and $\partial_a P|_{\alpha=\pi/4}<0$. In order to state that the minimum is inside the domain $D$, it remains to consider the edge face with $\alpha=\pi/2$. If $\alpha=\pi/2$ the function $P(h,t_1,\pi/2)$ has a minimum $P'$ in $(h^*,t_1^*,\pi/2)$ with

$$h^* = \frac{\pi\sigma_0 RT}{2\sqrt{2}(4\pi R^2 \sigma_0^2 (\mu_2 - \mu_1) + \pi\sigma_0^2 R - 2G^2)} ,
\quad t_1^* = \frac{1}{2} \left( T - \frac{\sqrt{2}GT}{\sqrt{4\pi R^2 \sigma_0^2 (\mu_2 - \mu_1) + \pi\sigma_0^2 R - 2G^2}} \right)$$

Since the one variable function $P(h^*,t_1^*,\alpha)$ is strictly increasing in $\alpha=\pi/2$, there exists an angle $\alpha < \pi/2$ in which $P(h^*,t_1^*,\alpha)<P'$, therefore the minimum of $P$ is in the interior of $D$. 

1057
It is not easy to find the roots of the gradient of $P$ so the minimum is estimated using some nonlinear optimization algorithms implemented with the software Mathematica [23] as a preliminary study. Both the Nelder-Mead “simplex” algorithm, the Random search method [24] and the Differential evolution procedure [25] give the same minimum point of $P$ for every specific cases analyzed; so it is reasonable to assume that the mean load $P$ has an unique minimum internally to the domain.

Table 1 reports the geometrical and material parameters of the cylindrical shells taken into account.

<table>
<thead>
<tr>
<th>#</th>
<th>$R$ (mm)</th>
<th>$T$ (mm)</th>
<th>$H$ (mm)</th>
<th>$\sigma_0$ (MPa)</th>
<th>$G$ (N/mm)</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.25</td>
<td>2.5</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>2</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>3</td>
<td>40.75</td>
<td>1.5</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>26.25</td>
<td>2.5</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>2</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>6</td>
<td>25.75</td>
<td>1.5</td>
<td>200</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>51.25</td>
<td>2.5</td>
<td>300</td>
<td>140</td>
<td>0.5</td>
<td>0.35</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 2 reports the mean load and the final crushing in the minimum configuration for the cases taken into account. In the table there are also the experimental data taken from real tests [9] and the committed percentage error. Despite the simplification adopted the proposed method is able to predict within ±20% the mean crushing load and the total crushing. As it is evident by table, the error tends to increase as the wall thickness decreases; this can be justified by the real deformation which, for thinner thickness, changes from axial splitting in splaying mode to fragmentation mode (Figure 5). This second mode of failure have a lower energy absorption capacity with also lower force values.

<table>
<thead>
<tr>
<th>#</th>
<th>P (kN)</th>
<th>s (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56.1</td>
<td>59.4</td>
</tr>
<tr>
<td>2</td>
<td>43.7</td>
<td>46.4</td>
</tr>
<tr>
<td>3</td>
<td>31.9</td>
<td>26.8</td>
</tr>
<tr>
<td>4</td>
<td>37.7</td>
<td>38.9</td>
</tr>
<tr>
<td>5</td>
<td>29.1</td>
<td>26.1</td>
</tr>
<tr>
<td>6</td>
<td>21.0</td>
<td>17.7</td>
</tr>
<tr>
<td>7</td>
<td>68.3</td>
<td>75.9</td>
</tr>
</tbody>
</table>

Table 2: Geometrical and material characteristics of the shells

<table>
<thead>
<tr>
<th>#</th>
<th>P (kN)</th>
<th>s (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56.1</td>
<td>59.4</td>
</tr>
<tr>
<td>2</td>
<td>43.7</td>
<td>46.4</td>
</tr>
<tr>
<td>3</td>
<td>31.9</td>
<td>26.8</td>
</tr>
<tr>
<td>4</td>
<td>37.7</td>
<td>38.9</td>
</tr>
<tr>
<td>5</td>
<td>29.1</td>
<td>26.1</td>
</tr>
<tr>
<td>6</td>
<td>21.0</td>
<td>17.7</td>
</tr>
<tr>
<td>7</td>
<td>68.3</td>
<td>75.9</td>
</tr>
</tbody>
</table>

Table 2: Mean crush load and final crushing
Figure 5: Different crushing mode: a) splaying with axial splitting, b) fragmentation

Figure 6 graphically represents the trend of the mean load for a specific configuration (case 1) in function of the bending angle $\alpha$ and of the thickness of the inside ply $t_1$, set $h$ equal to the minimum value 2.73 mm; the black point is the minimum of $P$. As it is observed from experimental evidence, the delamination in all the cylindrical tubes analyzed occurs at about half of the wall thickness.

5 CONCLUSIONS

A crashworthy problem was investigated from a mathematical and numerical point of view. In particular the energy absorption of composite cylindrical tubes subjected to axial loading was analyzed, defining analytically the external load as a function of three variables and identifying the minimum through a numerical approach. Despite some simplifications, the method adopted is able to predict the mean load and the final crushing of tubular specimens made of CFRP composite, once known parameters of the used material. This methodology can therefore be used as the first approach to follow during the design of specific CFRP impact attenuator. Forthcoming analysis will eliminated the assumption of a constant crush length both internally and externally to the axial splits and will consider different numerical methods.
REFERENCES


A COUPLED PROBLEM OF HEAT AND MASS TRANSFER APPLIED TO POROUS TEXTILE MEDIA SURROUNDING THE HUMAN FOOT

J. DURANY*, L. POCEIRO* AND F. VARAS*

*Dep. Matemática Aplicada II, E. I. Telecomunicación
Universidad de Vigo
Campus Marcosende, 36310 Vigo, Spain
e-mail: durany@dma.uvigo.es, web page: http://www.dma.uvigo.es/

Key words: Multiphysics, Foot Temperature, Foot-Sock Heat-Mass Fluxes, Finite Elements

Abstract. There have been many attempts in the literature to model human thermoregulation systems in order to predict central temperature and other heat stress indicators (see [3, 9], among many others). These models depend on a wide range of variables, including individual characteristics, surrounding textile and environmental factors. The main objective of this work is to formulate a stable, realistic and versatile 2-D mathematical model describing the heat transfer of the human foot (bare foot and foot surrounded by textile materials). The novelty, but also the difficulty, lies on the theoretical multiphysics that models the textile as a porous media, involving energy transport but also mass transport of liquid water, water vapour and gas, including evaporation phenomena.

The numerical solution to the global problem involves a segregated algorithm and fixed point techniques for the nonlinearities jointly with finite elements spatial discretizations. Implementation has been performed through commercial software COMSOL$^{TM}$ Multiphysics.

1 INTRODUCTION

Most of models to predict human thermoregulation systems are based on empirical properties that simplify human body global geometry, or simply apply theories in onedimensional models (see [4, 9], among others, and their references). There are some works that particularize the studies for some body parts. In the context of thermal modelling of the human foot followed by numerical simulation, it should be mentioned Covill et al. works [2, 3] using the finite element method.

The thermal model of the bare foot is based on the metabolic bio-heat equation, including the blood flow as a heat source and the effects of sweating. Here, this equation is
coupled to mass and energy transfer models at the textile that surrounds the foot (sock).
Final formulation leads to a highly non-linear partial differential equations system, with
adequate boundary conditions and heat transfer conditions between media.

The global model is solved by means of a segregated algorithm, fixed point techniques
for the non-linearities along with finite elements spatial discretizations. Implementation
is performed by using commercial software COMSOL\textsuperscript{TM} Multiphysics.

Finally, some simulations of several examples try to show that the numerical results
are qualitatively acceptable. This 2-D coupled model can be a first step for equations
and parameters adjustments, in order to apply it to more realistic three-dimensional
geometries but, in any case, allows to carry out certain analysis on geometry design and
on textiles composition that could be of great interest for the manufacturing industry.

2 THERMAL MODEL FOR THE FOOT-SOCK COUPLING

Usually, a hard simplification of the foot organs (bones, muscles, tendons, etc) is made
in order to reduce the complex internal composition to a single tissue, where the thermal
and biological properties are averaged. However, consideration of the foot skin as another
tissue does not add new formulation difficulties.

Here, it is assumed that the transverse heat flow is much smaller than the longitudinal
and vertical ones. So, the equation should be verified in a two-dimensional domain, $\Omega$, 
that corresponds to a longitudinal section of an average human foot of a length of 28cm, 
approximately (see Fig. 1, where it also shows the mesh used for the finite element 
method, the foot internal tissue, the 2mm thick skin and the 3mm thick sock). This 
equation in steady state form is given by:

$$-\nabla \cdot (k \nabla T) - \rho_b c_b \omega_b (T_b - T) = Q_m,$$

in $\Omega$ ,

(1)

where $k$ the is thermal conductivity ($W m^{-1} K^{-1}$), $\rho_b$ the blood density ($kg m^{-3}$),
c$_b$ the blood specific heat ($J kg^{-1} K^{-1}$), $\omega_b$ the blood perfusion ($s^{-1}$), $T_b$ the blood temperature
($K$), $Q_m$ the metabolic heat rate ($W m^{-3}$) and $T$ is the unknown temperature ($K$).

The textile medium $\Omega_c$ can be considered as an unsaturated porous medium. Therefore,
the model should consider (apart from the thermal equation) water and gas transport, in order to take into account sweating effects from the foot to the environment. In general, the thermal model consists of a diffusion equation for the unknown temperature, $T_c$, with a generation term that includes sweating/evaporative effects. That is:

$$-\nabla \cdot (k_c \nabla T_c) = -S_T, \quad \text{in } \Omega_c,$$

where $S_T$ is a heat source/sink due to sweating and is given by:

$$S_T = H_{lat} S_v,$$

being $H_{lat}$ the latent heat of vaporisation ($J kg^{-1}$) and $S_v$ the evaporation rate which is a priori unknown too. And finally, $k_c$ is the thermal conductivity ($W m^{-1} K^{-1}$) which will be determined from solid and gas properties of the textile, as follows:

$$k_c = k_t (1 - \phi) + k_w \theta_l \phi + k_g (1 - \theta_l) \phi,$$

where $k_t$, $k_w$ and $k_g$ are the thermal conductivity of fibers, liquid water and gas mixture in the sock, respectively, $\phi$ the textile porosity coefficient and $\theta_l$ the liquid water concentration in the sock.

To model liquid water transport in the porous medium, Richards equation is used (see [7], among others), which consists of a variant of Darcy’s law considering that both the hydraulic conductivity and the matric potential depend on the saturation level of the medium (Darcy’s law only considers saturated porous media):

$$-\rho_l \nabla \cdot (K_h(\theta_l) \nabla \psi_h(\theta_l)) = -S_v, \quad \text{in } \Omega_c,$$

with water volumetric concentration, $\theta_l$, expressed as per-unit, being the unknown of the problem. Moreover, $\rho_l$ denotes the density of liquid water ($kg m^{-3}$), $K_h$ the hydraulic conductivity ($ms^{-1}$), $\psi_h$ the hydraulic potential ($m$) and $S_v$ the vaporisation rate (a priori unknown).

The hydraulic potential can be written as the sum of two potentials (see [7] for details):

$$\psi_h = \psi_m + \psi_z,$$

where $\psi_m$ denotes the hydraulic potential and $\psi_z$ denotes gravitational potential that depends on the position of the terrestrial gravitational field:

$$\psi_z = \rho_l g (z - z_0),$$

with $g$ denoting the gravity acceleration ($ms^{-2}$) and $z$ the height, ($m$), measured in relation to a reference value, $z_0$.

Following van Genutchen approximation [6], the hydraulic conductivity is given by:

$$K_h(\Theta) = K_s K_r(\Theta) = K_s \Theta^\beta \left(1 - \left(1 - \Theta^\alpha \right)^m \right)^2,$$
where $K_s$ is the constant value of the hydraulic conductivity in saturation ($m\,s^{-1}$) and $K_r$ the relative hydraulic conductivity, that depends on the effective saturation $\Theta = \frac{\theta_l - \theta_r}{\theta_s - \theta_r}$ and, therefore, on the concentration of liquid water in the material, $\theta_l$. This effective saturation depends on $\theta_l$ and on the parameters $\theta_r$ and $\theta_s$, that denote the residual concentration of moisture and the water concentration in saturation, respectively. Moreover, in (8) it is found the parameter $m$, that is usually written as $m = 1 - \frac{1}{n}$, and both $m$ and $n$ are usually named as the shape parameters of the retention curve.

Finally, the matric potential, following van Genutchen proposal [6], is written:

$$\psi_m (\Theta) = -\left(\frac{\Theta^{\frac{1}{n}} - 1}{\alpha}\right)^{\frac{n}{m}} ,$$

where $\alpha$ is also a shape factor of the retention curve.

The equation of the water vapour transport can be written as follows:

$$-\nabla \cdot \left(\frac{k \rho_g}{\mu} \nabla P_v\right) - \nabla \cdot (D_v(T_c)\xi_g(\theta_a)\nabla m_v) = S_v, \quad \text{en } \Omega_c ,$$

where $m_v$, a priori unknown, denotes the mass concentration of the vapour in the total textile volume ($kg\,m^{-3}$), $k$ is the medium’s permeability ($m^2$), $\rho_g$ is the density of the gaseous mixture ($kg\,m^{-3}$), $\mu$ is the air viscosity ($kg\,m^{-1}\,s^{-1}$), $P_v$ is the partial pressure of the water vapour ($Nm^{-2}$), $D_v$ is the effective diffusivity of the vapour in the air ($m^2\,s^{-1}$) and $\xi_g$ is the tortuosity of the vapour transport.

The mass concentration of the vapour in the textile medium can be written as follows:

$$m_v = \rho_v \theta_a = \rho_v (\phi - \theta_l) ,$$

being $\rho_v$ the vapour density in the available volume (open pore) ($kg\,m^{-3}$), $\theta_a$ the concentration of the non-saturated porous space ($m^3\,m^{-3}$), and $\phi$ the porosity coefficient of the textile medium.

For calculating the vapour density, $\rho_v$, a chemical equilibrium hypothesis among the two water phases is imposed. This hypothesis implies a kinetic of phase change infinitely fast since mass interchanges occur instantaneously. In this case, classical Kelvin’s equilibrium equation is used (see [1], for instance):

$$\frac{P_v}{P_{v,\text{sat}}} = \exp \left(\frac{(\rho_g\gamma_{vm})M_l}{\rho_l RT_c}\right) ,$$

where $P_{v,\text{sat}}$ denotes the partial vapour pressure in saturation conditions ($Nm^{-2}$), $R$ is the constant of the ideal gases ($JK^{-1}mol^{-1}$) and $M_l$ denotes the molecular weight of the water ($kg\,mol^{-1}$). This expression (12) consists on a dimensionless volumetric relation measuring the saturation of the gas. Moreover, $\frac{P_v}{P_{v,\text{sat}}}$ is often called relative humidity, $h_r$, to give it a more physical sense. That is:

$$P_v = P h_r ,$$
where \( P \) is the pressure of the gaseous phase \((N\,m^{-2})\), that will be analyzed afterwards.

Now, using the ideal gas law, vapour density \( \rho_v \) in (11) is posed in terms of \( P_v, T_c \):

\[
\rho_v = \frac{P_v M_l}{RT_c},
\]

and, consequently, in terms of \( P, \theta_l \) and \( T_c \).

Fluxes of the gaseous phase are due to the gradient variations in temperature and pressure and are given by Darcy’s law:

\[
-\nabla \cdot \left( \frac{\kappa \rho_g}{\mu} \nabla P \right) = S_v, \quad \text{in } \Omega_c,
\]

where the main unknown is the pressure of the gaseous phase, \( P, (N\,m^{-2}) \). Moreover, (15) involves the permeability of the medium \( \kappa \) \((m^2)\), the air density, \( \mu \) \((kg\,m^{-1}\,s^{-1})\), and the gaseous phase density \( \rho_g \) \((kg\,m^{-3})\), that can be also written in terms of \( P \) and \( T_c \) using the ideal gas law:

\[
\rho_g = \frac{P M_g}{RT_c},
\]

where \( M_g \) is the molecular weight of the gaseous mixture.

Remark that equation (12) gives an implicit description of the source term \( S_v \). So, it can be eliminated as problem unknown combining (5) and (10), and using (12) to close a system in terms of \( T_c, \theta_l \) and \( P \). In short, the coupled mathematical model consists on the one hand on the bio-heating equation (1) in the domain that comprises the foot, being the unknown the temperature \((T)\). On the other hand, in the inner part of the textile medium the equations for energy flow (2), water transport (5) and gas transport (15) are considered. In this case, the unknowns are the temperature \((T_c)\), the volumetric concentration of the liquid water \((\theta_l)\), the pressure of the gaseous phase \((P)\) and the vaporisation rate \((S_v)\), that can be calculated in a direct way by coupling the system with the vapour transport equation (10).

3 CONTACT AND BOUNDARY CONDITIONS

Assuming perfect contact between both media (foot and textile medium), transmission conditions for the thermal problems, (1), (2), are given by the continuity of the temperatures and heat fluxes. For the gas transport equation (15), it is supposed that there are no gas flux from the foot. However, for the water transport equation (5), it is considered that water flux from the foot is given by the evaporative heat losses due to sweating:

\[
\mathbf{n} \cdot (-K_h \nabla \psi_h) = -\frac{Q_e}{H_{lat}},
\]

being \( Q_e \) the sweating rate defined by the same empirical formula that appears in [4] and [9], among others, following the Fanger classical model [5]. This formula links sweating
rate with the difference of the foot temperature at the skin surface $T$ ($K$) and an empirical expression related to the environmental water vapour pressure $P_{vatm}$ ($Kpa$). That is:

$$Q_e = Q_{eds} + Q_{ews},$$

(18)

where $Q_{eds}$ is the heat loss due to implicit sweat secretion when the skin is dry, and $Q_{ews}$ the corresponding heat loss due to evaporation of explicit sweat secretion:

$$Q_{eds} = 0.782 \left( T - \left(286.31 + \frac{P_{vatm}}{0.256}\right) \right);$$

$$Q_{ews} = h_e s_h 4.275 \left( T - \left(286.31 + \frac{P_{vatm}}{0.256}\right) \right),$$

(19)

with $s_h$ a nondimensional parameter denoting the skin humidity ($0 \leq s_h \leq 1$, from dry to wet). Moreover, $P_{vatm}$ can be written as follows:

$$P_{vatm} = P_{sat} H_r,$$

(20)

being $H_r$ the air relative humidity (on a per unit basis) and $P_{sat}$ the saturation pressure of the water vapour at the environment temperature ($kpa$). Using thermodynamical equilibrium, $P_{sat}$ is classically given by some empirical formula, as the following one (see [8], for example):

$$P_{sat} = 0.611 e^{\frac{17.27(T_a-273)}{T_a-273}},$$

(21)

In the upper location of the domain, where the foot is connected to the rest of the body, two parts of the boundary need to be distinguished. First, $B1$ the boundary in contact with the rest of the leg, where the same condition of null thermal flux is posed for the temperature $T$. On the other hand, the boundary between the textile medium and the environment that is denoted by $B1c$ and affects to the temperature $T_c$, to the water concentration $\theta_l$ and to the pressure $P$. As $B1c$ is extremely small ($3mm$), no heat nor mass transfer is considered through this boundary.

Finally, it is supposed that on the boundary textile-environment, $B2$, there are heat losses due to convection and radiation. Based on this, the boundary condition for the thermal equation (2) can be written as follows:

$$\mathbf{n} \cdot (-k_c \nabla T_c) = h \left( T_c - T_a \right) + \varepsilon \sigma \left( T_c^4 - T_a^4 \right),$$

(22)

where $\mathbf{n}$ is the normal, unitary and external vector to $B2$, $h$ is the convection coefficient ($W m^{-2} K^{-1}$), $T_c$ is the temperature on the surface of the textile medium ($K$), $T_a$ is the environmental temperature ($K$), $\varepsilon$ is the emissivity coefficient (dimensionless) and $\sigma$ is the Stefan-Boltzmann constant ($W m^{-2} K^{-4}$).

For the liquid water transport equation (5), a similar condition is given. That is, the flux is proportional to the difference of the vapour partial pressure in each media:

$$\mathbf{n} \cdot (-k_h \nabla \psi_h) = \frac{k_{atm} M_l}{R T_c} \left( P_{vatm} - P_v \right),$$

(23)

on $B2$. 

1067
where \( k_{atm} \) is the mass conductivity of the interface textile-air \((m \, s^{-1})\), \( M_l \) is the molecular mass of the water \((kg \, mol^{-1})\), \( R \) is the perfect gases constant \((J \, mol^{-1} \, K^{-1})\), \( P_{v_{atm}} \) is the vapour pressure in the atmosphere \((N \, m^{-2})\) and \( P_v \) is the vapour pressure in the porous medium \((N \, m^{-2})\).

For the gas transport equation (15) it is supposed that existing flux through the boundary is proportional to the pressure differences between atmosphere pressure and gas phase pressure in the porous medium, as proposed in the following equation:

\[
\mathbf{n} \cdot \left( -\frac{\kappa \rho_g}{\mu} \nabla P \right) = \frac{k'_{atm} M_g}{RT_c} (P_{atm} - P), \quad \text{on } B2, \tag{24}
\]

where \( k'_{atm} \) is the gas mass conductivity of the interface textile-air (supposed to be two orders of magnitude lower than the mass conductivity of the vapour flux) \((m \, s^{-1})\), \( M_g \) is the molar mass of the air and \( P_{atm} \) is the atmosphere pressure \((N \, m^{-2})\).

4 NUMERICAL SOLUTION AND EXAMPLES

For the numerical solution of this bi-dimensional coupled model, commercial software COMSOL\textsuperscript{TM} Multiphysics is used. This software is based on the finite element method for spatial discretizations and it seems adequate for solving the partial differential equations system of the coupled and strongly non-linear model presented in previous sections.

In order to facilitate the introduction of the model in the computer code, a series of functions, variables and constants have been defined, appearing in tables 1 and 2 of the appendix.

The first simulation example tries to highlight the numerical results obtained from the global model. So, Fig. 2 shows the temperature distribution for the bare foot and for the foot surrounded by a sock with a porosity of 50% and for an environmental temperature of 25°C and relative humidity of 35%. The temperature inside the foot slightly increases for being surrounded by the sock, also increasing the temperature on its external surface. Specifically, at the point of the tip of the toe with coordinates \((0.28 \, m, 0.015 \, m)\) it increases from 31.03°C to 32.49°C, while at the external point of the sock at the tip toe, with coordinates \((0.283 \, m, 0.015 \, m)\), it is of 29.01°C.

Next example shows the variations of the unknowns of the global system inside the sock for an extreme case of environmental temperature of 35°C. With the aim of visualizing these variations in a domain of such an small thickness, a zoom of the tip of the toe is considered (see Fig. 3) taking a line of nodes from the interior point of the sock in contact with the skin (point coordinates \((0.28 \, m, 0.015 \, m)\)) to the external surface (point coordinates \((0.283 \, m, 0.015 \, m)\)). The decrease of the temperature in this line inside the sock shown in Fig. 3 is due to the thermal gradient generated by the difference of temperature between the foot and the environment. A similar behaviour can be observed in Fig. 4 for the sock pressure though it is not quantitatively very relevant in this case.

Variations of water concentration at the same line are very small as showed the Fig. 5.
Finally, varying the values of the porosity coefficient of the textile medium, it is intended to validate the qualitative properties of the proposed coupled model. In fact, if the sock porosity tends to be very big, the skin temperature obtained from a bare foot model and the corresponding one using the coupled model with the textile medium should be very similar. And, it can be noted in Fig. 6 that for a porosity of 95% and an environmental temperature of 35°C, the temperature at the surface of the foot (point \(0.28 m, 0.015 m\)) is 36.05°C and at the external surface of the sock (point \(0.283 m, 0.015 m\)) is 35.89°C, while with the bare foot model the temperature obtained at the same point of the foot surface is 35.74°C. This validation test is very important since it sets the trend of the temperature respect to extreme porosities.

5 CONCLUSIONS

Literature sources does not show such a complete model as the one here proposed for the thermal coupling of a human foot to a textile medium (a sock or a pant). This model is configured from the classic equation of the bio-heating energy in the foot and the theoretical multi-physics on porous media that affects the textile medium, where mass transport is involved (water, vapour and evaporating phenomena).
Numerical simulations here presented show a results qualitatively acceptable. Thus, it seems that both the model and the finite elements discretizations selected for the numerical solution and the software used fulfil the objectives of this study.

ACKNOWLEDGEMENTS

This work has been partially funded by the Research Project MTM2010-21135-C02-02 of the Spanish Ministry of Education and FEDER Funds (EU).

REFERENCES


Figure 6: Temperature variation at the point of the skin surface with coordinates (0.28 m, 0.015 m) (top) and at the point of the surface of the sock with coordinates (0.283 m, 0.015 m) (bottom), for different porosities of the sock textile, and environmental temperature $T_a = 35^\circ$.


**Appendix: Data of the global model**

<table>
<thead>
<tr>
<th>Var.</th>
<th>Expression</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$k_f$ or $k_s$</td>
<td>$W/(mK)$</td>
<td>Foot thermal conductivity</td>
</tr>
<tr>
<td>$S_T$</td>
<td>$H_{lat}S_v$</td>
<td>$J/(s m^3)$</td>
<td>Heat Source\Sink due to sweating</td>
</tr>
<tr>
<td>$k_c$</td>
<td>$k_t(1-\phi)+k_w\theta_l\phi+k_g(1-\theta_l)\phi$</td>
<td>$W/(mK)$</td>
<td>Sock thermal conductivity</td>
</tr>
<tr>
<td>$k_g$</td>
<td>$k_{da}+k_v$</td>
<td>$W/(mK)$</td>
<td>Gas thermal conductivity</td>
</tr>
<tr>
<td>$k_{da}$</td>
<td>$0.024+7.73 \times 10^{-5}T-2.610^{-8}T^2$</td>
<td>$W/(mK)$</td>
<td>Dry air thermal conductivity</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>$\frac{\theta_2-\theta_1}{\theta_3-\theta_1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_v$</td>
<td>$P h_r$</td>
<td>$kPa$</td>
<td>Partial vapour pressure</td>
</tr>
<tr>
<td>$h_r$</td>
<td>$e^{\exp\left(\frac{(P_{atm}\psi_m)M_t}{mR_T}\right)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_h$</td>
<td>$K_sK_r(\Theta)$</td>
<td>$m/s$</td>
<td>Hydraulic conductivity</td>
</tr>
<tr>
<td>$K_r$</td>
<td>$\Theta^{0.5}(1-(1-\Theta^{1/m})^2)$</td>
<td>$m/s$</td>
<td>Residual hydraulic conductivity</td>
</tr>
<tr>
<td>$m$</td>
<td>$1-1/n$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\psi_h$</td>
<td>$\psi_m + \psi_z$</td>
<td>$m$</td>
<td></td>
</tr>
<tr>
<td>$\psi_z$</td>
<td>$(z-z_0)$</td>
<td>$m$</td>
<td></td>
</tr>
<tr>
<td>$\psi_m$</td>
<td>$-(\Theta^{-1/m}_0-1)^{1/n}$</td>
<td>$m$</td>
<td></td>
</tr>
<tr>
<td>$D_v$</td>
<td>$D_v(P_0/P_{atm})(T_c/T_a)^{1.75}$</td>
<td>$m^2/s$</td>
<td>Vapour diffusivity</td>
</tr>
<tr>
<td>$\xi_g$</td>
<td>0.66 $\theta_a$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_v$</td>
<td>$\rho_v\theta_a$</td>
<td>$kg/(m^3)$</td>
<td>Vapour mass concentration</td>
</tr>
<tr>
<td>$\rho_v$</td>
<td>$\frac{P_{sat}M_t}{RT_c}$</td>
<td>$kg/m^3$</td>
<td>Water vapour density</td>
</tr>
<tr>
<td>$\theta_a$</td>
<td>$\phi-\theta_l$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_g$</td>
<td>$\frac{P_{sat}M_g}{RT_c}$</td>
<td>$kg/m^3$</td>
<td>Gas density</td>
</tr>
<tr>
<td>$Q_{ed}$</td>
<td>$Q_{eds}+Q_{ews}$</td>
<td>$W/m^2$</td>
<td>Evaporation heat losses</td>
</tr>
<tr>
<td>$Q_{eds}$</td>
<td>$0.782 \left(T-\left(286.31+\frac{P_{atm}}{0.256}\right)\right)$</td>
<td>$W/m^2$</td>
<td>Heating losses due to implicit sweat secretion</td>
</tr>
<tr>
<td>$Q_{ews}$</td>
<td>$h_c s_h 4.275 \left(\frac{Q_{eds}}{0.782}\right)$</td>
<td>$W/m^2$</td>
<td>Heating losses due to explicit sweat secretion</td>
</tr>
<tr>
<td>$P_{v_{atm}}$</td>
<td>$P_{v_{sat}}H_r$</td>
<td>$kPa$</td>
<td>Environmental water vapour pressure</td>
</tr>
<tr>
<td>$P_{v_{sat}}$</td>
<td>$0.611 \exp\left(\frac{17.27(T_c-273)}{273.3+(T_c-273)}\right)$</td>
<td>$kPa$</td>
<td>Environmental water vapour saturation pressure</td>
</tr>
</tbody>
</table>

**Table 1:** Variables defined to eased model understanding (in order of appearance in the paper)
### Table 2: Constants defined in the model (in order of appearance in the paper).

<table>
<thead>
<tr>
<th>Param.</th>
<th>Value</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_f$</td>
<td>0.5</td>
<td>$W/(m,K)$</td>
<td>Thermal conductivity of the internal tissue of the foot</td>
</tr>
<tr>
<td>$k_s$</td>
<td>0.2</td>
<td>$W/(m,K)$</td>
<td>Skin thermal conductivity</td>
</tr>
<tr>
<td>$\rho_b$</td>
<td>1060</td>
<td>$kg/m^3$</td>
<td>Blood density</td>
</tr>
<tr>
<td>$c_b$</td>
<td>3850</td>
<td>$J/(kg,K)$</td>
<td>Blood specific heat</td>
</tr>
<tr>
<td>$\omega_b$</td>
<td>$0.54 \times 10^{-3}$</td>
<td>$s^{-1}$</td>
<td>Blood perfusion volumetric rate</td>
</tr>
<tr>
<td>$T_b$</td>
<td>310.15</td>
<td>$K$</td>
<td>Blood temperature</td>
</tr>
<tr>
<td>$Q_m$</td>
<td>300</td>
<td>$W/m^3$</td>
<td>Metabolic heat of the foot tissues</td>
</tr>
<tr>
<td>$H_{lat}$</td>
<td>$2.25 \times 10^6$</td>
<td>$J/kg$</td>
<td>Water condensation conductivity</td>
</tr>
<tr>
<td>$k_t$</td>
<td>0.06</td>
<td>$W/(m,K)$</td>
<td>Fibers thermal conductivity</td>
</tr>
<tr>
<td>$k_w$</td>
<td>$2.7 \times 10^{-1}$</td>
<td>$W/(m,K)$</td>
<td>Liquid water thermal conductivity</td>
</tr>
<tr>
<td>$k_v$</td>
<td>$2.5 \times 10^{-2}$</td>
<td>$W/(m,K)$</td>
<td>Water vapour thermal conductivity</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.5</td>
<td>$m^3/m^3$</td>
<td>Sock porosity</td>
</tr>
<tr>
<td>$\rho_l$</td>
<td>1000</td>
<td>$kg/m^3$</td>
<td>Liquid water density</td>
</tr>
<tr>
<td>$\theta_r$</td>
<td>0.095</td>
<td>$m^3/m^3$</td>
<td>Sock residual saturation</td>
</tr>
<tr>
<td>$\theta_s$</td>
<td>0.41</td>
<td>$m^3/m^3$</td>
<td>Maximum sock saturation</td>
</tr>
<tr>
<td>$n$</td>
<td>1.31</td>
<td></td>
<td>Shape parameter of water retention curve</td>
</tr>
<tr>
<td>$K_s$</td>
<td>$7.22 \times 10^{-7}$</td>
<td>$m/s$</td>
<td>Saturated hydraulic conductivity</td>
</tr>
<tr>
<td>$g$</td>
<td>9.82</td>
<td>$m/s^2$</td>
<td>Gravity acceleration</td>
</tr>
<tr>
<td>$z_0$</td>
<td>0</td>
<td>$m$</td>
<td>Reference height for gravitational potential</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.019</td>
<td>$m^{-1}$</td>
<td>Shape parameter of water retention curve</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>$10^{-14}$</td>
<td>$m^2$</td>
<td>Air absolute permeability</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$1.7 \times 10^{-5}$</td>
<td>$kg/(m,s)$</td>
<td>Air viscosity</td>
</tr>
<tr>
<td>$D_{so}$</td>
<td>$2.92 \times 10^{-5}$</td>
<td>$m^2/s$</td>
<td>Vapour diffusivity (at 0°C and 1 bar)</td>
</tr>
<tr>
<td>$P_0$</td>
<td>101</td>
<td>$kPa$</td>
<td>Reference pressure</td>
</tr>
<tr>
<td>$P_{atm}$</td>
<td>101</td>
<td>$kPa$</td>
<td>Atmospheric pressure</td>
</tr>
<tr>
<td>$R$</td>
<td>8.31</td>
<td>$J/mol,K$</td>
<td>Ideal Gases constant</td>
</tr>
<tr>
<td>$h_e$</td>
<td>10</td>
<td>$W/(m^2,K)$</td>
<td>Convection coefficient</td>
</tr>
<tr>
<td>$s_h$</td>
<td>$s_h \in [0, 1]$</td>
<td></td>
<td>Skin humidity</td>
</tr>
<tr>
<td>$T_a$</td>
<td>298</td>
<td>$K$</td>
<td>Environmental temperature</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>0.97</td>
<td></td>
<td>Skin emissivity</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>$5.67 \times 10^{-8}$</td>
<td>$W/(m^2,K^4)$</td>
<td>Stefan-Boltzman constant</td>
</tr>
<tr>
<td>$H_r$</td>
<td>0.35</td>
<td></td>
<td>Environmental relative humidity (on a per unit basis)</td>
</tr>
<tr>
<td>$M_l$</td>
<td>18</td>
<td>$g/mol$</td>
<td>Water molecular mass</td>
</tr>
<tr>
<td>$M_g$</td>
<td>18</td>
<td>$g/mol$</td>
<td>Gas molecular mass</td>
</tr>
<tr>
<td>$k_{atm}$</td>
<td>0.137</td>
<td>$m/s$</td>
<td>Mass conductivity at the interface sock-air</td>
</tr>
<tr>
<td>$k'_{atm}$</td>
<td>0.001</td>
<td>$m/s$</td>
<td>Gas mass conductivity at the interface sock-air</td>
</tr>
</tbody>
</table>
A COUPLED THERMO-FLEXIBLE MULTI-BODY APPROACH FOR VIRTUAL RIG TESTING OF BRAKE DISCS

N. STRÖMBERG
Jönköping University
Department of Mechanical Engineering
SE-551 11 Jönköping, Sweden
e-mail: niclas.stromberg@jth.hj.se, web page: http://www.fema.se

Key words: Multi-Body, Thermo-mechanical, Brake discs

Abstract. In this paper a coupled thermo-flexible multi-body model of a test rig including a brake disc with pads is proposed. The dynamics of the system is modeled by two rotors connected via a torsional spring. The rotors represent the brake disc and the flywheel of the rig. The brake moment acting on the disc is governed by thermo-mechanical finite element analysis of the frictional contact between the pad and the disc. The finite element model is formulated by treating the disc in an Eulerian framework. The frictional power is calculated most accurately by solving Signorini’s contact conditions and Coulomb’s law of friction by using the augmented Lagrangian approach. This is done by including thermal expansion of the pad and the disc in the equilibrium equations and using a temperature dependent friction coefficient. The heat due to the frictional power is then transported with convection defined by the angular velocity of the disc. Distortions in the solution due to the non-symmetry of the convection matrix is stabilized by using the streamline-upwind approach. Summarized, the proposed virtual test rig of brake discs is modeled by three system of equations, i.e. the dynamic equations of the multi-body system, the energy balance of the pad and the disc, and the frictional contact of the pad and the disc. In order to obtain a robust and efficient approach, these three equation systems are solved sequentially by using Newmark’s method, the trapezoidal approach and Newton’s method. In such manner, temperatures, brake power and angular velocities can be generated accurately at low computational costs for different braking scenarios. This is demonstrated for a real pad-disc system to a heavy truck.

1 INTRODUCTION

In the development of disc brake systems physical rig testing of these systems is most important in order to meet design requirements. For disc brake systems to heavy trucks,
the rigs for performing these experiments are large investments for the developers and consequently the tests are expensive. In addition, a test usually takes long time to set up and perform. Thus, to study several design proposals using physical rig tests quickly becomes very costly. Therefore, it would be very beneficial if virtual rig tests could be performed during the development process. In this paper, a thermo-flexible multi-body approach for performing such tests is suggested and implemented.

The dynamics of the rig is represented by a flexible multi-body model. The disc and the flywheel are modeled by two mass moments of inertia and are connected via a torsional spring. The brake moment on the disc is obtained by using a thermo-mechanical finite element model including frictional contact. Recently, Strömberg [1] developed a Eulerian framework for this class of problems. Instead of formulating the disc in a Lagrangian frame, the disc material flows through a fixed mesh, where the convective terms are defined by the angular velocity from the flexible multi-body model. Due to the fixed grid, a node-to-node contact formulation between the disc and the pad can easily be established. The contact between the disc and the pad is then solved by the augmented Lagrangian formulation using Newton’s method. One of the first implementation of this approach can be found in Strömberg [2].

A similar work as presented herein, but without temperatures, can be found in Klarbring et al. [3], where the dynamic transmission error in spur gears was studied.

![Diagram](image)

**Figure 1:** The thermo-flexible multi-body model of a test rig for disc brakes.

2 A GLOBAL FLEXIBLE DYNAMIC MODEL

Let us study the system depicted in Figure 1. The system consists of two rotors, one brake pad and a torsional spring with a stiffness $k$. The first rotor, with mass moment of inertia $J_1$, represents the brake disc and the second rotor with mass moment of inertia $J_2$ is the flywheel of the test rig. The two rotors are connected via the torsional spring and
the kinematics of the rotors are defined by the angles \( \theta_i = \theta_i(t) \) along the \( e_3 \)-direction as functions of time \( t \). A brake force \( F = F(t)e_3 \) is applied on the pad which results in frictional contact between the pad and the brake disc, which in turn implies a brake moment. The frictional contact is modeled most accurately by applying a coupled thermomechanical finite element analysis. Details about this analysis is presented in the next section.

The frictional heating and the corresponding thermal expansion will imply that the distribution of the frictional forces will change in time. This is included in the global flexible dynamic model by letting the brake moment \( M_b = M_b(t)e_3 \) acting on the brake disc be calculated using this distribution. That is, letting \( r^A \) be the radius to a contact node \( A \) and \( P^t_A \) be the friction force at this node,

\[
M_b = M_b(t) = \sum_{A=1}^{n_c} r^A P^t_A, \tag{1}
\]

where \( n_c \) is the number of contact nodes.

The dynamics of the system can now be formulated as

\[
J\ddot{\theta} + K\dot{\theta} = M, \tag{2}
\]

where

\[
J = \begin{bmatrix}
J_1 & 0 \\
0 & J_2
\end{bmatrix},
\]

\[
K = \begin{bmatrix}
k & -k \\
-k & k
\end{bmatrix},
\]

\[
M = \begin{bmatrix}
M_b(t) \\
0
\end{bmatrix}, \quad \dot{\theta} = \begin{bmatrix}
\dot{\theta}_1(t) \\
\dot{\theta}_2(t)
\end{bmatrix}.
\]

Equation (2) is treated in time by applying the average acceleration method. Letting \( \theta_n, \dot{\theta}_n, \ddot{\theta}_n \) and \( \theta_{n+1}, \dot{\theta}_{n+1}, \ddot{\theta}_{n+1} \) represent the approximations of \( \theta(t) \) at time \( t_n \) and \( t_{n+1} \), respectively, the procedure for this is governed by

\[
\begin{bmatrix}
J + \frac{\Delta t^2}{4} K
\end{bmatrix} \ddot{\theta}_{n+1} = M_{n+1} - \dot{M}_n, \tag{4}
\]

where

\[
\dot{M}_n = K\theta_n + \Delta t K\dot{\theta}_n + \frac{\Delta t^2}{4} K\ddot{\theta}_n,
\]

\[
\dot{\theta}_{n+1} = \dot{\theta}_n + \frac{\Delta t}{2} (\ddot{\theta}_n + \ddot{\theta}_{n+1}), \tag{5}
\]

\[
\theta_{n+1} = \theta_n + \Delta t \dot{\theta}_n + \frac{\Delta t^2}{4} (\ddot{\theta}_n + \ddot{\theta}_{n+1}).
\]
3 A LOCAL THERMO-MECHANICAL MODEL

The friction forces appearing in (1) is determined by performing a thermo-mechanical finite element analysis as outlined in this section. This is done by formulating the brake disc in an Eulerian framework. In such manner heat is transported through the mesh by convection defined by the angular velocity $\dot{\theta}(t)$ of the disc. The heat balance in the disc is then governed by

$$C \dot{T} + (O + N + R)T = Q_c,$$

(6)

where $C$ is the heat capacity matrix, $O$ represents the conduction in the disc, $N = N(\dot{\theta})$ is the convection matrix and $R = R(\dot{\theta})$ is a matrix of artificial conduction in accordance to the streamline-upwind approach in order to stabilize distortions in the solution caused by the nonsymmetric convection matrix. Furthermore, $T$ is a vector of nodal temperatures and $Q_c = Q_c(P_n, T, \dot{\theta})$ is a vector of nodal heat fluxes at the contact surface representing the contact conductance and the frictional power.

The rates of the temperature vector are treated by applying the trapezoidal rule, i.e. the following approximation is utilized ($0 \leq \alpha \leq 1$):

$$T_{n+1} = T_n + \Delta t \left( (1 - \alpha) \dot{T}_n + \alpha \dot{T}_{n+1} \right),$$

(7)

where subscripts $n$ and $n + 1$ again represent approximations of states at times $t_n$ and $t_{n+1}$, respectively. Typically, we set $\alpha = 2/3$ according to the Galerkin approach.

Figure 2: The temperature dependent friction model.
The frictional power is governed by the normal contact forces \( P_n \), a temperature dependent friction coefficient \( \mu = \mu(T) \) and the tangential velocity \( r^A \dot{\theta}_1 \) at each contact node. That is, at a contact node \( A \), the frictional power can be written as
\[
Q_f^A = \mu(T^A)P_n^A r^A \dot{\theta}_1.
\]

The temperature dependent friction model is defined by the friction coefficient \( \mu_{\text{room}} \) at the room temperature \( T_{\text{room}} \), a peak value \( \mu_{\text{peak}} \) at a certain temperature \( T_{\text{peak}} \), and a lower bounded value \( \mu_{\text{bound}} \) at high temperatures above \( T_{\text{bound}} \). These assumptions are illustrated by Figure 2. A quadratic regression model,
\[
\mu(T) = \mu_0 + \mu_1 T + \mu_2 T^2,
\]
is then fitted according to the values obtained from experiments of \( \mu_{\text{room}}, T_{\text{room}}, \mu_{\text{peak}}, T_{\text{peak}}, \mu_{\text{bound}} \) and \( T_{\text{bound}} \).

The contact forces are obtained by solving equilibrium for linear elastic bodies with thermal expansion included. In addition, the contact forces are in accordance to Signorini’s unilateral contact conditions and Coulomb’s law of friction for global sliding. Equilibrium of the disc reads
\[
K \mathbf{d} + \dot{\mathbf{K}}T + (C_n + C_t(T))^T \mathbf{P}_n = \mathbf{F}_\omega,
\]
where \( \mathbf{d} \) is the displacement vector, \( K \) is the stiffness matrix, \( \dot{\mathbf{K}} \) represents the thermal expansion properties, \( C_n \) contains normal directions, \( C_t(T) \) is defined by the tangential sliding directions as well as the temperature dependent friction model in (9), and, finally, \( \mathbf{F}_\omega = \mathbf{F}_\omega(\theta_1) \) is a vector of centripetal forces.

Signorini’s contact conditions are treated by the well-known augmented Lagrangian approach. By letting \( \mathbf{d}_n = C_n \mathbf{d} \), this can be written as
\[
\mathbf{P}_n = \frac{\mathbf{P}_n + r \mathbf{d}_n + |\mathbf{P}_n + r \mathbf{d}_n|}{2},
\]
where \( r > 0 \) is a penalty coefficient and it is assumed that the initial gaps between the pad and the disc are negligible.

4 A SEQUENTIAL APPROACH

Summarized, for a given state defined by \( \mathbf{d}_n, \mathbf{P}_n, \mathbf{T}_n, \theta_n, \dot{\theta}_n, \ddot{\theta}_n \) at a time \( t_n \), our coupled thermo-flexible system is treated by solving the following equations sequentially in order to find the next state at time \( t_{n+1} \):
\[
\begin{align*}
\mathbf{h}(\mathbf{d}_{n+1}, \mathbf{P}_{n+1}; \mathbf{T}_n, \theta_n) &= 0, \\
J_{\text{eff}} \ddot{\theta}_{n+1} &= M_{\text{eff}}(\mathbf{P}_{n+1}, \theta_n, \dot{\theta}_n, \ddot{\theta}_n), \\
A(\mathbf{P}_{n+1}, \dot{\theta}_{n+1}) \mathbf{T}_{n+1} &= Q_{\text{eff}}(\mathbf{P}_{n+1}, \mathbf{T}_n, \dot{\theta}_{n+1}).
\end{align*}
\]
The first equation system in (12)\(_1\) represents the augmented Lagrangian formulation of the thermo-mechanical contact between the pad and the disc. This was presented in more details for the disc in (10) and (11). For given \(T_n\) and \(\dot{\theta}_n\), the solution to (12)\(_1\), defined by \(d_{n+1}\) and \(P_{n+1}\), is obtained by applying a non-smooth Newton method.

Figure 3: The sequential approach.

Figure 4: A heat band on the disc and the corresponding temperatures on the pad.
(12) is a compact formulation of (4) using

\[ J_{\text{eff}} = \left[ J + \frac{\Delta t^2}{4} K \right], \]

\[ M_{\text{eff}}(P_{n+1}, \theta_n, \dot{\theta}_n, \ddot{\theta}_n) = M_{n+1} - \dot{M}_n. \]  

Finally, (12) represents heat balance of the system, which was discussed previously in detail for the disc, see (6) and (7). One might notice that the contact conductances have here been included in \( A = A(P_{n+1}, \dot{\theta}_{n+1}) \) together with \( C, O, N \) and \( R \).

The sequentially procedure for solving the equations in (12) is also illustrated by Figure 3.

![Graphs](image-url)

(a) Clamping force  (b) Angular velocity

(c) Brake moment  (d) Maximum temperature

**Figure 5:** History of data generated by the toolbox.

5 NUMERICAL EXAMPLES

The thermo-flexible multi-body approach presented above is implemented in an in-house toolbox using Matlab and Fortran. A GUI is also developed such that a virtual rig test is easily set up by a user. An example of such a test is demonstrated in Figures 4 and 5. The test represents hard braking of a truck moving downhill. In a short time a
heat band of high temperatures develops on the disc surface. This is shown in Figure 4, where also a plot of the corresponding temperatures on the pad is given. In addition, the toolbox also generates histories of different data such as e.g. angular velocity, braking moment and maximum temperature, see Figure 5. In a near future, the virtual tests generated with the toolbox will be compared to real experiments. This will be discussed at the conference and is a topic of a forthcoming paper.

6 CONCLUDING REMARKS

In this work a new thermo-flexible multi-body approach for simulating rig tests of disc brakes is presented. The proposed method is robust and efficient. This is demonstrated by simulating a rig test of a real disc brake to a heavy truck.

REFERENCES


A LARGE DEFORMATION AND THERMOMECHANICALLY COUPLED INTERFACE APPROACH

M. KALISKE*, R. FLEISCHHAUER*, R. BEHNKE* AND I. M. ZREID*

*Institute for Structural Analysis
Technische Universität Dresden
01062 Dresden, Germany
e-mail: michael.kaliske@tu-dresden.de

Key words: thermomechanics, interface element, finite deformation

Abstract. Interfaces are formed e.g. by the contact surface of different materials of heterogeneous solids or by crack flanks within damaged bodies. Since the combination of temperature evolution and mechanical loadings influences significantly the deformation and thermal behavior of interfacial layers, these failure layers are thermally and mechanically described in the presented approach in a fully coupled sense. Thermomechanical interface descriptions can be used for prediction of crack propagation and, as soon as a designated failure layer exists, to predict the thermomechanical behavior of the observed solid. The presented interface approach for finite deformation introduces a consistent framework derived from principle thermodynamical laws.

1 INTRODUCTION

Interfacial layers are caused by e.g. the manufacturing process of heterogeneous solids, where they determine a designated failure layer between the different components of such a solid. Also crack flanks in homogeneous bodies determine an interface. In order to predict crack propagation within homogeneous solids or within heterogeneous solids, thermomechanical interface descriptions can be used (see [1]). The presented framework is introduced in a spatial setting of the thermodynamical balance equations. Consequently, the true mechanical and true thermal quantities are constitutively described. Bonds between the opening flanks are the main phenomena of a separation process. The transmission of tractions and heat is constitutively described in terms of elastic and inelastic thermomechanical behavior. The mechanical part of the material description is obtained by considering the equilibrium of forces within the bonds between the interface surfaces, while the thermal part of the constitutive formulations is derived from the balance of thermal energy within the interface. Finally, numerical examples are shown in order to demonstrate the capabilities of the element formulation.
2 KINEMATICS OF INTERFACES

Consider $B$ as the observed solid body having different subbodies $B_i$, for which $B_i \subset B$ holds. Let $\varphi : B \times \mathbb{R}^3 \to \mathbb{R}^3$ be the nonlinear deformation map of $B$ at material point $X \in B$ and time $t \in \mathbb{R}$. Therewith, $\varphi$ maps points $X \in B$ onto points $x = \varphi(X, t)$. $F := \partial_x \varphi(X, t)$ is the deformation gradient with the Jacobian $J := \det[F] > 0$. The velocity field is given with $\dot{\varphi} = v = \partial_t \varphi(X, t)$ and defines the spatial velocity gradient $l := \partial_x \dot{\varphi}(X, t)$. In terms of an interface between two subbodies $B_i$ and $B_k$, the boundaries $\partial B_i \supset \Omega$ and $\partial B_k \supset \Omega$ are piecewise identical along the connective interface $\Omega \subset B$. The opening displacement between two points $+(\in \Omega$ and $-\in \Omega$ can be described by the opening vector $\Delta = x^+ - x^-$, where $x^+ \in B_i$ and $x^- \in B_k$. The related points are initially connected having the same coordinates $X^+ \in B_i$, $X^- \in B_k$ at $\Omega^+_ref$ and $\Omega^-_ref$. Both points start to separate as soon as the bonding between them is smaller than its connective forces. The unified kinematics of not separating into tangential and normal directions are firstly introduced in [2].

3 THERMOMECHANICS OF AN INTERFACE AND RELATED FINITE ELEMENT EQUATIONS

In the following section, the continuum balance principles of linear momentum and energy are extended to the interfacial description. The expressions have to be supplemented by additional boundary terms related to $\Omega$.

3.1 Mechanical equilibrium

The balance of linear momentum

$$\rho \ddot{v} = \rho \ddot{b} + \text{div}(\sigma)$$

represents the mechanical equilibrium condition of any deformation process. The Galerkin method is applied on its spatial form

$$\int_B \delta u \cdot \text{div}(\sigma) \, dv - \int_B \delta u \cdot (\rho \ddot{u} - \rho \dot{b}) \, dv = 0 ,$$

where $\delta u$ is a test function, satisfying $\delta u := \{ \delta u \in B_0 | \delta u = 0 \text{ on } \partial B_0 \}$. Since the Cauchy stresses are symmetric and the divergence theorem $\text{div}(\sigma \delta u) = \text{div}(\sigma) \cdot \delta u + \sigma : \text{grad} (\delta u)$ as well as the Gauss theorem holds, one can express Eq. (2) as

$$\int_B \sigma : \text{grad} (\delta u) \, dv - \int_{\partial B} (\sigma \delta u) \cdot n \, da + \\ \int_B \delta u \cdot (\rho \ddot{u} - \rho \dot{b}) \, dv - \int_{\Omega^+} \delta u^+ \cdot t^+ \, d\Omega^+ - \int_{\Omega^-} \delta u^- \cdot t^- \, d\Omega^- = 0 .$$
The boundary term of Eq. (3) contains the external tractions \((\sigma \delta u \cdot n) = \delta u \cdot (\sigma n) = \delta u \cdot t\). Together with the body forces \(b\) as well as the inertia term \(\rho \ddot{u}\), the external forces are represented. The integrals over \(\Omega^+\) and \(\Omega^-\) are the contributions coming from the opening interface. The test function \(\delta u^+\) is defined at \(\Omega^+\) and the test function \(\delta u^-\) is defined at \(\Omega^-\). Subsequently, Eq. (3) has to be discretized, using e.g. linear ansatz functions \(N(\xi)\) for the shape of the elements, according to the isoparametric concept of the finite element method with local coordinates \(\xi\). The displacement field \(u = x - X\) is discretized according to \(u = N(\xi)dE\), where \(dE\) denotes the nodal displacements of one element. The test function is also discretized by \(\delta u = N(\xi)\delta dE\). The gradient of the test function is discretized by \(\text{grad}(u) = \partial_x N(\xi) dE\) = \(B(\xi)dE\). Now, the finite element equations of the discretized equilibrium condition can be formulated

\[
G_M = \sum_{E=1}^{N} \left[ \int_{\Omega_E^+} B^T \sigma \, dv + \left( \int_{\Omega_E} N^T \rho N \, dv \right) \delta dE - \int_{\partial \Omega_E} N^T (\rho \dot{b}) \, dv - \int_{\partial \Omega_E} N^T t_e \, da \right.
- \left. \int_{\Omega_E^-} N^T_t \, d\Omega^- \right] = \sum_{E=1}^{N} \left[ f_{int}^E + m^E \ddot{d}E - f_{ext}^E - f_{if}^E \right] = 0 \tag{4}
\]

The test function defined at the boundaries \(\Omega^+\) and \(\Omega^-\), are discretized by

\[
\delta u^+ = N(\xi_1, \xi_2, \xi_3 = +1)\delta dE \tag{6}
\]
\[
= N_{\Omega^+} \delta dE \tag{7}
\]
\[
\delta u^- = N(\xi_1, \xi_2, \xi_3 = -1)\delta dE \tag{8}
\]
\[
= N_{\Omega^-} \delta dE \tag{9}
\]

### 3.2 Thermal equilibrium

Before dealing with the thermal equilibrium, an evolution equation for the temperature is needed. This property can be obtained by the following briefly described steps. Starting, the spatial description of dissipation is defined \(\rho \gamma \theta \geq 0\). The dissipation has to be additively split into local \(D_{loc}\) and conductive \(D_{con}\) parts, \(D = D_{loc} + D_{con}\). It is common to require a strict positiveness for both parts. Therefrom, the Clausius-Planck inequality

\[
D_{loc} := \rho \gamma \dot{\theta} - \rho r + \text{div}(q) \geq 0 \tag{10}
\]

is introduced. One can incorporate the spatial energy balance

\[
\rho \dot{e} = \sigma : (gl) + \rho r - \text{div}(q) \tag{11}
\]
into Eq. (10) resulting in

\[ D_{\text{loc}} := \rho \theta \dot{\eta} - \rho \dot{e} + \sigma : (g l) \geq 0 . \]  

(12)

Following, the Helmholtz free energy \( \psi \) has to be defined through the partial Legendre transformation \( \psi := e - \theta \eta \), which implies certain dependencies of the Helmholtz free energy. One can consider a homogeneous inelastic material, where the free energy \( \psi = \psi(g, F, \theta, G, I) \) is a function of the temperature gradient \( G \), the deformation gradient \( F \), the absolute temperature \( \theta \) and a set of internal variables \( I \). The spatial metric \( g \) is needed in order to map deformation measures onto the reference configuration. Consequently, the time derivative of the free energy reads

\[ \dot{\psi} = \partial_{F} \psi : \dot{F} + \partial_{g} \psi : \dot{G} : I + \partial_{I} \psi : \dot{I} \geq 0 . \]  

(13)

Using the Legendre transformation, the time derivative of the internal energy yields the form

\[ \dot{e} := \dot{\psi} + \dot{\theta} \eta + \eta \dot{\theta} . \]  

(14)

Incorporating Eq. (14) into Eq. (12) results in

\[ \sigma : (g l) - \rho \dot{\psi} - \rho \eta \dot{\psi} \geq 0 . \]  

(15)

Next, the time derivative of the free energy (Eq. (13)) has to be plugged into Eq. (15). Using the definition of the spatial velocity gradient \( l = \dot{F} F^{-1} \) and the conservation of mass, \( J \cdot \rho = \rho_{0} \) as well as the relation between Cauchy and first Piola-Kirchhoff stress, \( \sigma = 1/J \cdot P F^{T} \), the resulting thermodynamic restriction is defined as

\[ [g P - \rho_{0} \partial_{F} \psi] : F - \rho_{0} [\eta + \partial_{g} \psi] : \dot{G} - \rho_{0} [\partial_{I} \psi] : \dot{I} \geq 0 . \]  

(16)

Following [3, 4], the thermodynamic restriction should be fulfilled for an arbitrary rate of deformation gradient, temperature and temperature gradient, which yields in Eq. (16),

\[ \rho_{0} \partial_{F} \psi = g P , \]  

(17)

\[ \eta = -\partial_{g} \psi , \]  

(18)

\[ \partial_{g} \psi = 0 . \]  

(19)

Considering Eqs. (17), (18) and (19), Eq. (10) reduces to

\[ D_{\text{loc}} := -\rho_{0} \partial_{I} \psi : \dot{I} \geq 0 . \]  

(20)

Furthermore, inserting Eq. (13) into Eq. (14) reads

\[ \dot{e} = \partial_{F} \psi : \dot{F} + \partial_{g} \psi : \dot{G} + \partial_{I} \psi : \dot{I} + \eta \dot{\theta} + \theta \dot{\eta} . \]  

(21)
Inserting this expression into the balance of energy, under the use of Eqs. (17), (18) and (19), the balance of entropy equation
\[
\rho \theta \dot{\theta} = -\rho \cdot [\partial_z \psi] : \mathbf{T} + \rho r - \text{div} \, q
\]  
(22)
is obtained. The last step is the insertion of the rate of entropy
\[
\dot{\theta} = - \partial^2_{\eta \theta} \psi \cdot \dot{\theta} - \partial^2_{\mathbf{F} \theta} \psi : \mathbf{F} - \partial^2_{\mathbf{Z} \theta} \psi : \mathbf{T}
\]  
(23)
into Eq. (22). Under consideration that the heat capacity is defined as \( c = -\theta \cdot \partial^2_{\eta \theta} \psi \), the transient heat conduction equation
\[
\rho c \dot{\theta} = -\text{div}(q) + \rho r + \rho \cdot [\theta \cdot \partial^2_{\mathbf{F} \theta} \psi] : \mathbf{F} - \rho \cdot [\partial_z \psi - \theta \cdot \partial^2_{\mathbf{Z} \theta} \psi] : \mathbf{T}
\]  
(24)
is formulated. Furthermore, one has to apply the Galerkin method to the transient heat conduction equation in order to achieve the finite element equations. Therefore, Eq. (24) is multiplied with an arbitrary test function \( \delta \theta \) satisfying \( \delta \theta := \{ \delta \theta \in \mathcal{B}_c | \delta \theta = 0 \text{ on } \partial \mathcal{B}_{c0} \} \). The integration over the current volume yields
\[
\int_{\mathcal{B}} \delta \theta \rho c \dot{\theta} \, dv - \int_{\mathcal{B}} \delta \theta r \, dv + \int_{\partial \mathcal{B}} \delta \theta \text{div}(q) \, dv - \int_{\mathcal{B}} \delta \theta w_{\text{ext}} \, dv + \int_{\mathcal{B}} \delta \theta w_{\text{int}} \, dv = 0 .
\]  
(25)
The divergence theorem \( (\delta \theta \text{div}(q)) = (\text{div}(\delta \theta q)) - (\text{grad}(\delta \theta) \cdot q) \) can be applied and, together with the Gauss theorem, results in
\[
\int_{\mathcal{B}} \delta \theta \rho c \dot{\theta} \, dv - \int_{\partial \mathcal{B}} \delta \theta r \, dv + \int_{\partial \mathcal{B}} \delta \theta \cdot (q \cdot n) \, da - \int_{\partial \mathcal{B}} \text{grad}(\delta \theta) \cdot q \, dv - \int_{\mathcal{B}} \delta \theta \cdot w_{\text{ext}} \, dv + \\
\int_{\Omega^+} \delta \theta \cdot w_{\text{int}} \, dv + \int_{\Omega^-} \delta \theta^+ \cdot (q^+ \cdot n^+ - w^+) \, d\Omega^+ + \int_{\Omega^-} \delta \theta^- \cdot (q^- \cdot n^- - w^-) \, d\Omega^- = 0 .
\]  
(26)
The external thermal conduction energy \( (q \cdot n) = h_e \) is identified as the boundary term. The additional contributions at the boundaries \( \Omega^+ \) and \( \Omega^- \) have to be considered, since they are caused by the opening crack flanks. The thermal conduction energies \( h^+ = (q^+ \cdot n^+) \) and \( h^- = (q^- \cdot n^-) \), as well as the power terms \( w^+ = w_{\text{ext}}^+ - w_{\text{int}}^+ \) and \( w^- = w_{\text{ext}}^- - w_{\text{int}}^- \) describe the energy of the surfaces \( \Omega^+ \) and \( \Omega^- \), since they are originated in heat flux and power terms of the bonds between the opening flanks. Furthermore, the temperature field is discretized as \( \theta = \mathcal{N}_\theta(\xi) \theta^E \), as well as the test function \( \delta \theta = \mathcal{N}_\theta(\xi) \delta \theta^E \). The test function defined at the boundaries \( \Omega^+ \) and \( \Omega^- \), are discretized by
\[
\delta \theta^+ = \delta \theta^E
\]  
(27)
\[
\delta \theta^- = \delta \theta^E
\]  
(28)
\[
\delta \theta^+ = \delta \theta^E
\]  
(29)
\[
\delta \theta^- = \delta \theta^E
\]  
(30)
Finally, one can formulate the finite element equations of the thermal equilibrium

\[
G_T = \sum_{E=1}^{N} \left[ \int_{\Omega^+} N^T_{\theta \rho c \dot{\theta}} \, dv - \int_{\Omega^+} N^T_{\theta r} \, dv + \int_{\Omega^+} N^T_{\theta h_e} \, da \\
- \int_{\Omega^+} B^T_{\theta} q \, dv - \int_{\Omega^+} N^T_{\theta w_{\text{ext}}} \, dv + \int_{\Omega^+} N^T_{\theta w_{\text{int}}} \, dv \\
+ \int_{\Omega^-} N^T_{\theta \rho c \dot{\theta}} (q^+ \cdot n^+ - w^+) \, d\Omega^+ + \int_{\Omega^-} N^T_{\theta r} (q^- \cdot n^- - w^-) \, d\Omega^- \right] = 0 .
\]

4 THERMAL ENERGY AND TRACTION-SEPARATION-LAW (TSL)

After having derived the finite element equations, the constitutive descriptions of an interface can be defined. The main mechanism of debonding is the development of connective bonds between opening interface surfaces as depicted in Fig. 1. In order to have equilibrium of forces (see Fig. 1(a)) within the interface element, \( t^+ := -t \cdot J_A^+ \) and \( t^- := +t \cdot J_A^- \) are introduced. The traction quantities \( t^+ \) and \( t^- \) can have different values on the opposite surfaces \( \Omega^+ \) and \( \Omega^- \), since in general different changes of the area of these surfaces are allowed. This is achieved by scaling the current tractions \( t \) according to the change of their related differential area. The area change is described by \( J_A^+ = d\Omega_{\text{ref}}^+/d\Omega^+ \) and \( J_A^- = d\Omega_{\text{ref}}^-/d\Omega^- \) and can be generally determined from the area map \( n \, da = J F^{-T} N \, dA \) and the current normal definition \( n = F^{-T} N \), that leads to the
expressions
\[ n \, da = J F^{-T} N \, dA = J n \, dA , \]  
\[ J_A = \frac{dA}{da} = \frac{1}{J} . \]  

Therewith, the condition for the equilibrium of forces
\[ f^+ = - f^- , \]  
\[ \int_{\Omega^+} t^+ d\Omega^+ = \int_{\Omega^-} t^- d\Omega^- \]  
is fulfilled for the same initial area \( \Omega_{ref} \) of both surfaces. It is possible to think of the current traction vectors \( t^+ \) and \( t^- \) as Cauchy tractions at the crack surfaces, while the transformation with \( J_A^+ \) and \( J_A^- \) results in Kirchhoff tractions \( +t \) and \( -t \). This transformation yields a change of the integration areas, while evaluating the reaction forces at the ends of a delamination bond (compare Fig. 1(a)).

Considering an opened crack within the observed body \( B \), the crack surfaces as well as the bonds between them undergo different thermal conduction and convection mechanisms, respectively. The thermal convection at the crack surfaces, that transfers heat between the subbodies \( B_i \) and the surrounding environment is not considered in the proposed element formulation. The thermal conduction through the connective bonds between the crack surfaces is denoted by a zero sum of the total heat in- and outflux of the \( \Omega^+ \) and the \( \Omega^- \) surfaces, due to a zero loss or gain of thermal heat flux energy on an interface element. Therefore, the balance of thermal conduction energy (see Fig. 1(b)) is obtained according to
\[ h^+ = - h^- , \]  
\[ \int_{\Omega^+} (q^+ \cdot n^+) d\Omega^+ = \int_{\Omega^-} (q^- \cdot n^-) d\Omega^- . \]  
The balance condition in Eq. (37) can be fulfilled for \( (q \cdot n)^+ := -(q \cdot n) \cdot J_A^+ \) and \( (q \cdot n)^- := (q \cdot n) \cdot J_A^- \). Similar to the traction quantities \( t^+ \) and \( t^- \), the thermal conduction quantities \( (q \cdot n)^+ \) and \( (q \cdot n)^- \) are spatial objects. The constitutive descriptions of the introduced quantities \( t^+ \), \( t^- \) and \( (q \cdot n)^+ \), \( (q \cdot n)^- \) are given in the following subsections.

4.1 Traction-separation-laws
The tractions are defined as \( t^+ := - t \cdot J_A^+ \) and \( t^- := + t \cdot J_A^- \). The relation between the spatial tractions \( t \) and current separation \( \Delta \) can be obtained from a TSL-potential.
4.1.1 Elastic TSL-potential

The proposed elastic potential is an extension of the one in [5] and it is defined per unit reference area. The specific form of the proposed potential reads

$$\Phi(\Delta, [\theta]) := \phi - \frac{\phi}{\delta} (\| \Delta \| + \delta) \exp \left( -\frac{\| \Delta \|}{\delta} \right) - \frac{1}{2} \alpha_t \phi \delta^2 \| \Delta \|^2 , \quad (38)$$

where \( \phi \) is a material parameter, representing the work or energy of separation, \( \delta \) is a characteristic opening length and \( \alpha_t \) is the thermal expansion coefficient. To derive the TSL, the derivative \( \partial_{\Delta} \Phi(\Delta, [\theta]) \) is needed. In case of Eq. (38), it reads

$$t(\Delta, [\theta]) = \partial_{\Delta} \Phi(\Delta, [\theta]) = \frac{\phi}{\delta^2} \exp \left( -\frac{\| \Delta \|}{\delta} \right) \Delta - \alpha_t \phi \delta^2 [\theta] \Delta . \quad (39)$$

4.1.2 Inelastic TSL-potential

The inelastic traction separation law is derived from a viscoelastic potential \( \Phi = \Phi^e(\theta, \Delta) + \Phi^v(\theta, \Delta, \Delta^v) \), that consists of an elastic \( \Phi^e \) part and a viscous \( \Phi^v \) part. The total separation \( \Delta = \Delta^e + \Delta^v \) is split into an elastic and a viscous part. Consequently, the total tractions are determined by an addition of the elastic and viscous tractions, \( t = t^e + t^v \). The elastic part \( \Phi^e(\theta, \Delta) \) of the TSL potential reads

$$\Phi^e(\Delta, \theta) = \phi - \frac{\phi}{\delta} (\| \Delta \| + \delta) \exp \left( -\frac{\| \Delta \|}{\delta} \right) \exp \left( -\vartheta \frac{\theta}{\theta_b} \right) , \quad (40)$$

Similar to Eq. (38), \( \phi \) denotes the work of separation and \( \delta \) is the characteristic opening length. The change of the temperature \( \vartheta = \theta - \theta_b \) relative to the reference temperature \( \theta_b \) at the middle of the interface is related to characteristic temperature \( \theta_b \) at the middle of the interface. The viscous part of the inelastic TSL is given as

$$\Phi^v(\Delta, \Delta^v, \theta) = \frac{1}{2} \beta \phi \delta^2 \exp \left( -\frac{\| \Delta \|}{\delta} \right) \exp \left( -\vartheta \frac{\theta}{\theta_b} \right) \| \Delta - \Delta^v \|^2 , \quad (41)$$

where \( \beta \) is a material parameter representing the relation between the elastic and the viscous part of the work of separation. The elastic tractions can be derived from the elastic part of \( \Phi \) and read

$$t^e(\Delta, \theta) = \frac{\partial}{\partial \Delta} \Phi^e = \frac{\phi}{\delta^2} \exp \left( -\frac{\| \Delta \|}{\delta} \right) \exp \left( -\vartheta \frac{\theta}{\theta_b} \right) \Delta , \quad (42)$$

while the viscous tractions are given as

$$t^v(\Delta, \Delta^v, \theta) = -\frac{\partial}{\partial \Delta^v} \Phi^v = \beta \phi \delta^2 \exp \left( -\frac{\| \Delta \|}{\delta} \right) \exp \left( -\vartheta \frac{\theta}{\theta_b} \right) (\Delta - \Delta^v) . \quad (43)$$
The motivation for the chosen viscous potential and the resulting tractions is, to achieve the viscous tractions as a function of the elastic tractions. It can be seen that
\[ t^v = \beta \| t^e \| \frac{(\Delta - \Delta^v)}{\| \Delta \|}. \] (44)

The characteristic length \( \delta \) is calculated from the material parameters, according to
\[ \delta = \phi \cdot \exp \left( \frac{-\vartheta}{\theta_b} \right) T_{\text{max}} \cdot \exp (1). \] (45)

The evolution law for the viscous part of the separation reads
\[ \dot{\Delta}^v = \frac{\delta}{\eta} t^v, \] (46)

where \( \eta \) is the viscosity and the characteristic length \( \delta_v \) for the viscous part is determined by
\[ \delta_v = \beta \delta. \] (47)

The viscosity is temperature dependent and may be described by Williams-Landel-Ferry equation
\[ \eta = \eta_0 \left( \frac{-C_1 (\theta - \theta_r)}{C_2 + \theta + \theta_r} \right), \] (48)

where \( \eta_0 \) is the viscosity at the reference temperature and \( C_1 \) and \( C_2 \) are additional material parameters.

4.2 Thermal conduction energy and rate of work

As stated in Eq. (37), \((q \cdot n)^+\) and \((q \cdot n)^-\) have to be expressed. Additionally, the rate of work done by the tractions at the \(\Omega^+\) and \(\Omega^-\) surfaces is derived. The proposed formulations of the thermal conduction energy per unit reference areas are given by
\[ (q \cdot n)^+ := - (q \cdot n) \cdot J_A^+ = - \left( -k \frac{\| \theta \|}{\| \Delta \|} \right) \cdot J_A^+, \] (49)
\[ (q \cdot n)^- := (q \cdot n) \cdot J_A^- = \left( -k \frac{\| \theta \|}{\| \Delta \|} \right) \cdot J_A^-. \] (50)

One can assume that the heat is mainly transferred along the connective bonds, as soon as the separation process between the two materials has started \(\| \Delta \| > 0\) and as soon as a difference between the temperatures of both sides \(\| \theta \| = \theta^+ - \theta^-\) at the initially connected points is found. The introduced formulations of Eq. (49) and Eq. (50) are phenomenological descriptions. The computation of the derivatives \(\partial_{\| \theta \|} (q \cdot n)\) and \(\partial_{\Delta} (q \cdot n)\)
are required for the incorporation of the element formulation into a FE-framework and are given as

$$ \partial_\Delta (q \cdot n) = \left( k \theta \| \Delta \|^3 \right), $$

(51)

$$ \partial_{\theta t} (q \cdot n) = \left( -k \frac{1}{\| \Delta \|} \right). $$

(52)

The total rates of work \( w^+ = w_{ext}^+ - w_{int}^+ \) and \( w^- = w_{ext}^- - w_{int}^- \) at the opposite flanks of the interface are the result of the subtraction of the internal power from the external power (compare Eq. (24)). Their specific form in terms of the elastic TSL potential reads

$$ w^+ := \frac{1}{2} w^{Dc} t^+ \Delta J^+ = \frac{1}{2} \left[ \theta \right] \left[ \partial_{\theta t} t \right] \cdot \Delta J^+, $$

(53)

$$ w^- := -\frac{1}{2} w^{Dc} t^- \Delta J^- = \frac{1}{2} \left[ \theta \right] \left[ \partial_{\theta t} t \right] \cdot \Delta J^-, $$

(54)

and is assigned equally with the coefficient 1/2 to the sides \( \Omega^+ \) and \( \Omega^- \) in a phenomenological way. The contributions of the rates of work in terms of the viscoelastic TSL potential are given as

$$ w^+ := \frac{1}{2} \theta \left[ \partial_\Delta^2 \Phi \right] \cdot \Delta J^+ = \frac{1}{2} \left[ \partial_\Phi^v \Phi^v - \partial_\Delta^2 \Phi^v \right] \cdot \Delta^v J^+, $$

(55)

$$ w^- := \frac{1}{2} \theta \left[ \partial_\Delta^2 \Phi \right] \cdot \Delta J^- = -\frac{1}{2} \left[ \partial_\Phi^v \Phi^v - \partial_\Delta^2 \Phi^v \right] \cdot \Delta^v J^-, $$

(56)

5 NUMERICAL EXAMPLES

Subsequently, two numerical model problems are given for the verification of the proposed element formulation.

5.1 Thermal conduction

The first example is a pure thermal conduction simulation, where the top surface is loaded with a constant temperature rate of \( \dot{\vartheta}_e = 1.5 \) K/s for a time of 100 s and, subsequently, the temperature is kept constant for the next 300 s. At the bottom, the thermal boundary condition is a free surface. The solid elements are thermomechanical linear elements. The geometry of 10x10x20 mm is discretized with 128 solid elements and 16 of the proposed thermomechanical elastic interface elements at the middle of the height of the specimen. The mechanical boundary conditions are set such that a stress free state is ensured. The results of the computation containing interface elements are illustrated in Fig. 2, where Fig. 2(a), Fig. 2(b) and Fig. 2(c) contain the graphical illustration of the change in the temperature field at three different times for the computation with the proposed thermoelastic interface elements. Interpreting the obtained results, it is concluded, that the thermal conductivity behavior of the elastic interface element is expressed in a correct manner.
Figure 2: Illustration of thermal conduction simulation

5.2 Predictive tensile simulation

The second example is a tensile test simulation. The geometry of the specimen of 10x10x20 mm is discretized with 54 thermomechanical linear solid elements containing thermoviscoelastic Neo-Hooke material. In the middle of the height of the specimen 9 thermoviscoelastic interface elements are placed. The bottom surface is totally fixed, while a constant rate of displacement of 0.16 mm/s is applied to the top surface for 25 s and afterwards held constant for 75 s. The resulting stress relaxation behavior is depicted in Fig. 3, where the initial stress free configuration is depicted in Fig. 3(a), the maximum stress situation is shown in Fig. 3(b) and relaxation of stress can be seen from Fig. 3(c). The stress relaxation within the interface generates an increase in the opening of the interface. Therefrom, a physically correct behavior of the proposed element and constitutive formulation is concluded.

6 CONCLUSIONS

- A thermomechanically coupled interface element formulation in the spatial setting is introduced.

- Elastic and inelastic thermomechanical TSL-potentials for deriving tractions and internal, as well as external work terms are introduced.

- Numerical examples are shown in order to demonstrate the potential capabilities of the proposed formulations.
Figure 3: Illustration of predictive tensile test simulation

REFERENCES


A MULTISCALE METHOD TO ANALYZE THE DETERIORATION DUE TO ALKALI SILICA REACTION CONSIDERING THE EFFECTS OF TEMPERATURE AND RELATIVE HUMIDITY

Tao Wu*, İlker Temizer†, Peter Wriggers*

*Institute of Continuum Mechanics
Leibniz Universität Hannover
Appelstraße 11, 30167 Hannover, Germany
e-mail: wu,wriggers@ikm.uni-hannover, www.ikm.uni-hannover.de/

†Department of Mechanical Engineering
Bilkent University
06800 Bilkent Ankara, Turkey
e-mail: temizer@bilkent.edu.tr, www.me.bilkent.edu.tr

Key words: Multiscale, Coupling, Concrete, Alkali Silica Reaction

Abstract. This work presents a three-dimensional multiscale framework to investigate the deterioration resulting from alkali silica reaction (ASR) in the concrete. In this contribution, 3D micro-CT scan of hardened cement paste (HCP) and aggregates with a random distribution embedded in a homogenized cement paste matrix represent the microscale and mesoscale of the concrete respectively. A 3D hydro-chemo-thermo-mechanical model based on staggered method is developed at the mesoscale of the concrete, yet taking into account the deterioration at the microscale due to ASR.

1 INTRODUCTION

Concrete is an extremely complex heterogeneous material with random microstructures at different length scales. At the macroscale level, it is treated as a homogeneous material. While going down one scale, the mesoscale includes a binding matrix, aggregates and pores with broad size distributions as well as interfacial zones between the aggregates and the matrix. At a lower level, the microscale is represented by the microstructure of HCP, which is comprised of hydration products, unhydrated residual clinker and micropores [1,2].

ASR discovered in the 1940s, is a long-term chemical reaction, while it is detrimental to the concrete structure. ASR is one chemical reaction between reactive forms of silica
in the aggregates and alkali ions in the pore solution. The gel as the reaction product can swell in the presence of water and produce internal stresses, which leads to the formation of the micro-crack network in the concrete \([3,4,5,6]\).

### 2 Kinetics of Chemical Reaction

Chemical extent \(\xi\) based on first-order kinetics is used to describe the progression of ASR, where \(\xi = 1\) denotes the beginning of ASR and \(\xi = 0\) indicates the end of ASR. The explicit equation of the chemical extent \(\xi\) firstly explained in \([6]\) is expressed through

\[
\xi(t) = \frac{1 - \exp(-t/\tau_{ch})}{1 + \exp(-t/\tau_{ch} + \tau_{lat}/\tau_{ch})}
\]  

(2.1)

where \(\tau_{lat}\) is the latency time and \(\tau_{ch}\) is the characteristic time, respectively corresponding to the initiation and the development period of ASR from a practical point of view \([4,5,6]\). In addition, Larive \([6]\) addressed the influences of the temperature and the relative humidity on \(\tau_{lat}\) and \(\tau_{ch}\). The dependencies of the latency time and the characteristic time on the temperature and the relative humidity are illustrated in Fig.1(a) and Fig.(b) respectively.

![Figure 1](image)

**Figure 1:** (a) Latency time with function of temperature and relative humidity (b) characteristic time with function of temperature and relative humidity.

### 3 Analysis of ASR at Microscale and Mesoscale

In this work, the cement specimen with an length of 1750 \(\mu m\) was conducted to obtain the microstructural geometry of HCP using three-dimensional micro-CT scans with a resolution of 1\(\mu m\). Therefore, the micro-CT scan of HCP is comprised of 1750\(^3\) data points, where each point corresponds to a voxel of 1\(\mu m^3\). The natural element to use within
the finite element method to discretize the microstructure is an 8-node brick where each element is assigned to a single material phase [1,2]. Many continuum damage models can depict the failure of HCP, but the simplified constitutive model developed by Hain and Wriggers [1,2] is employed in this work to save the computational cost, due to too complex three-dimensional micro-structural geometry of HCP. In addition, it is only feasible to define the simplified damage model in hydration products because of its high volume fraction in the HCP. The observation that the chemical property of the gel is similar to calcium-silicate-hydrate (C-S-H), enables the gel to be conceived as an incompressible material with the poission ratio of 0.49975. Fig.2 illustrates the distribution and evolution of the damage in HCP with the extent of ASR, where the damage occurs in the hydration products in the vicinity of micropores triggered by the expansion of the gel in micropores. Computational homogenization is an efficient tool to link macro-microscale of the

Figure 2: (a)-(d) damage distribution on cross section of HCP with various values of chemical extent $\xi$, (f) material distribution on cross-section of HCP.
material [7]. In this contribution, 30 randomly obtained representative volume elements (RVEs) from micro-CT scan of HCP are performed and the resulting effective damage of statistical tests due to the expansion of the gel through computational homogenization approach with respect to days and chemical extent are shown in Fig.3(a) and Fig.3(b) respectively. The correlations between chemical extent and effective damage in Fig.3(b) can be approximated by polynomial curve, which can be directly applied to the next length scale.

The take-and-place method is employed to generate the three-dimensional geometrical representation of concrete at the mesoscale, where randomly distributed aggregates are embedded in the homogenized HCP [8]. Aggregates are assumed to be elastic material in the present work, and a visco-plastic model of the classical PERZYNA-type combined with an isotropic damage is defined for HCP, in order to capture the experimentally nonlinear phenomenon of the concrete [1,2].

4 Coupling

It is observed that ASR is depending on the temperature and the relative humidity from experiments in [6] and the associated influences are explained. The small size of the microstructure can skip over the hydro-thermo-chemo coupling work at the microscale, and the correlation between effective damage and chemical extent $\xi$ is obtained through computational homogenization, which can be applied at the next scale directly. However, the hydro-thermo-chemo coupling has to be carried out at the mesoscale due to its relatively large size, so that the influences of transient temperature and relative humidity on the chemical extent $\xi$ can be analyzed. The obtained chemical extent at the mesoscale enables to obtain the deterioration due to ASR originating from microscale through the
Table 1:

Multiscale Hydro-Thermo-Chemo-Mechanical Coupling based on Staggered Method in the Concrete.

1. **Diffusion field.** Update the relative humidity $S^{n+1}$ through the instationary diffusion equation, once the diffusion part can numerically converge.

2. **Thermal field.** Update the temperature $\theta^{n+1}$ through the instationary thermal conduction, once the convergence criterion is satisfied.

3. **Chemical extent.** Adopt $S^{n+1}$ and $\theta^{n+1}$ to calculate the new chemical extent $\xi^{n+1}$ based on the back-Euler approach in a analytical manner, see Eq.(2.1), Fig.1(a) and Fig.1(b).

4. **Upscale damage.** Since temperature and humidity mediated chemical extent as the scalar multiscale variable is projected as constants from upper to the lower scales, the correlation between effective damage and chemical extent can be applied to the mesoscale, see Fig.3(b), therefore, the deterioration due to ASR at the microscale is determined through the value of chemical extent at the mesoscale.

5. **Mechanical field.** Apply the sum of the mechanical damage $D^u$ and chemical damage $D^c$ to the nonlinear constitutive law of HCP in order to obtain $u^{n+1}$, until the convergence is achieved.

6. **Increase time step.** Update all the field variables and set the time step forward to go back to step 1.

The correlation between effective damage and chemical extent through computation homogenization, thus presenting the multiscale investigation on ASR in the concrete. Staggered approach adopted in this contribution can overcome the trouble of too expensive cost in the computation while solving the problem of three-dimensional multiphysics, since it permits to solve the equations of all fields sequentially [9]. Supposing that the equilibrium states for diffusion, conduction and mechanical problems have been obtained at time $t_n$, the procedure for searching for the solutions at time $t_{n+1} = t_n + \delta t$ can be described, see Table 1.

5 CONCLUSIONS

The objective of the present work is to establish a multiscale model to quantitatively predict the deterioration due to ASR in the concrete structure. In this contribution, 3D micro-CT scan of hardened cement paste (HCP) and aggregates with a random distribution embedded in a homogenized cement paste matrix are used to represent the microscale and mesoscale of concrete respectively. One significant assumption is adopted that gels are evenly produced in micropores of HCP and exert uniform pressure on the surrounding
cement paste. The correlation between effective damage and chemical extent is sought through computational homogenization with statistical tests and the obtained expansion coefficient of the gel. Another assumption that temperature and humidity mediated chemical extent as the scalar multiscale variable is projected as constants from upper to the lower scales, ensures the reliability of the transition of the ASR induced damage between two scales.

A visco-plastic model of the classical PERZYNA-type combined with an isotropic damage is defined for HCP. At the mesoscale, hydro-chemo-thermal-mechanical coupling is implemented based on the staggered method, where the transient temperature and relative humidity are upscaled through instationary thermal conduction and diffusion, which contribute to the chemical extent $\xi$ through back-Euler method in an analytical manner. The deterioration due to ASR originating from microscale at the mesoscale is obtained the correlation between effective damage and chemical extent, qualifying the chemical extent. Therefore, a 3D multiscale hydro-chemo-thermo-mechanical is illustrated to predict the failure due to ASR in the concrete, yet it analyze the reaction at the microscale.

References


ANALYTICAL SOLUTIONS OF NONSTATIONARY ADIABATIC PROCESSES FOR COMPRESSION (EXTENSION) OF VISCOPLASTIC SPHERICAL AND CYLINDRICAL SHELLS, SPHERICAL AND CYLINDRICAL LAYERS FROM VISCOUS LIQUID

Alexey B. Kiselev

Mechanics and Mathematics Faculty of M.V. Lomonosov State University
Leninskie Gory, 1, MSU, Main Bldg, 119992 Moscow, Russia
e-mail: akis2006@yandex.ru

Key words: Viscoplastic Spherical and Cylindrical Shells, Compression, Extension.

Summary. In first part of this paper present original analytical solutions for one-dimensional problems of adiabatic compression and extension of viscoplastic (Sokolovsky-Perzyna type model) thick-walled spherical and cylindrical shells. Solutions obtained in Lagrange coordinates and at hypothesis of incompressibility of shells material. In second part of paper present original analytical solutions for one-dimensional problems of adiabatic compression and extension of viscous (Navier-Stokes model) incompressible spherical and cylindrical layers. These solutions may be, in particular, used for testing computational programs and estimate of effectiveness of new numerical methods.

Russian Foundation for Basic Research (grant 12-01-00425a) is acknowledged for financial support.
1 INTRODUCTION

It is known a few analytical solutions for dynamical problems of elastoviscoplasticity in view of its particular complication (see [1-5] and references in their publications). In contrast to papers [2, 3], where were investigated problems of compression and extension of spherical pores from incompressible viscoplastic material under permanent loading, in first part of present paper solutions obtained under dynamical external loading. In addition analogous solutions were obtained for cylindrical thick-walled shells.

Classical Zababaxin problem about fill out of bubbles in viscous liquid under action of permanent pressure on infinity [6], in second part of present paper this problem generalized for case of dynamical both external and internal pressure. In addition solution obtained for spherical layer finite thickness. Analogous solutions were obtained for cylindrical layer of viscous liquid.

2 COMPRESSION OF SPHERICAL SHELL

In one-dimensional approximation (all parameters dependent on radial Lagrange coordinate $R$ and time $t$) consider process of adiabatic compression of spherical shell which internal and external radius change in time on low $r_0 = r(R_0, t)$ and $r_1 = r(R_1, t)$ respectively, where $R_0$ and $R_1$ - internal and external radius of shell in initial time $t = 0$ (figure 1).

![Figure 1](image)

Do next simplifying assumption:

1) Behavior of shell material described of equations of elastoviscoplastic model Sokolovsky-Perzyna type [7]:

$$\dot{\varepsilon}_y = \frac{S_y}{2\mu} + \frac{S_y(S_y-S_y)^2}{2\eta\sqrt{S_y}} H(S_y-S_y-\frac{2}{3}Y_0)$$

(1)
Here $\dot{e}_i$ and $S_j$ – deviators of strain rates and stresses; $Y_0$ – yield limit under simple tension; $H(x)$ – Heaviside function; $\mu$ and $\eta$ – shear module and dynamic viscosity.

2) Elastic deformations can be neglected: $\dot{\varepsilon}^e = 0, \dot{\varepsilon}_i^p = \dot{\varepsilon}_i^p (\dot{\varepsilon}_i^e, \dot{\varepsilon}_i^p, \dot{\varepsilon}_i = \dot{\varepsilon}_i^e + \dot{\varepsilon}_i^p$ – elastic, plastic and total strain rates respectively); plastic flow is incompressible: $\dot{\varepsilon}_{ik} = 0$.

Therefore equations (1) reduced to one equation

$$\sigma_R - \sigma_0 = Y_0 + 2\eta(\dot{\varepsilon}_R - \dot{\varepsilon}_0)$$

Here $v$ – radial velocity; $\dot{\varepsilon}_R = \partial v / \partial R$, $\dot{\varepsilon}_0 = v / R$ – radial and ring strain rates; $\sigma_R, \sigma_0$ – radial and ring stresses.

Hypothesis of material incompressibility $\dot{\varepsilon}_R + 2\dot{\varepsilon}_0 = 0$ give equation for determination of velocity distribution in shell:

$$\frac{\partial v}{\partial R} + 2v / R = 0$$

Solution of equation (3) has next form:

$$v = C(t) / R^2$$

where $C(t) \leq 0$ because take place compression of shell and velocity $v \leq 0$.

The equation of momentum for spherical shell is

$$\rho_0 \dot{v} = \frac{\partial \sigma_R}{\partial R} + \frac{2(\sigma_R - \sigma_0)}{R}$$

Here $\rho_0$ – density of shell material; the dot over symbols indicates the material derivative with respect to time.

Put next boundary conditions:

$$\sigma_R \big|_{R=R_0} = \Sigma(t) < 0, \quad \sigma_R \big|_{R=R_0} = 0$$

If magnitude of external loading in initial moment $\Sigma(0)$ more than some critical value, which bill be find below, than material completely passed in plastic state. Precisely this loading is considered.

Substitute (2), (4) to equation of momentum (5), and integrate over $R$ with taking account of boundary conditions (6), we get ordinary differential equation for function $C(t)$:

$$\dot{C}(t) + \alpha C(t) = \beta \Sigma(t) + \gamma$$

where

$$\alpha = \frac{4\eta(R^2_1 + R_1R_0 + R^2_0)}{\rho_0 R_1^2 R_0^2}, \quad \beta = \frac{R_1 R_0}{\rho_0 (R_1 - R_0)}, \quad \gamma = \frac{2Y_0 R_1 R_0}{\rho_0 (R_1 - R_0)} \ln \frac{R_1}{R_0}$$

Solution of equation (7) with initial condition $C(0) = 0$ (velocity of shell under $t = 0$ equal zero) is:
Taking into account that acceleration in initial moment under compression is negative, i.e. \( \dot{C}(0) < 0 \), we take condition for initial external pressure to shell:

\[
\Sigma(0) = -\gamma \beta = -2Y_0 \ln \frac{R}{R_0}
\]

Value \( \Sigma_{\text{min}} = 2Y_0 \ln \frac{R}{R_0} \) is effective yield limit under multifold compression. This result was obtained earlier and in paper [2].

When external pressure to shell in constant, i.e. \( \Sigma(t) = \Sigma_0 = \text{const} < -2Y_0 \ln \frac{R}{R_0} \), we find that

\[
C(t) = \frac{1}{\alpha} (\beta \Sigma_0 + \gamma)(1 - e^{-\alpha t})
\]

Current value of Euler coordinate for material particle with Lagrange coordinate \( R \) is:

\[
r(R,t) = R \left[ 1 + \frac{\beta \Sigma_0 + \gamma}{\alpha R^3} \left( t + \frac{e^{-\alpha t} - 1}{\alpha} \right) \right]
\]

Obtain equation for moment of collapse of incompressible spherical shell \( t = t^*_s \) (i.e. then \( r(R_0,t^*_s) = 0 \)), for which obtained solution have physical meaning:

\[
e^{-\alpha t^*_s} + \alpha t^*_s = 1 - \frac{R_0^3 \alpha^2}{\beta \Sigma_0 + \gamma}
\]

Under condition (9) equation (12) always have uniqueness solution which is not has a form of elemental functions; however numerical solution of equation (12) is not a problem.

3 COMPRESSION OF CYLINDRICAL SHELL

In case of cylindrical shell under made above assumptions and since condition that displacements along axis \( z \) are lacking, i.e. strain rates \( \dot{\varepsilon}_z = 0 \), equations (1) result to one equationility

\[
\sigma_R - \sigma_\theta = \frac{2}{\sqrt{3}} Y_0 + 2\eta(\dot{\varepsilon}_R - \dot{\varepsilon}_\theta)
\]

Taking into account hypothesis of incompressibility of shell material \( \dot{\varepsilon}_R + \dot{\varepsilon}_\theta = 0 \), then similarly as in case of spherical shell instead of formulas (3)-(5), (7)-(12) we received

\[
\frac{\partial v}{\partial R} + \frac{v}{R} = 0
\]
\[ v = \frac{B(t)}{R}, \quad B(t) \leq 0 \] (15)

\[ \rho_0 \dot{\nu} = \frac{\partial \sigma_R}{\partial R} + \frac{\sigma_R - \sigma_0}{R} \] (16)

\[ \dot{B}(t) + \dot{\alpha} B(t) = \frac{\beta \Sigma(t)}{\alpha} + \frac{Y_0}{\sqrt{3\rho_0}} \] (17)

\[ \dot{\alpha} = \frac{2\eta(R_i^2 - R_o^2)}{\rho_0 R_i R_o^3 \ln(R_o / R_i)}, \quad \dot{\beta} = \frac{1}{\rho_0 \ln(R_i / R_o)}, \quad \bar{\gamma} = \frac{Y_0}{\sqrt{3\rho_0}} \]

\[ B(t) = \dot{\beta} e^{-\dot{\alpha}} \int_0^t e^{\dot{\alpha} \Sigma(t)} dt + \frac{\bar{\gamma}}{\alpha} (1 - e^{-\dot{\alpha} t}) \] (18)

\[ \Sigma(0) < -\frac{\bar{\gamma}}{\beta} = -\frac{Y_0}{\sqrt{3\rho_0}} \ln \frac{R_i}{R_o} \] (19)

\[ \Sigma(t) = \Sigma_0 = \text{const} < -\frac{Y_0}{\sqrt{3\rho_0}} \ln \frac{R_i}{R_o}: \quad B(t) = \frac{1}{\alpha} (\dot{\beta} \Sigma_0 + \bar{\gamma})(1 - e^{-\dot{\alpha} t}) \] (20)

\[ r(R,t) = R \left[ 1 + \frac{\dot{\beta} \Sigma_0 + \bar{\gamma}}{\alpha R_i^2} \left(t + \frac{e^{-\dot{\alpha} t} - 1}{\dot{\alpha}}\right) \right] \] (21)

\[ e^{-\dot{\alpha} t_i^*} + \dot{\alpha} t_i^* = 1 - \frac{R_i^2 \dot{\alpha}^2}{\beta \Sigma_0 + \bar{\gamma}} \] (22)

4 ABOUT EXTENSION OF SPHERICAL AND CYLINDRICAL SHELLS

In case of extension shells instead of boundary conditions (6) we state next boundary conditions:

\[ \sigma_R \mid_{R=R_0} = 0, \quad \sigma_R \mid_{R=R_1} = \Sigma(t) > 0 \] (23)

Easily note that obtained solutions of one-dimensional problems about compression of thick-walled spherical and cylindrical shells from incompressible viscoplastic material may be use for extension of shells. For that sufficiently in all formulas to substitute \( Y_0 \) on \((-Y_0)\).

In case of extension of shells \( C(t) \geq 0, \quad B(t) \geq 0 \). Give off meaning concept of moment of collapse of shells \( t_i^*, \quad t_i^{*} \). And conditions on value of external loading in initial moment for spherical and cylindrical shells (9), (19) take respectively next form:

\[ \Sigma(0) > -\frac{\bar{\gamma}}{\beta} = \frac{2Y_0 \ln \frac{R_i}{R_0}}{R_0}, \quad \Sigma(0) > -\frac{\bar{\gamma}}{\beta} = \frac{Y_0}{\sqrt{3\rho_0}} \ln \frac{R_i}{R_0} \] (24)

Consider example of extension spherical shell then loading have graduated form:

\[ \Sigma(t) = \Sigma_0 H(T - t), \quad \Sigma_0 = \text{const} > 2Y_0 \ln \frac{R_i}{R_0} \] (25)

where \( T \) – time of loading.
Permutate (25) in (8) and substitute \( Y_0 \) on \((-Y_0)\) we obtain

\[
C(t) = \frac{(\Sigma_0 - 2Y_0 \ln(R_l / R_0))R_0^3}{4\eta(1 - (R_0 / R_l)^3)}(1 - e^{-\alpha t}), \quad 0 \leq t \leq T
\]  

(26)

\[
C(t) = C(T) + \frac{Y}{\alpha}(e^{-\alpha T} - e^{-\alpha t}), \quad T < t \leq t_t^s
\]

(27)

In formula (27) \( C(T) \) determine from (26) and \( t_t^s \) (moment of stopping of extension for spherical shell) determine from (27) provided that \( C(t_t^s) = 0 \):

\[
t_t^s = -\frac{1}{\alpha} \ln\left(\frac{Y}{\gamma}C(T) + e^{-\alpha t}\right)
\]

(28)

Function which determine formulas (26) and (27) schematically show on figure 2.

![Figure 2](image)

Distribution of ring deformation in moment of stopping of extension for spherical shell \( t = t_t^s \) describe by next formule:

\[
\varepsilon_{\theta} = \frac{1}{R^3} \left[ (t_t^s(1 - e^{-\alpha t}) + T e^{-\alpha t})\frac{\Sigma_0 - 2Y_0 \ln(R_l / R_0)R_0^3}{4\eta(1 - (R_0 / R_l)^3)} + \frac{Y}{\alpha}(t_t^s - T)e^{-\alpha t} \right]
\]

(29)

In case of extension of cylindrical shell under action of loading graduated form

\[
\Sigma(t) = \Sigma_0 H(T - t), \quad \Sigma_0 = \text{const} > \frac{Y_0}{\sqrt{3}} \ln \frac{R_l}{R_0}
\]

(29)

we find, substitute \( Y_0 \) on \((-Y_0)\), that

\[
B(t) = \frac{(\Sigma_0 - Y_0 \ln(R_l / R_0) / \sqrt{3})R_0^2}{2\eta(1 - (R_0 / R_l)^3)}(1 - e^{-\alpha t}), \quad 0 \leq t \leq T
\]

(30)
In formula (31) \( B(T) \) determine from (30) and \( t_c^* \) (moment of stopping of extension for spherical shell) determine from (31) provided that \( B(t_c^*) = 0 \):

\[
t_c^* = -\frac{1}{\alpha} \ln\left(\frac{\alpha}{\gamma} B(T) + e^{-\beta t} \right)
\]

Distribution of ring deformation in moment of stopping of extension for cylindrical shell \( t = t_c^* \) describe by next formula:

\[
\varepsilon_\theta \bigg|_{t=t_c^*} = \frac{1}{R^2} \left( t_c^* (1 - e^{-\beta t}) + T e^{-\beta t} \right) \sum_{n=0}^{\infty} Y_0 \ln\left( \frac{R_1}{R_0} \right) / \sqrt{3} R_0^2 + \frac{\gamma}{\alpha} (t_c^* - T) e^{-\beta t} \right]
\]

**5 GENERALIZATION OF ZABABAXIN PROBLEM**

Consider spherical layer of viscous incompressible liquid. Internal and external radius change in time on low \( r_0 = r(R_0, t) \) and \( r_1 = r(R_1, t) \) respectively, where \( R_0 \) and \( R_1 \) - internal and external radius of layer in initial time \( t = 0 \) (figure 1).

Constitutive equations of model Navier-Stokes in one-dimensional spherical case are:

\[
\sigma_R = -p + 2\eta \dot{\varepsilon}_R, \quad \sigma_\theta = -p + 2\eta \dot{\varepsilon}_\theta
\]

Here \( p = -(\sigma_R + 2\sigma_\theta) / 3 \) - pressure in liquid; the other notations coincide with introduced above for viscoplastic media.

Put next boundary conditions:

\[
\sigma_R \big|_{r=R_i} = -P_i(t) < 0, \quad \sigma_R \big|_{r=R_0} = -P_0(t) < 0
\]

Here \( P_i(t), P_0(t) \) - pressure on external and internal surfaces of spherical layer.

Substitute (33), (4) to equation of momentum (5), and integrate over \( R \) with taking account of boundary conditions (34), we get ordinary differential equation for function \( C(t) \):

\[
\dot{C}(t) + \alpha C(t) = \beta (P_0(t) - P_i(t))
\]

where

\[
\alpha = \frac{4\eta (1 / R_0^3 - 1 / R_1^3)}{\rho_0 (1 / R_0 - 1 / R_1)}, \quad \beta = \frac{1}{\rho_0 (1 / R_0 - 1 / R_1)}
\]

Solution of equation (35) with initial condition \( v \big|_{t=0} = 0 \) (i.e. \( C(0) = 0 \)) is

\[
C(t) = \beta e^{-\alpha t} \int_0^t e^{\alpha t'} (P_0(t') - P_i(t')) dt', \quad v(R, t) = \frac{C(t)}{R^2}
\]
Under conditions $R_i \to +\infty$, $P_0(t) = 0$, $P(t) = P_\infty = \text{const}$ (Zababaxin problem [6]) we have:

$$\nu(R,t) = -P_\infty \frac{R_0^3}{4\eta R^2} \left(1-\exp\left(-\frac{4\eta}{\rho_0 R_0^2} t\right)\right)$$

Solution for cylindrical layer is

$$B(t) = \bar{\beta} e^{-\bar{\alpha} t} \int_0^t e^{\bar{\alpha} \xi} \left(P_0(t) - P_i(t)\right) d\xi, \quad \nu(R,t) = \frac{B(t)}{R}$$

where

$$\bar{\alpha} = \frac{2\eta \left(1/R_0^2 - 1/R_i^2\right)}{\rho_0 \ln(R_i/R_0)} \quad \bar{\beta} = \frac{1}{\rho_0 \ln(R_i/R_0)}$$

6 CONCLUSIONS

Obtained analytical solutions may be used, in particular, for testing computational programs and estimate of effectiveness of new numerical methods, as it is ware made in paper [8] used results of paper [1].

REFERENCES

COUPLED QUANTUM-CLASSICAL TRANSPORT IN SILICON NANOWIRES

O. MUSCATO AND V. DI STEFANO

Dipartimento di Matematica e Informatica
Università degli Studi di Catania
Viale A. Doria 6, 95125 Catania, Italy
e-mail: muscato@dm.unict.it, vdistefano@dm.unict.it

Key words: Silicon nanowires, Schrödinger-Poisson-Boltzmann system, hydrodynamic model, maximum entropy principle

Abstract. We present an extended hydrodynamic model describing the transport of electrons in the axial direction of a silicon nanowire. This model has been formulated by closing the moment system derived from the Boltzmann equation on the basis of the maximum entropy principle of Extended Thermodynamics, coupled to the Schrödinger-Poisson system. Explicit closure relations for the high-order fluxes and the production terms are obtained without any fitting procedure, including scattering of electrons with acoustic and non polar optical phonons. We derive, using this model, the electron mobility.

1 INTRODUCTION

Silicon nanowires (SiNWs) are quasi one-dimensional structures in which the electrons are spatially confined in the two transversal directions and free to move in the longitudinal one. SiNW devices have been fabricated recently using lithographic techniques [1]. They have attracted significant interest due to their potential to function as logic devices, thermoelectric devices, and sensors. Therefore, it is crucial to accurately model and estimate the performance of these devices.

By shrinking the dimension of electronic devices, effects of quantum confinement are observed and the wave nature of the electrons must be taken into account. The Non-Equilibrium Green Function formalism is the most advanced transport model for the simulation of SiNW devices, but it necessitates rather intensive computational efforts since it requires detailed information on the propagation of the electron wave packet injected in the device. Under reasonable hypothesis, transport in low-dimension semiconductors can be tackled coupling quantum and semiclassical tools. In fact, the main quantum transport phenomena in SiNW transistors at room temperature, such as the source-to-drain tunneling, and the conductance fluctuation induced by the quantum interference,
become significant only when the longitudinal length (called channel) is smaller than 10nm [2]. Therefore, for longer channels, semiclassical formulations based on the 1-D Multiband Boltzmann Transport Equation (MBTE) can give reliable simulation results when it is solved self-consistently with the 3-D Poisson and 2-D Schrödinger equations in order to obtain the self-consistent potential and subband energies and wavefunctions [1]. Another simplification comes from the use of the Effective Mass Approximation (EMA), which is supposed to be still a good solution in the confining direction in the presence of disorder, which is probably valid for semiconductor nanowires down to 5 nm in diameter, below which atomistic electronic structure models need to be employed. Solving the MBTE numerically is not an easy task, because it forms an integro-differential system in two dimensions in the phase-space and one in time, with a complicate collisional operator. The full solution of the MBTE can be obtained or by using the Monte Carlo (MC) method [3]-[10] or by using deterministic numerical solvers [11],[12],[13] at expense of huge computational times. Another alternative is to obtain from the MBTE hydrodynamic models that are a good engineering-oriented approach. This can be achieved by taking moments of the MBTE, and by closing the obtained hierarchy of balance equations as well as modeling the production terms (i.e. the moments on the collisional operator).

2 Transport equations

In the following we shall consider a SiNW with rectangular cross section. For a quantum wire with linear expansion in z-direction, and confined in the plane x-y, the normed electron wave function $\psi(x, y, z)$ can be written in the form

$$\psi(x, y, z) = \chi_\alpha(x, y)e^{ik_z z}/\sqrt{L_z}$$  \hspace{1cm} (1)

where $\chi_\alpha(x, y)$ is the wave function of the $\alpha$-th subband and the term $e^{ik_z z}/\sqrt{L_z}$ describes an independent plane wave in z-direction confined to the normalization length, where $z \in [0, L_z]$ and $k_z$ is the wave vector number. In general the electron is subject to external confining potential $U$, such as by a discontinuity in the band gap at an interface between two materials, and also to the effect of the other electrons in the system. The simplest approximation, called Hartree approximation, is to assume that the electrons as whole produce an average electrostatic energy potential $V_{tot}$, and that a given electron feels the resulting total potential

$$V_{tot} = U(x, y) - e\Phi(x, y, z)$$  \hspace{1cm} (2)

The normed wave function satisfies the Schrödinger equation in the Effective Mass Approximation, i.e.

$$\left[E_c - \frac{\hbar^2}{2m^*} \Delta + V_{tot}(x, y, z)\right] \psi = E \psi$$  \hspace{1cm} (3)

where $E$ is the total energy, $E_c$ the conduction band edge energy, and $m^*$ denotes the effective mass of the electron in the conduction band. By inserting eq.(1) into eq.(3), in
each z-th cross section of the device, one obtains the following equation for the envelope function $\chi_{az}(x, y)$

$$\left[-\frac{\hbar^2}{2m^*} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + U - e\Phi\right] \chi_{az} = \varepsilon_{az} \chi_{az} \quad , \quad E_{az} = \varepsilon_{az} + \frac{\hbar^2 k_z^2}{2m^*} + E_c$$

(4)

where $\varepsilon_{az}$ is the kinetic energy associated with the confinement in the x-y plane, and we have assumed parabolic band approximation. The term $\Phi$ satisfies the Poisson equation

$$\nabla \cdot [\varepsilon \nabla \Phi(x, y, z)] = e(n - N_D + N_A)$$

(5)

where $N_D, N_A$ are the doping profile (due to donors and acceptors) and $n(x, y, z, t)$ is the electron density, which depends on $\chi_{az}$

$$n(x, y, z, t) = \sum \rho^\alpha(z, t) |\chi_{az}(x, y, t)|^2$$

(6)

where $\rho^\alpha$ is the subband linear density in the z-direction

$$\rho^\alpha(z, t) = \frac{2}{2\pi} \int f^\alpha(z, k_z, t) dk_z$$

(7)

$f^\alpha$ being the electron distribution function in the $\alpha$-subband. For an assigned confining potential, one has to solve a coupled problem formed by eqs.(4), (5) and (6) to find $\varepsilon_{az}, \chi_{az}$ in each cross-section.

The transport in the z-th direction is described using the MBTE [1]

$$\frac{\partial f^\alpha}{\partial t} + v_z(k_z) \frac{\partial f^\alpha}{\partial z} - \frac{e}{\hbar} \varepsilon_z \frac{\partial f^\alpha}{\partial k_z} = \sum_{\alpha'} \sum_\eta C_\eta[f^\alpha, f^\alpha']$$

(8)

where $e$ is the absolute value of the electron charge, $\hbar$ the Planck constant divided by $2\pi$, and

$$v_z = \frac{1}{\hbar} \frac{\partial E_{az}}{\partial k_z} = \frac{\hbar k_z}{m^*} \quad , \quad \varepsilon_z = -\frac{1}{e} \frac{\partial E_{az}}{\partial z}$$

(9)

are respectively the electron group velocity and the electric field. In the low density approximation (not-degenerate case), the collisional operator writes

$$C_\eta[f^\alpha, f^\alpha'] = \frac{L_z}{2\pi} \int dk_z' \{ w_\eta(k', k)f^\alpha'(k_z') - w_\eta(k, k')f^\alpha(k_z) \}$$

(10)

where $w_\eta(k, k') = w_\eta(\alpha, k_z, \alpha', k'_z)$ is the $\eta$-th scattering rate. When $\alpha = \alpha'$ we have an intra-subband scattering, otherwise we have an inter-subband scattering.

Scattering mechanisms in SiNWs must comprise acoustic phonon scattering (bulk and confined), non-polar optical phonon scattering, surface scattering, scattering with ionized
impurities, as well as dielectric screening [4], [5]. However in this preliminary study, for
the sake of simplicity, we shall limit ourselves to consider just scattering with optical and
acoustic phonons. For the bulk acoustic phonon scattering, in the elastic equipartition
approximation, the transition rate is given by [1]

$$w_{ac}(k, k') = s_{ac} G^{\alpha\alpha'} \delta(E_{\alpha'} - E_{\alpha})$$

where $D_A$ is the acoustic deformation potential (9 eV), $T_L$ the lattice temperature, $\rho$ the
mass density ( 2.33 gr/cm$^3$), $v_s$ the sound speed (6960 m/sec), and $G^{\alpha\alpha'}$ the confinement
factor

$$G^{\alpha\alpha'} = \int |\chi_{\alpha'}(x, y)|^2 |\chi_{\alpha}(x, y)|^2 dx dy.$$  \(11\)

For the optical phonons we have

$$w_{op}(k, k') = s_{op} \left[ g_0 + \frac{1}{2} \pm \frac{1}{2} \right] G^{\alpha\alpha'} \delta(E_{\alpha'} - E_{\alpha} \mp \hbar \omega_0)$$

where $D_0$ is the optical deformation potential (11.4 $10^8$ eV/cm), $\hbar \omega_0$ the effective optical
phonon energy (63 meV), and $g_0$ the Bose-Einstein phonon occupation number.

3 Extended Hydrodynamic model

By multiplying the MBTE (8) by the weight functions $\psi_A = \{1, v_z, \varepsilon_z, v_z \varepsilon_z\}$, and
integrating in the $k_z$ space, one obtains the following hydrodynamic-like equations

$$\frac{\partial \rho^\alpha}{\partial t} + \frac{\partial (\rho^\alpha V^\alpha)}{\partial z} = \rho^\alpha \sum_{\alpha'} C_{\rho}^{\alpha\alpha'}$$  \(14\)

$$\frac{\partial (\rho^\alpha V^\alpha)}{\partial t} + \frac{2}{m^*} \frac{\partial (\rho^\alpha W^\alpha)}{\partial z} + \frac{e}{m^*} \rho^\alpha \varepsilon_z = \rho^\alpha \sum_{\alpha'} C_{V}^{\alpha\alpha'}$$  \(15\)

$$\frac{\partial (\rho^\alpha W^\alpha)}{\partial t} + \frac{\partial (\rho^\alpha S^\alpha)}{\partial z} + \rho^\alpha \varepsilon_z V^\alpha = \rho^\alpha \sum_{\alpha'} C_{W}^{\alpha\alpha'}$$  \(16\)

$$\frac{\partial (\rho^\alpha S^\alpha)}{\partial t} + \frac{\partial (\rho^\alpha F^\alpha)}{\partial z} + \frac{3 e}{m^* \rho^\alpha} \varepsilon_z W^\alpha = \rho^\alpha \sum_{\alpha'} C_{S}^{\alpha\alpha'}$$  \(17\)

in the unknowns (called moments)

$$V^\alpha = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \int f_{\alpha}(z, k_z, t) v_z dk_z \quad \text{(subband velocity)},$$  \(18\)

$$W^\alpha = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \int f_{\alpha}(z, k_z, t) \varepsilon_z dk_z \quad \text{(subband energy)},$$  \(19\)

$$S^\alpha = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \int f_{\alpha}(z, k_z, t) \varepsilon_z v_z dk_z \quad \text{(subband energy-flux)}$$  \(20\)
and the higher-order flux $F^\alpha$, and the production terms

$$F^\alpha = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \int f_{\alpha} v_z^2 \varepsilon_z dk_z \quad (21)$$

$$C'^\rho_{\alpha\alpha'} = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \sum_\eta \int C_\eta[f_{\alpha}, f_{\alpha'}] dk_z \quad (22)$$

$$C'^V_{\alpha\alpha'} = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \sum_\eta \int C_\eta[f_{\alpha}, f_{\alpha'}] v_z dk_z \quad (23)$$

$$C'^W_{\alpha\alpha'} = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \sum_\eta \int C_\eta[f_{\alpha}, f_{\alpha'}] \varepsilon_z dk_z \quad (24)$$

$$C'^S_{\alpha\alpha'} = \frac{2}{(2\pi)} \frac{1}{\rho^\alpha} \sum_\eta \int C_\eta[f_{\alpha}, f_{\alpha'}] \varepsilon_z v_z dk_z \quad . (25)$$

This system of PDEs is of hyperbolic type and it is not closed, i.e. there are more unknowns than equations. The Maximum Entropy Principle leads to a systematic way for obtaining constitutive relations on the basis of the information theory [14], as already proved successfully in the bulk case [15]-[19], and for quantum well structures [20], [21]. Actually, in a semiconductor electrons interact with phonons describing the thermal vibrations of the ions placed at the points of the crystal lattice. However, since we are considering the phonon gas as a thermal bath, one has to extremize only the electron component of the entropy. We define the entropy of the electronic system as

$$S_e = \sum_\alpha |\chi_\alpha(x, y, t)|^2 S_\alpha^e \quad (26)$$

$$S_\alpha^e = -\frac{2}{(2\pi)} k_B \int_R (f_\alpha \log f_\alpha - f_\alpha) dk_z \quad , (27)$$

and, according to MEP, we estimate the $f_\alpha$’s as the distributions that maximize $S_e$ under the constraints that the basic moments, which we have previously considered, are assigned. In a neighborhood of local thermal equilibrium, this distribution function writes [22]

$$\hat{f}_\alpha = \exp \left( -\frac{\lambda_\alpha}{k_B} - \lambda_\alpha^V \varepsilon_z \right) \left\{ 1 - \tau \left( \hat{\lambda}_\alpha^W v_z + \hat{\lambda}_\alpha^S v_z \varepsilon_z \right) \right\} \quad (28)$$

where the quantities ($\lambda_\alpha, \lambda_\alpha^W, \lambda_\alpha^V, \lambda_\alpha^S$) are known functions of the moments $\{\rho^\alpha, V^\alpha, W^\alpha, S^\alpha\}$. By using the distribution function (28) it is possible to evaluate the unknown functions appearing in the balance equations by integration. In this way the higher-order flux term writes

$$F^\alpha = \frac{6(W^\alpha)^2}{m^*} \quad (29)$$
as well as the production terms $C_{\alpha}^{\alpha'}, C_{V}^{\alpha'}, C_{W}^{\alpha'}, C_{S}^{\alpha'}$ have been determined. We want underline that this Extended Hydrodynamic model has been closed by using first principles, and it is free of any fitting parameters.

4 Electron Mobility

The mobility is one of the most important parameters that determine the performance of a field-effect transistor. At low electric field, the carrier drift velocity is proportional to the electric field strength, and the proportionality constant is defined as the mobility. Hence a higher mobility material is likely to have higher frequency response, because carriers take less time to travel through the device. When the fields are sufficiently large, nonlinearities in the mobility and saturation in the drift velocity are observed. In fact, the scattering of the carriers with the lattice, the impurities, and the surface is more active for higher fields, and the charges lose the energy gained by the electric field.

Now we want to prove that our Extended Hydrodynamic model is able to predict such behaviours. We shall assume that the wire is surrounded by an oxide which gives rise to an infinitely deep potential barrier. In such a case, the following analytical relations for the bottom energies and envelope functions can be used [1]

$$\varepsilon_{\alpha} = \frac{\hbar^2 \pi^2}{2m^*} \left( \frac{n^2}{L_x^2} + \frac{m^2}{L_y^2} \right), \quad \chi_{\alpha} = \sqrt{\frac{2}{L_x}} \sin \left( \frac{n \pi}{L_x} x \right) \sqrt{\frac{2}{L_y}} \sin \left( \frac{m \pi}{L_y} y \right), \quad n, m \in \mathbb{N}. \quad (30)$$

To obtain the drift velocity and the mobility, we have performed a numerical integration of our hydrodynamic model in the stationary homogeneous case with a constant electric field along the $z$ direction. In this case the unknowns $(\rho_{\alpha}, V_{\alpha}, W_{\alpha}, S_{\alpha})$ depend on the time only. The initial data are the equilibrium values, obtained with a global maxwellian with lattice temperature $T_0$ (300 K), i.e.

$$V_{\alpha}(0) = 0, \quad W_{\alpha}(0) = \frac{1}{2} k_B T_0, \quad S_{\alpha}(0) = 0 \quad (31)$$

$$\rho_{\alpha}(0) = L_x L_y N_D \frac{\exp \left( -\frac{\varepsilon_{\alpha}}{k_B T_0} \right)}{\sum_{\alpha} \exp \left( -\frac{\varepsilon_{\alpha}}{k_B T_0} \right)}, \quad (32)$$

where $N_D$ is the number of donor impurities. The average drift velocity is defined as

$$\langle V \rangle = \frac{\sum_{\alpha} \rho_{\alpha} V_{\alpha}}{\sum_{\alpha} \rho_{\alpha}}. \quad (33)$$

In the figure 1 we plot the subband velocities $V_{\alpha}$ ($\alpha = 1, \ldots, 4$) as well as the average drift velocity versus the simulation time, for an electric field of 1000 V/cm, with $L_x = L_y = 10$ nm. The stationary regime is reached in a few picoseconds, and the typical phenomenon of saturation is qualitatively and quantitatively well described.
In the figure 2 we plot the average drift velocity, obtained in the stationary regime, versus the electric field from which we can evaluate the saturation velocity $v_s = 6.7 \times 10^6$ cm/sec.
The average mobility is

$$\mu = \frac{\sum_\alpha \rho^\alpha \mu^\alpha}{\sum_\alpha \rho^\alpha} , \quad \mu^\alpha = \frac{V^\alpha}{E_z}$$  \hspace{1cm} (34)

where $\mu^\alpha$ is the subband mobility. In the figure 3 we plot the average mobility as function of the electric field. We notice for low fields ($\lesssim 1000$ V/cm) the mobility is constant (i.e. $\mu_0 = 406$ cm$^2$/V/sec) whereas, for high fields, the mobility decreases because the scattering processes become more active. In the latter figure for comparison we have also reported the mobility given by the Caughey-Thomas formula [23]

$$\mu_C = \mu_0 \left[ 1 + \left( \frac{\mu_0 E_z}{v_s} \right)^2 \right]^{-\frac{1}{2}}.$$ \hspace{1cm} (35)

Similar results have been obtained using MC simulations [4], but with more expensive computational times.
Figure 3: The average mobility (34)\textsubscript{1} (circles) versus the electric field with $L_x = L_y = 10$ nm, and the mobility evaluated using the Caughey-Thomas formula (diamonds).

5 Conclusions

An extended hydrodynamic model for SiNWs has been formulated with the use of the maximum entropy principle, where the transport coefficients are completely determined without any fitting procedure. Using this model we have evaluated the electron mobility (low and high-field), which is in agreement with MC simulation results. However, our model must be improved by including other relevant scattering mechanisms such as scattering with impurities, surface roughness, acoustic confined phonons. Simulation of real device as well as the study of thermoelectric effects according to the guideline in [24]-[28] are under investigation, and they will be published in the next future.

Acknowledgment

We acknowledge the support of the Università degli Studi di Catania, MIUR PRIN 2009 ”Problemi Matematici delle Teorie Cinetiche e Applicazioni”, and the CINECA Award N. HP10COORUX 2012.

REFERENCES


COUPLED PROBLEMS IN ANALYSIS OF QUANTUM DOTS WITH MULTIBAND MODELS

Sanjay Prabhakar* and Roderick Melnik*,†

*M²NeT Laboratory, Wilfrid Laurier University
Waterloo, ON Canada, N2L 3C5

†Gregorio Millan Institute, Universidad Carlos III de Madrid,
Leganes, Spain, 28911

Key words: Finite Element Method, Coupled multiphysics problems, 8-band $k \cdot p$ method, Piezoelectric fields and potentials

Abstract. We investigate the influence of piezoelectromechanical effects on the band structures of electron (hole) states in wurtzite quantum dots. We apply the 8-band $k \cdot p$ method and solve the corresponding eigenvalue (partial differential equations) problem for quantum dots with wetting layers based on the Finite Element Method. The coupled multiphysics model includes the piezoelectromechanical part and the band structure calculation part for electrons (holes) in quantum dots. We show that the piezoelectromechanical effects bring the localization of electron states at the top of the dots and hole states at the bottom of the dots.

1 Introduction

Studies on low dimensional systems have attracted considerable attention, spurred on by the development of smaller and faster electronic devices and by the exploitation of their extraordinary properties for improved performance in various areas of science and technology, including nano- and micro-electronics, thermoelectricity and magnetism [1, 2, 3, 5]. Today’s technology allows us to grow finite size semiconductor quantum dots with wetting layers. One can expect a straightforward analogy to the planar electronic/optoelectronic industry to extrapolate that complex compositionally modulated quantum dots structures could greatly increase the versatility and power of these building blocks in nanoscale applications [1, 2]. Low dimensional semiconductor nanostructures such as quantum dots, quantum wells and quantum wires can be grown in the laboratory by Stranski-Krastanov growth technique [1, 2, 3, 5, 6, 7, 8, 9, 10]. Devices based on these nanostructures are used in optoelectronics as light emitting diodes, lasers, etc [1, 2]. The band structures of nanostructures can be engineered and modified by several different schemes such as gate
controlled electric fields and n- or p-type doped semiconductors as well as other techniques. Piezoelectromechanical effects provide another efficient way to modify the band diagram of semiconductor materials.

In this paper, based on the fully coupled model and 8-band $k \cdot p$ method, we analyze in detail the influence of electromechanical effects on the band structures of wurtzite GaN quantum dots with wetting layers.

2 Computational Method

We consider GaN quantum dots embedded into the AlN matrix of hexagon shape (see Fig. 1). In all results reported here, each side of the hexagon for AlN matrix has its length of 17 nm and height of 20 nm. The height of the pyramidal shape GaN wurtzite quantum dots is 5 nm. Such quantum dots are grown over 1 nm thickness of the wetting layer. In the first part of the solution procedure, we have solved the system of partial differential equations (PDEs) describing piezoelectromechanical interactions in order to find the corresponding piezoelectric fields and potentials. In the second part, we diagonalize the strain dependent 8-band $k \cdot p$ Hamiltonian based on the Finite Element Method (FEM) [14]. In particular, in a typical simulation run, an AlN/GaN quantum dot with wetting layer contains the total number of 15945 elements, of which (from bottom to top) 4698 elements are in the lower barrier, 4255 elements are in the wetting layer, 3615 elements are in the quantum dots, and 3377 elements are in the upper barrier (see Fig. 1). We impose the Neumann boundary condition at the interface of AlN/GaN and the Dirichlet boundary conditions on the rest of the boundary. We utilize the UMFPACK solver in the
COMSOL multiphysics package [14] to find the piezoelectromechanical fields, eigenvalues and eigenfunctions of the corresponding eigenvalue PDE problem. Following [11, 12], in the next section we provide further details on the Hamiltonian and two main parts of the mathematical model.

3 Physics-Based Mathematical Model

3.1 Piezoelectromechanical effects in Cartesian coordinates

The starting point for the analysis of the influence of piezoelectromechanical effect on the band structure calculation of low dimensional semiconductor nanostructures is the coupled system of the Navier equations for stress and Maxwell’s equations for piezoelectric fields as [11]

\[
\begin{align*}
\partial_j \sigma_{ik} &= 0, \\
\partial_i D_i &= 0.
\end{align*}
\]

The stress tensor components \(\sigma_{ik}\) and the electric displacement vector components \(D_i\) are given by

\[
\begin{align*}
\sigma_{ik} &= C_{iklm} \varepsilon_{lm} + \epsilon_{niki} \partial_n V, \\
D_i &= \epsilon_{ilm} \varepsilon_{lm} - \hat{\epsilon}_{in} \partial_n V + P_{sp} \delta_{iz},
\end{align*}
\]

where \(C_{iklm}\) are the elastic moduli constants, \(\epsilon_{ik}\) is the piezoelectric constant, \(\epsilon_{in}\) is the permittivity, \(V\) is the piezoelectric potential, \(P_{sp}\) is the spontaneous polarization and
Figure 3: (Color online) Probability distributions of heavy holes and light holes in an AlN/GaN quantum dot without accounting for piezoelectromechanical effects.

\[ E = -\frac{\partial V}{\partial z} \]

is the built in piezoelectric field. Also, \( \varepsilon_{ij} \) are the components of strain tensors which are written as

\[ \varepsilon_{ij} = \varepsilon_{ij}^u + \varepsilon_{ij}^0, \quad (5) \]

where \( \varepsilon_{ij}^0 \) are the local intrinsic strain tensor components due to lattice mismatch and \( \varepsilon_{ij}^u \) is position dependent strain tensor components. These two can be written as

\[ \varepsilon_{ij}^0 = (\delta_{ij} - \delta_{iz}\delta_{jz}) a + \delta_{iz}\delta_{jz} c, \quad (6) \]

\[ \varepsilon_{ij}^u = \frac{1}{2} (\partial_i u_j + \partial_j u_i), \quad (7) \]

where \( a = (a_0 - a)/a_0 \) and \( c = (c_0 - c)/c_0 \) are the local intrinsic strains along \( a \)- and \( c \)-directions, respectively (which are nonzero in the quantum dots and zero otherwise). Here, \( a_0 \) and \( c_0 \) are the lattice constants of the quantum dots and the barrier material of the NWSLs.

### 3.2 8-band k · p model in Cartesian coordinates

The steady state Schrödinger equation of the Kane model for the electrons in the conduction band and holes in the valence band can be written as [11]

\[ \mathbf{H}\psi = E\psi, \quad (8) \]

where

\[ \mathbf{H} = \begin{pmatrix} H_c & H_{cv} \\ H_{cv}^* & H_v \end{pmatrix}, \psi = \begin{pmatrix} \psi_c \\ \psi_v \end{pmatrix}, \quad (9) \]
with $\psi_c = \psi_c(r)$ and $\psi_v = \psi_v(r)$ are the position dependent conduction and valence band envelope function.

The total wave function $\Psi$ is (see [11])

$$\Psi = \sum_{j=c,x,y,z} f_j \psi_j = f \psi,$$

where $f = (f_c f_x f_y f_z)$ and $\psi = (\psi_c \psi_x \psi_y \psi_z)^T$. The functions $f$ are spinless and $\psi$ is a spinor:

$$\psi_j = \left( \begin{array}{c} \psi^1_j \\ \psi^2_j \end{array} \right), \quad j = c, x, y, z.$$

Hence, the basis functions of the Hamiltonian (9) take the following form:

$$\left( f_c \psi^1_c, f_c \psi^2_c, f_x \psi^1_x, f_x \psi^2_x, f_y \psi^1_y, f_y \psi^2_y, f_z \psi^1_z, f_z \psi^2_z \right)^T$$

The next step is the description of the matrix Hamiltonian $H$ of (9). The diagonal element of the conduction band Hamiltonian $H_c$ can be written as

$$H_c = A'_1 k_x^2 + A'_2 (k_x^2 + k_y^2) + U_c + a_1 \varepsilon_{zz} + a_2 (\varepsilon_{xx} + \varepsilon_{yy}),$$

where $U_c = U_c(r)$ is the position dependent edge of the conduction band $\Gamma_1$, $a_1$ and $a_2$ are deformation potentials for the conduction band. The parameters $A'_1$ and $A'_2$ are expressed
Figure 5: (Color online) Probability distributions of ground and first excited states of electrons (holes) in an AlN/GaN quantum dot with piezoelectromechanical effects. Notice that the piezoelectromechanical effects bring the localization of electron wavefunctions to the top of the dot and hole wavefunctions to the bottom of the dot.

via the components $1/m_\parallel$ and $1/m_\perp$ of the tensor of the reciprocal effective masses for the conduction band in the single-band approximation and the Kane parameters $P_1 = -i\hbar\langle \phi_c|\hbar k_z|\phi_z \rangle/m_0$ and $P_2 = -i\hbar\langle \phi_c|\hbar k_x|\phi_x \rangle/m_0$. They are given by [11]

\[
A'_1 = \frac{\hbar^2}{2m_\parallel} - \frac{P_1^2}{E_g},
\]

\[
A'_2 = \frac{\hbar^2}{2m_\perp} - \frac{P_2^2}{E_g},
\]

where $E_g$ is the band gap of semiconductor materials.

The intra-valence-band Hamiltonian $H_v$ can be written as

\[
H_v = H^{(0)} + H^{(so)} + H^{(e)} + H^{(k)}.
\]

The Hamiltonian $H^{(0)}$ entering Eq. (15) represents the position-dependent potential energy of an electron:

\[
H^{(0)} = \begin{pmatrix}
U_{v6} & 0 & 0 \\
0 & U_{v6} & 0 \\
0 & 0 & U_{v1}
\end{pmatrix},
\]

where $U_{v6} = U_{v6}(r)$ and $U_{v1} = U_{v1}(r)$ are the position dependent edges of the valence bands $\Gamma_6$ and $\Gamma_1$, respectively.
The spin-orbit Hamiltonian $H^{(so)}$ in Eq. (15) can be treated as a perturbation term and can be written as \[7, 8\]

$$H^{(so)} = i \begin{pmatrix} 0 & -\Delta_2 \sigma_z & \Delta_3 \sigma_y \\ \Delta_2 \sigma_z & 0 & -\Delta_3 \sigma_x \\ -\Delta_3 \sigma_y & \Delta_3 \sigma_x & 0 \end{pmatrix},$$

(17)

where $\Delta_2 = \Delta_2(r)$ and $\Delta_3 = \Delta_3(r)$ are the parameters of the valence-band spin-orbit splitting and $\sigma_i (i = x, y, z)$ are the Pauli spin matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$  

(18)

The kinetic energy Hamiltonian $H^{(k)}$ in Eq. (15) can be written as

$$H^{(k)} = \begin{pmatrix} L_1' k_x^2 + M_1 k_y + M_2 k_z^2 & N_1' k_x k_y & N_2' k_x k_z \\ N_1' k_x k_y & M_1 k_x^2 + L_1' k_y^2 + M_2 k_z^2 & N_2' k_y k_z \\ N_2' k_x k_z & N_2' k_y k_z & M_3 (k_x^2 + k_y^2) + L_3 k_z^2 \end{pmatrix},$$

(19)

where

$$L_1' = L_1 + \frac{P_1^2}{E_g}, \quad L_2' = L_2 + \frac{P_2^2}{E_g},$$

$$N_1' = N_1 + \frac{P_1^2}{E_g}, \quad N_2' = N_2 + \frac{P_1 P_2}{E_g}.$$  

(20)

(21)

Also,

$$L_1 = \frac{\hbar^2}{2m_0} (A_2 + A_4 + A_5),$$

$$M_1 = \frac{\hbar^2}{2m_0} (A_2 + A_4 - A_5),$$

$$N_1 = \frac{\hbar^2}{2m_0} 2A_5, \quad L_2 = \frac{\hbar^2}{2m_0} A_1,$$

$$M_2 = \frac{\hbar^2}{2m_0} (A_1 + A_3), \quad N_2 = \frac{\hbar^2}{2m_0} \sqrt{2} A_6,$$

$$M_3 = \frac{\hbar^2}{2m_0} A_2, \quad N_3 = i \sqrt{2} A_7,$$

with $A_1, A_2, \ldots, A_7$ being real material parameters in conventional notations \[7, 8, 11\], $m_0$ is the free electron mass. Wurtzite structure has six fold rotational symmetry and thus we use the relation $L_1' - M_1 = N_1'$. 


Finally, the strain tensor components are written as [11]

\[
\mathbf{H}^{(e)} = \begin{pmatrix}
    l_1 \varepsilon_{xx} + m_1 \varepsilon_{yy} + m_2 \varepsilon_{zz} & n_1 \varepsilon_{xy} & n_2 \varepsilon_{xz} \\
    n_1 \varepsilon_{xy} & m_1 \varepsilon_{xx} + l_1 \varepsilon_{yy} + m_2 \varepsilon_{zz} & n_2 \varepsilon_{yz} \\
    n_2 \varepsilon_{xz} & n_2 \varepsilon_{yz} & m_3 (\varepsilon_{xx} + \varepsilon_{yy}) + l_2 \varepsilon_{zz}
\end{pmatrix},
\]

(23)

where material constants \( l_1, l_2, m_1, m_2, n_1, \) and \( n_2 \) are expressed via conventional deformation potential tensor components as follows: [8, 7, 11]

\[
l_1 = D_2 + D_4 + D_5, \quad m_1 = D_2 + D_4 - D_5, \\
n_1 = 2D_5, \quad l_2 = D_1, \quad m_2 = D_1 + D_3, \\
n_2 = \sqrt{2}D_6, \quad m_3 = D_2.
\]

(24)

We can also apply six fold rotational symmetry in the strain Hamiltonian of wurtzite structure which holds the relation \( l_1 - m_1 = n_1 \).

The Hamiltonian \( \mathbf{H}_{cv} \) of (9) (\( \mathbf{H}_{cv}^\dagger \) is its Hermitian conjugate) can be written as

\[
\mathbf{H}_{cv} = (H_{cx} \quad H_{cy} \quad H_{cz}),
\]

(25)

where

\[
H_{cx} = iP_2k_x, \quad H_{cy} = iP_2k_y, \quad H_{cz} = iP_1k_z.
\]

(26)

4 Results and Discussions

In Fig. 2, we have plotted the probability distributions of ground and first excited states of electrons in AlN/GaN quantum dots without accounting for piezoelectromechanical effects. In Fig. 3, we have plotted the probability distributions of heavy holes and light holes in AlN/GaN quantum dots without accounting for piezoelectromechanical effects. Since we consider the flat band of AlN/GaN quantum dots in our 8-band \( \mathbf{k} \cdot \mathbf{p} \) Hamiltonian, we see that the localization of the wave functions of the electrons and holes in quantum dots reside in the center of the dots.

Next, we have solved the piezoelectromechanical part and plotted the piezoelectric fields and piezoelectric potentials in AlN/GaN quantum dots in Fig. 4. We couple the piezoelectromechanical and band structure calculation parts via the 8-band \( \mathbf{k} \cdot \mathbf{p} \) Hamiltonian.

Then, we have diagonalized the strain dependent 8-band \( \mathbf{k} \cdot \mathbf{p} \) Hamiltonian based on Finite Element Method and investigate the influence of piezoelectromechanical effects in wurtzite AlN/GaN quantum dots. In Fig. 5, we have plotted the probability distributions of ground and first excited states of electrons and holes in AlN/GaN quantum dots. It can be seen that the influence of piezoelectromechanical effects bring the electron states to the top of the dots and hole states (both heavy and light holes) to the bottom of the dots near the wetting layers.
5 Conclusion

We have solved the three-dimensional coupled multiphysics model of strain dependent 8-band $k \cdot p$ Hamiltonian in Cartesian coordinates based on the Finite Element Method. We have shown that without piezoelectromechanical effects, electron and hole wavefunctions are located in the center of the quantum dots. By accounting for piezoelectromechanical effects, we have found that in the analyzed wurtzite AlN/GaN quantum dots these effects push the distribution of electron wavefunctions to the top of the quantum dots and hole wavefunctions to the bottom of the quantum dots, near the wetting layer.

6 Acknowledgement

This work was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada and Canada Research Chair (CRC) programs.

REFERENCES


COUPLED SWELLING AND LARGE STRAIN MODEL FOR HYDROGELS: APPLICATION TO THE NUCLEUS PULPOUS OF THE INTERVERTEBRAL DISC

V. BRULLIARD*, S. LE CORRE†† AND E. VERRON††

*††LUNAM Université, Ecole Centrale de Nantes
GeM-UMR CNRS 6183, 1 rue de la Ncé, BP 92101
44321 Nantes, France
e-mail: viclette.brulliard@ec-nantes.fr, web page: http://gem.ec-nantes.fr/

††LUNAM Université, Polytech’ Nantes
LTN-UMR CNRS 6607, rue Christian Pauc, BP 50906
44306 Nantes, France
e-mail: steven.le-corre@univ-nantes.fr, web page: http://www.polytech.univ-nantes.fr/ltn/fr/

Key words: Hydrogel, Coupled Swelling and Large Strain, Intervertebral Disc, Nucleus Pulposus

Abstract. The main constituents of the nucleus pulposus, the central gelatinous part of the intervertebral disc, are water and a solid extracellular matrix of macromolecules. Based on this observation, the nucleus pulposus can be seen as a gel-like material in which swelling - due to the diffusion of the fluid molecules into the macromolecular network - and large strain elasticity, induced by the macromolecular chains, occur. In recent literature, many authors have been interested in describing such a coupled deformation-diffusion problem for gels. Most of the time, these works were formulated with respect to the dry configuration of the material. However, for the nucleus pulposus, the dry configuration does not exist. Thus, the formulation of the deformation-diffusion problem should be modified in order to consider a reference configuration, which is undeformed and unconstrained (but already swollen).

The theoretical aspects of the derivation of this coupled deformation-diffusion model is proposed. Then, this model is numerically implemented in the finite element commercial software ABAQUS. The states of equilibrium are investigated on simple homogeneous examples. Finally, the human nucleus pulposus of the intervertebral disc is chosen as a representative complex example of application for this approach. The associated annulus fibrosus is modelled using an anisotropic hyperelastic material law.
1 INTRODUCTION

The role of the intervertebral disc (IVD) is to provide flexibility of the spine, and transmit and distribute large loads through the spine. To carry out these tasks the IVD consists in a particularly complex structure made of a central gelatinous nucleus pulposus (NP) surrounded circumferentially by the annulus fibrosus (AF) as shown in Figure 1. However, due to aging or pathological process, many people show degenerative changes in the IVD, and especially in the NP. These changes affect both its composition and structure, as well as its mechanical functions. Back pain is often a clinical consequence of disc degeneration.

\[ \text{Figure 1: Representation of the IVD structure (issued from [1]).} \]

The full understanding of the role of each one of the components of the IVD in the mechanical response of the IVD, as well as the mechanism of disc degeneration, is of major interest. Indeed, it would help proposing solutions to the problem of disc degeneration. In that purpose, finite element models are a very useful tool. Several finite element models have already been used to predict the behavior of the IVD, to investigate the process of disc degeneration, or to understand the complex exchange mechanisms that occur within the IVD. Anisotropic highly non-linear constitutive models reflecting the physical structure of the AF have been reported [2, 3]. However, relevant constitutive models for the NP are not straightforward. Indeed, for relatively simple purposes, the NP has often been modelled as a non-linear incompressible solid or a fluid [4]. More recently, the NP has been considered as a biphasic material, comprising an incompressible fluid phase that saturates and flows through an elastic isotropic solid phase, leading to complex equations [5]. Our goal is to propose a new approach for modelling the NP. Based on the observation of its constituents - mainly water (about 80\%) and a solid extracellular matrix of structural macromolecules (proteoglycans, collagens) - the NP can be seen as a gel-like system which couples large strain elasticity induced by the macromolecular chains, and swelling properties related to the migration of the solvent through the network.

In recent literature, some authors have been interested in describing such a coupled deformation-diffusion problem for gels [6, 7, 8]. In most of these works, the problem was formulated with respect to the dry configuration of the material. However, keeping in mind the application to the NP, the derivation of problem should be modified in order to consider a reference configuration which would not be dry, but already swollen (since the
NP initially contains a large quantity of water). To our knowledge, only Lucantonio and coworkers have introduced a reference state which is initially swollen, homogeneous and stress-free, but they do not completely overcome the initial dry configuration [9].

Here, the theoretical aspects of a new constitutive model for gel-like material such as the NP of the IVD, based on an already swollen reference, are developed. This model is then numerically implemented in the finite element commercial software ABAQUS for investigating the states of chemical long-term equilibrium. As benchmark examples, simple homogeneous cases, such as stress-free swelling and uniaxial tension of a cubic sample are studied. Finally, a finite element model of a simplified human intervertebral disc is proposed: the present approach is applied to the central NF, whereas the surrounding annulus fibrosus is classically modelled using an anisotropic hyperelastic material law.

2 A CONSTITUTIVE MODEL OF COUPLED SWELLING AND LARGE STRAIN FOR A GEL-LIKE MATERIAL

2.1 Kinematics

The gel-like material is considered as single continuum body, subjected to both large strain and solvent diffusion. As usual for the elasticity of soft materials, the general framework of large strain hyperelasticity is adopted, and the material is considered homogeneous and isotropic at the Continuum Mechanics scale.

The reference configuration ($\Omega_0$) of the material body is assumed to be already swollen, undeformed and unconstrained. At time $t > 0$, the body occupies a different region ($\Omega$). The motion of any material point $P$ of position $X$ in the reference configuration is described by the mapping $x = \varphi(X, t)$, where $x$ denotes the spatial position of $P$. The deformation gradient is given by $F = \text{Grad}\varphi$.

A multiplicative decomposition of the deformation gradient $F$ is adopted:

$$F = f \ F_s, \quad \tag{1}$$

where $F_s \lambda_s I$ represents the homogeneous, isotropic deformation due to swelling, from the reference state ($\Omega_0$) to the equilibrium (fully swollen) state ($\Omega_s$), $\lambda_s$ being the equilibrium swelling stretch, and $f$ stands for the elastic deformation of the gel-like material. Let denote $V_0$ the volume of the material in the reference state, $V_s$ the volume in the swollen state and $V$ the volume in the final state. Thus, we have the following relationship describing the volume changes between the different configurations:

$$J = \det F = \frac{V}{V_0}, \quad J_s = \det F_s = \frac{V_s}{V_0}, \quad j = \det f = \frac{V}{V_s}, \quad \text{and} \quad J = j J_s. \quad \tag{2}$$

Assuming that both the solid network and the water are incompressible, we have $j = 1$, and $J = J_s$ will describe the swelling degree from the reference configuration.

Assuming that there is no free volume, the volume in the reference configuration $V_0$ and the volume in the swollen configuration $V_s$ are given respectively by:
\[ V_0 = V_m + V_{f,0}, \quad \text{and} \quad V_s = V_m + V_f = V_m + V_{f,0} + V_{f,a}, \] (3)

where \( V_m \) is the volume of the solid matrix, \( V_{f,0} \) the initial volume of fluid, \( V_f \) the total volume of the fluid molecules in the swollen state, and \( V_{f,a} \) is the volume of fluid molecules added between the reference state and the swollen state. Thus, the swelling degree writes:

\[ J_s = \frac{V_m + V_f}{V_m + V_{f,0}} = 1 + \frac{V_{f,a}}{V_0}. \] (4)

Introducing \( \nu \) the volume of a fluid molecule, \( n_{f,a} \) the number of fluid molecules added between the reference state and the final one, we have

\[ J_s = 1 + \frac{\nu n_{f,a}}{V_0} = 1 + \nu c, \] (5)

where \( c = n_{f,a}/V_0 \) is the number of fluid molecules added per unit of reference volume. Equation (5) is often referred to as the molecular incompressibility constraint [6].

2.2 Free energy and constitutive theory

2.2.1 General form of the free energy density, thermodynamic restrictions

First, the deformation between \((\Omega_0)\) and \((\Omega_s)\) is due to the diffusion of the fluid molecules in the material. Thus, the Helmholtz free energy which includes the chemical potential of the pure solvent has to be considered, as well as the change in free energy due to mixing of the solvent with the solid network. Secondly, the strain energy between \((\Omega_s)\) and \((\Omega)\) should also be introduced. The total free energy density of the material written per unit reference volume is the sum of these contributions:

\[ \Psi(F) = \Psi(f, J_s) = \Psi_{\text{pure}}(J_s) + \Psi_{\text{mix}}(J_s) + \Psi_{\text{mech}}(f, J_s). \] (6)

Considering the general framework of Thermodynamics of Irreversible Processes (see for example [10]) for deformation and diffusion, the Clausius–Duhem inequality issued is:

\[ \mathcal{D}_{\text{int}} = P : \dot{\mathbf{F}} + \mu \dot{\mathbf{c}} - \dot{\Psi} - \mathbf{j} \cdot \text{Grad} \mu \geq 0, \] (7)

where \( \mathcal{D}_{\text{int}} \) is the internal dissipation, \( P \) is the first Piola-Kirchhoff (nominal) stress tensor, \( \mu \) the chemical potential and \( \mathbf{j} \) represents the fluid flux. Recalling that \( \Psi \) can be written as a function of both the deformation gradient \( \mathbf{f} \) and the variable \( J_s \), the internal dissipation can be expressed as:

\[ \mathcal{D}_{\text{int}} = P : \dot{\mathbf{F}} + \mu \dot{\mathbf{c}} - \frac{\partial \Psi}{\partial \mathbf{f}} \bigg|_{J_s} : \dot{\mathbf{f}} - \frac{\partial \Psi}{\partial J_s} \bigg|_{\mathbf{f}} J_s - \mathbf{j} \cdot \text{Grad} \mu \geq 0. \] (8)

Deriving the multiplicative decomposition Eq. (1) with respect to time, we have:
\[ \dot{\mathbf{F}} = \frac{1}{3} J_s^{-2/3} \dot{J}_s \mathbf{f} + J_s^{1/3} \mathbf{f}. \] 

Then, using Eq. (9) and the incompressibility constraint Eq. (5) in the second term, Eq. (8) becomes

\[ \left( J_s^{1/3} \mathbf{P} - \frac{\partial \Psi}{\partial \mathbf{f}} \right) : \dot{\mathbf{f}} + \left[ \frac{\mu}{\nu} + \frac{1}{3} J_s^{-2/3} \mathbf{f} : \mathbf{f} - \frac{\partial \Psi}{\partial J_s} \right] \dot{J}_s - \mathbf{j} \cdot \text{Grad}\mu \geq 0. \] 

The next step of the derivation consists in invoking the classical Coleman-Noll procedure.

### 2.2.2 Constitutive equations

The details of the derivation are not given here; only the final constitutive equations are given.

**Stress-strain relationship** Introducing the right Cauchy-Green strain tensor of the purely elastic deformation \( \mathbf{c} = \mathbf{f}^T \mathbf{f} \), the first Fiola-Kirchhoff stress tensor is

\[ \mathbf{P} = -q \mathbf{F}^{-T} + 2J_s^{-1/3} \mathbf{f} \frac{\partial \Psi_{\text{mech}}}{\partial \mathbf{c}} = -q \mathbf{F}^{-T} + 2J_s^{-2/3} \mathbf{F} \frac{\partial \Psi_{\text{mech}}}{\partial \mathbf{c}}, \] 

where \( q \) is an arbitrary scalar (classically referred to as hydrostatic pressure). Recalling that the Cauchy (true) stress tensor \( \sigma \) is related to the first Fiola-Kirchhoff stress tensor \( \mathbf{P} \) by \( \sigma = J_s^{-1} \mathbf{P} \mathbf{F}^T \), we have also

\[ \sigma = -q J_s^{-1} \mathbf{I} + J_s^{-1} \frac{\partial \Psi_{\text{mech}}}{\partial \mathbf{c}} \mathbf{f}^T. \] 

**Chemical potential-fluid content relationship** The chemical potential is given by

\[ \frac{\mu}{\nu} = \left( \frac{d\Psi_{\text{pure}}}{dJ_s} + \frac{d\Psi_{\text{mix}}}{dJ_s} + \frac{\partial \Psi_{\text{mech}}}{\partial J_s} \right) - \sigma^S, \] 

\( \sigma^S := \frac{1}{3} \mathbf{f}^T \mathbf{f} = \frac{1}{3} J_s^{-2/3} \mathbf{P} : \mathbf{f} \) being the isotropic hydrostatic stress.

### 2.3 Specialization of the theory

The very general form of the constitutive equations have been derived. Now, the theory has to be specialized by imposing additional constitutive assumptions on the free energy densities.

Denoting \( \mu_0 \) the chemical potential of the pure solvent, the free energy density of the pure solvent per unit reference volume is simply:
\[ \Psi_{\text{pure}} = \mu_0 c = \frac{\mu_0}{\nu} (J_s - 1). \] (14)

The standard Flory-Huggins free energy \[11, 12\] for mixing is originally defined per unit dry volume by
\[ \Psi_{\text{mix, dry}}(\phi_m) = \frac{kT}{\nu} \frac{1}{\phi_m} [(1 - \phi_m) \ln (1 - \phi_m) + \chi \phi_m (1 - \phi_m)], \] (15)
where \( k \) is the Boltzmann constant, \( T \) is the temperature, \( \chi \) is a dimensionless parameter denoting the interaction between the fluid and the solid matrix, and \( \phi_m = V_m/V \) is the volume fraction of the solid matrix. Equation (15) may be also written in terms of \( J_m = 1/\phi_m \):
\[ \Psi_{\text{mix, dry}}(J_m) = \frac{kT}{\nu} (J_m - 1) \left[ \ln \left( 1 - \frac{1}{J_m} \right) + \frac{\chi}{J_m} \right]. \] (16)

Introducing also the relationship \( J_m = J_{s,0} J_s \), the free energy density of mixing per unit reference volume becomes:
\[ \Psi_{\text{mix}}(J_s) = \frac{1}{J_{s,0}} \Psi_{\text{mix, dry}}(J_m) = \frac{kT}{\nu} \frac{1}{J_{s,0}} (J_{s,0} J_s - 1) \left[ \ln \left( 1 - \frac{1}{J_{s,0} J_s} \right) + \frac{\chi}{J_{s,0} J_s} \right]. \] (17)

For the deformation of the solid matrix, a simple neo-Hookean formulation is adopted:
\[ \Psi_{\text{mech}}(J_s) = \frac{C_0}{2} (I_1 - 3 - 2 \ln J_s), \] (18)
\( C_0 \) being the shear modulus of the material defined per unit reference volume, and \( I_1 \) being the first invariant of transformation \( \mathbf{F} \), related to the first invariant of transformation \( \mathbf{f} \) by the relationship \( I_1 = \text{tr} (\mathbf{F}^T \mathbf{F}) = J_s^{2/3} I_1 \).

Thus, adding the three contributions given by Eqs. (14), (17), (18), the total free energy density per unit reference volume is:
\[ \Psi(\mathbf{f}, J_s) = \frac{\mu_0}{\nu} (J_s - 1) + \frac{kT}{\nu} \frac{1}{J_{s,0}} (J_{s,0} J_s - 1) \left[ \ln \left( 1 - \frac{1}{J_{s,0} J_s} \right) + \frac{\chi}{J_{s,0} J_s} \right] \]
\[ + \frac{C_0}{2} (I_1 - 3 - 2 \ln J_s). \] (19)

Deriving Eqs. (14), (18) and (17) with respect to \( J_s \), and Eq. (18) with respect to \( c \), and substituting them into the general expression of the stresses Eqs. (11) and (12) and into the general expression of the chemical potential Eq. (13), the final set of constitutive equations for a gel-like material is:

- Stress
\( \mathbf{P} = -q \mathbf{F}^{-T} + C_0 \mathbf{F} \), \quad and \quad \sigma = -q J_s^{-1} \mathbf{I} + C_0 J_s^{-1} \mathbf{B}, \) (20)
where \( \mathbf{B} = \mathbf{F} \mathbf{F}^T \) is the right Cauchy-Green stress tensor of transformation \( \mathbf{F}. \)
- Chemical potential

\[ \frac{\mu - \mu_0}{kT} = \ln \left( 1 - \frac{1}{J_{s,0}J_s} \right) + \frac{1}{J_{s,0}J_s} + \frac{\chi}{(J_{s,0}J_s)^2} \]  
\[ + \frac{\nu}{kT} J_s^{-1} (q - C_0). \]  

(21)

3 FINITE ELEMENT IMPLEMENTATION

This constitutive theory for gel-like materials has been implemented into the finite element software ABAQUS, through the user-defined subroutine UHYPHER, as proposed in [13]. For that purpose, another form of the free energy density than Eq. (19) has been used, obtained by a Legendre transformation:

\[ \Psi(f, \mu) = \mu C - \Psi(f, J_s) = \frac{\mu}{\nu} (J_s - 1) - \Psi(f, J_s), \]  

(22)

keeping in mind the incompressibility constraint Eq. (5). Noting that ABAQUS uses \( \bar{I}_1 = J_s^{-2/3} I_1 \) rather than \( I_1 \), and normalising the expression by \( kT/\nu \), the free energy density expression that has been implemented is the following:

\[ \Psi_{\text{ncm}}(f, \mu) = \frac{\mu - \mu_0}{kT} (J_s - 1) - \frac{1}{J_{s,0}J_s} (J_{s,0}J_s - 1) \left[ \ln \left( 1 - \frac{1}{J_{s,0}J_s} \right) + \frac{\chi}{J_{s,0}J_s} \right] \]
\[ - \frac{1}{2} \frac{C_0 \nu}{kT} \left( J_s^{2/3} \bar{I}_1 - 3 - 2 \ln J_s \right). \]  

(23)

The material parameters \( C_0, \chi \) and \( J_{s,0} \) are passed by the user to the subroutine. Practically, rather than \( J_{s,0} \), the fluid volume fraction in the reference state defined as \( \phi_f = 1 - 1/J_{s,0} \) is used.

To study the equilibrium states, the "normalized" chemical potential variation \( (\mu - \mu_0)/kT \) is imposed as a uniform predefined field within the material at each increment; its values is incremented gradually as a loading parameter, from an initial value \( (\mu - \mu_0)/kT|_{\text{init}} \) to 0 (corresponding to the chemical equilibrium \( \mu = \mu_0 \)). The initial value should be specified in agreement with the fact that the material is isotropic, and supposed undeformed and unconstrained initially.

4 APPLICATIONS

First, two simple homogeneous cases (stress-free swelling and uniaxial tension of a cubic sample) are studied as benchmark examples, in order to emphasize the influence of the reference state. Then, the constitutive approach is applied to the central NP of the IVD. For the different cases, \( kT = 4 \cdot 10^{-21} \) J at room temperature, the molecular volume of water is \( \nu = 3.0 \cdot 10^{-29} \) m\(^3\)/molecule, and \( \chi = 0.1 \) is fixed.
4.1 Homogeneous examples

4.1.1 Free swelling

Cubic samples with different initial fluid content \( \phi_f \) are immersed in a fluid bath of chemical potential \( \mu_0 \) and allowed to swell freely, isotropically, and without constraint. The deformation gradient and the state of stress are simply

\[
F = \lambda_s I, \quad \text{and} \quad \sigma = 0.
\]  

(24)

Using Eq. (24) together with Eqs. (20) and (21) permits to write the equation satisfied by the free-swelling stretch \( \lambda_s \):

\[
\frac{\mu - \mu_0}{kT} = \ln \left( 1 - \frac{1 - \phi_f}{\lambda_s^3} \right) + \frac{1 - \phi_f}{\lambda_s^3} + \frac{\chi (1 - \phi_f)^2}{\lambda_s^6} \right) + \frac{C_{0\nu}}{kT} \left( \lambda_s^{-1} - \lambda_s^{-3} \right).
\]  

(25)

Figure 2 shows the free-swelling isotropic stretch \( \lambda_s \) as a function of the "normalized" chemical potential variation, for \( \phi_f = 30, 60, 80\% \) (the shear modulus in the reference state was taken as \( C_0 = 4 \cdot 10^4 \) Pa). As expected, the free-swelling stretch \( \lambda_s \) increases when one approaches the equilibrium \( \mu = \mu_0 \), whatever the value of the initial fluid content \( \phi_f \). Moreover, as the initial fluid content \( \phi_f \) increases, the swelling degree at equilibrium \( J_s^e = (\lambda_s^e)^3 \) decreases.

![Figure 2: Isotropic stretch as a function of the chemical potential for three reference states.](image)

4.1.2 Uniaxial tension

Bar sample with different initial fluid content \( \phi_f \) are immersed in a fluid bath of chemical potential \( \mu_0 \) and subjected to uniaxial tension \( P_{11} \) in the longitudinal direction \( e_1 \). The deformation gradient and the first Piola-Kirchhoff stress tensor are given by:
\[
F = \text{diag}(\lambda_1, \lambda_2, \lambda_2), \quad \text{and} \quad P = P_{11} e_1. \tag{26}
\]

Substituting Eq. (26) into Eqs. (20)_1 and (21) gives the following expressions relating the longitudinal stretch \(\lambda_1\), the transverse stretch \(\lambda_2\) and the applied load \(P_{11}\):

\[
\frac{\mu - \mu_0}{kT} = \left[ \ln \left( 1 - \frac{1 - \phi_f}{\lambda_1 \lambda_2^2} \right) + \frac{1 - \phi_f}{\lambda_1 \lambda_2^2} + \frac{\chi (1 - \phi_f)^2}{\lambda_1^2 \lambda_2^4} \right] + \frac{C_0 \nu}{kT} \frac{1}{\lambda_1} \left( 1 - \frac{1}{\lambda_2^2} \right), \tag{27}
\]

and

\[
P_{11} = C_0 \left( \lambda_1 - \frac{\lambda_2^2}{\lambda_1^2} \right). \tag{28}
\]

The equilibrium is attained when \(\mu = \mu_0\). Figure 3 compares the transverse stretch \(\lambda_2\) and the "normalized" nominal stress \(P_{11}\) for imposed longitudinal stretch \(\lambda_1\) for various initial fluid volume fraction in the reference state \(\phi_f = 30, 60, 80\%\), at \(C_0 = 4 \times 10^4\) Pa. Figure 3 (left) shows that the transverse stretch \(\lambda_2\) is smaller as the value of the initial fluid content \(\phi_f\) gets greater. The isotropic stretch at swelling equilibrium \(\lambda^*_s\) for each reference state is obtained at the intersection with the dotted curve: we have \(\lambda^*_s = 3.7, 3.0, 2.3\) for \(\phi_f = 30, 60, 80\%\) respectively. This is consistent with the previous observations regarding the free-swelling case. Figure 3 (right) shows the traction/compression behaviour of the different materials from the swollen state \(\lambda_1/\lambda^*_s\). When the initial fluid content \(\phi_f\) increases, the longitudinal stress is smaller. Indeed, the material gets softer as the fluid contents gets higher.

![Graph](image1.png)

**Figure 3**: Transverse stretch (left) and "normalized" nominal stress (right) versus longitudinal stretch for three reference states.
4.2 Simplified human intervertebral disc

The human IVD consists in three main structures: a central cavity filled by the gelatino- nous NP, a surrounding AF, and the cartilaginous endplates that sandwich the whole. The finite element (FE) model is based on a simplified geometry: the NP and the AF are considered as coxial cylinders. The radius of the NP is set to 10 mm, the radius of the total IVD is set to 20 mm, the height of the AF and the NP is set to 12 mm, and the thickness of the endplates are 1 mm. Due to the numerous symmetries of the problem, only 1/8th of the geometry can be modelled. The endplate, much stiffer than the AF and the NP, are considered as rigid bodies. The NP is modelled using the gel-like model developed here, while the AF is modelled using an anisotropic hyperelastic Helzapfel strain energy potential, proposed in [14] and available in ABAQUS. The material parameters applied for the simulations are listed in Table 1. The inner NP and the outer AF are assumed to be tied together, as well as the top NP and AF faces with the endplate. As loading conditions, an uniaxial displacement of 1 mm is applied on the top endplate.

<table>
<thead>
<tr>
<th>Material</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF ground substance</td>
<td>$C_{10} = 0.05$ MPa</td>
<td>$D = 0$</td>
</tr>
<tr>
<td>AF fibers</td>
<td>$k_1 = 3$ MPa $^a$</td>
<td>$k_2 = 10$ $^a$</td>
</tr>
<tr>
<td>NP</td>
<td>$C_0 = 0.1$ MPa</td>
<td>$\chi = 0.1$</td>
</tr>
<tr>
<td></td>
<td>$\gamma = \pm 30^\circ$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 shows the mechanical response of the IVD. The order of magnitude of the FE results is consistent with the experimental data issued from [15].

![Graph showing axial compressive force versus axial displacement](image1)

**Figure 4:** Axial compressive force versus axial displacement (left) and radial displacement (right).

It should be noted that for this model, several assumptions were made; among them: (i) the geometry was considered cylindrical while the IVD has more a bean-like shape;
(ii) the AF material parameters were considered constant and uniform, while, as shown in [2, 16], the fibres orientation angle as well as the AF material parameters $k_1$ and $k_2$ strongly depend on the anatomical IVD region: (iii) the contacts between the different disc components were considered tied, which is probably not the case in reality. Taking into account all these realistic features is currently in progress.

5 CONCLUSIONS

In the present paper, a theoretical constitutive model for gel-like materials has been derived. The originality of this approach lies in the fact that the reference configuration is already swollen, undeformed and unconstrained. The model has been implemented in a finite element software, and successfully tested on simple homogeneous examples. Finally, the human nucleus pulposus was chosen as a complex example of application of this approach. Early simulations of a simplified intervertebral disc under compressive load show good promises regarding the prediction of the mechanical response of a healthy disc as compared as experimental data. As future work, it is believed that this model could be used to model disc degeneration through appropriate kinetik laws describing the change in stiffness and the fluid loss due to ageing.

REFERENCES


FUNDAMENTAL SOLUTIONS FOR A COUPLED FORMULATION OF POROUS BIPHASIC MEDIA WITH COMPRESSIBLE SOLID AND FLUID PHASES

R. SERPIERI*, F. TRAVASCIO† AND S. ASFOUR†

*Dipartimento di Ingegneria (DING)
Università degli Studi del Sannio
Piazza Roma, 21 - I, 82100, Benevento, Italy
e-mail: roberto.serpieri@unisannio.it, web page: http://www.ding.unisannio.it/

†Biomechanics Research Laboratory, College of Engineering
University of Miami
1251 Memorial Drive, MCA268, Coral Gables, FL, 33130, USA
e-mail: f.travascio@miami.edu, sasfour@miami.edu - web page: http://www.ie.miami.edu/brl/

Key words: porous media, compressibility, TMCPM, least action, poroelasticity

Summary. A general biphasic poroelastic formulation at finite strains with intrinsic compressibility of phases, whose governing equations are inferred on account of a least-action variational principle, has been recently proposed (TMCPM). Hereby, a theoretical, analytical, and numerical assessment is presented on the capability of linearized TMCPM to recover, in the limit of vanishing porosity, a traditional single phase continuum model.

1 INTRODUCTION

During the past sixty years, significant effort has been made for describing the mechanical behavior of fluid saturated poroelastic media [5]. Interest in this research field is motivated by the capability of this theory to describe the mechanics of several physical systems (e.g., consolidation of soils, biological tissues, polymeric foams, etc.) [9].

In applications where phases can be reasonably modeled as intrinsically incompressible continua, Incompressible Theory of Mixtures (ITM) [4, 2, 12] have found widespread deployment. In contrast, the achievement of a general consistent theory of poroelasticity, capable of addressing media with any degree and range of compressibility of the constituent phases, turns out to be a still intensively debated issue [6, 14, 11].

In biphasic theories, the two phases are most frequently treated as superposed continua each endowed with a separate energy potential. As shown by several authors (e.g., [17]), this general approach requires, alongside of traditional linear momentum and mass bal-
ances, an additional governing equation, simply referred to henceforth as closure equation, for which very differentiated candidates have been proposed.

In a study specifically addressing intrinsically compressible phases, tracing back to Cosserat’s theory [10], Bowen [3] identified this equation in momentum of momentum balance. Porosity was added as an additional independent state field by Wilmansky: in [19, 1] several additional balance equations have been investigated in the form of a porosity balance, or as an integrability condition for the deformation of the solid skeleton. In [8] and [6], a geometric saturation constraint is employed as a closure equation, and is combined with a multiplicative decomposition of the deformation gradient.

Alternatively, the lacking closure equations has been also inferred tracing back to the fundamental least-action principle of Hamiltonian mechanics. This approach was exploited in [13] to derive a porosity-associated additional governing equation by pairing an additional porosity field introduced as an independent state field.

A least-action principle was also exploited recently in [15, 16] to derive a biphasic formulation at finite strains that fully accounts for the compressibility of both phases. This framework, termed Theory of Microscopically Compressible Porous Media (TM-CPM), presents the following distinctive features: 1) the employment of an additional macroscopic state field of effective volumetric strain; 2) the identification of the lacking additional equation as the stationarity condition that is action-conjugated with such effective volumetric strain; 3) the explicit presence of the fluid mass balance among the governing PDEs, as an equation completely independent from the remaining ones.

The present study addresses a systematization of the linearized form of TMCPM within the range of infinitesimal perturbations. The linearized model is specifically investigated within quasi stationary conditions that address biphasic behavior under slow consolidation phenomena in soil mechanics or fast deformations in soft hydrated tissues. A reduced \( u - p \) form of governing equations is thus obtained, which differs from its ITM counterpart by the presence of a dimensionless microstructural parameter, termed \( \bar{k}_p \).

A simple homogenization procedure is proposed for retrieving estimates and bounds for this parameter. The estimates obtained are exploited to investigate the general behavior of the quasi stationary-system in the limit of vanishing porosity, to assess the consistent recovery of single-phase behavior of a solid continuum filling the entire space. This study is supported by analytical and numerical solutions of a set of stress relaxation test problems where the full range of porosity is spanned.

2 THEORETICAL BACKGROUND

2.1 Kinematics

Linearized TMCPM is briefly recalled. Volumetric fractions are customarily defined as \( \phi^\alpha = \frac{\hat{V}^{(\alpha)}}{\bar{V}} \) (\( \alpha = s, f \)), where \( \bar{V}, \hat{V}^{(s)}, \) and \( \hat{V}^{(f)} \) represent, respectively, the volumes of the mixture, of the solid phase, and of the fluid phase contained in a Representative Volume
Element (RVE). At reference configuration, the mass distribution is defined by the fields $\bar{\rho}^\alpha_o$ and $\hat{\rho}^\alpha_o$, which represent apparent and true mass densities of the components of the mixture. The complete saturation hypothesis implies the two following relations:

$$\phi^f_o + \phi^s_o = 1, \quad \bar{\rho}^\alpha_o = \phi^\alpha_o \hat{\rho}^\alpha_o, \quad \alpha = s, f$$

(1)

A second configuration of the solid phase, infinitesimally displaced from the reference one, is macroscopically described by the displacement field $\bar{\mathbf{u}}^{(s)}$. For this configuration, linearization of equations (1) immediately yields:

$$d\phi^f = -d\phi^s, \quad d\bar{\rho}^\alpha = d\phi^\alpha \hat{\rho}^\alpha + \phi^\alpha d\hat{\rho}^\alpha, \quad \alpha = s, f$$

(2)

Characteristic feature of TMCPM is the presence of two volumetric strain measures associated with each phase composing the mixture: the apparent volumetric dilatation $\bar{e}^\alpha$, and the effective volumetric dilatation $\hat{e}^\alpha$. The latter is introduced as a primary state field completely independent from the former. In linearized TMCPM, as a result of linearization of finite volumetric strains, $\bar{e}^s$ and $\hat{e}^s$ are defined as follows:

$$\bar{e}^s = \frac{d\bar{V}}{\bar{V}}, \quad \hat{e}^s = \frac{d\hat{V}^{(s)}}{\hat{V}^{(s)}}$$

(3)

In particular, the apparent solid volumetric dilatation is the trace of the solid strain tensor: $\bar{e}^s = \text{tr} \bar{\varepsilon}^{(s)} = \nabla \cdot \bar{\mathbf{u}}^{(s)}$. Volumetric strains are related to density differentials by:

$$\bar{e}^\alpha = -\frac{d\bar{\rho}^\alpha}{\bar{\rho}^\alpha_o}, \quad \hat{e}^\alpha = -\frac{d\hat{\rho}^\alpha}{\hat{\rho}^\alpha_o}, \quad \alpha = s, f$$

(4)

In addition, a relationship among the variation of the fluid volumetric fraction, and the apparent and intrinsic solid deformation has been reported [16]:

$$d\phi^f = (\bar{e}^s - \hat{e}^s)\phi^s_o$$

(5)

Combination of equations (2), (4) and (5) allows deriving a fundamental kinematic relation in linearized TMCPM relating apparent and intrinsic deformations of the mixture:

$$\phi^f_o \bar{e}^f + \phi^s_o \hat{e}^s = \phi^f_o \hat{e}^f + \phi^s_o \hat{e}^s$$

(6)

This equation, derived in [16] for the special case $\phi^f_o = \phi^s_o = 0.5$, represents a dimensionless saturation constraint for a biphasic compressible medium.

2.2 Mechanics

A hyperelastic isotropic formulation is recalled. A macroscopic solid strain energy density is introduced as a function of the infinitesimal strain measures of solid $\bar{\psi}^{(s)} =$
The stress measures are the \textit{extrinsic} stress tensor, $\tilde{\sigma}^{(s)}$, conjugated to $\tilde{\varepsilon}^{(s)}$, and a single scalar stress conjugated to $\tilde{\varepsilon}^{s}$, termed \textit{intrinsic solid stress}, $\tilde{\sigma}^{s}$:

$$\tilde{\sigma}^{(s)} = \frac{\partial \tilde{\psi}^{(s)}}{\partial \tilde{\varepsilon}^{(s)}}, \quad \tilde{\sigma}^{s} = \frac{\partial \tilde{\psi}^{s}}{\partial \tilde{\varepsilon}^{s}}$$  \hspace{1cm} (7)

As previously reported [16], a volumetric-deviatoric stress-strain uncoupling yields the following representation for solid stress in the isotropic case

$$\tilde{\sigma}^{(s)} = 2\tilde{\mu}\tilde{\varepsilon}^{(s)} + \tilde{Z}\tilde{\varepsilon}^{s}I + \tilde{k}_{sM}\tilde{\varepsilon}^{s}I, \quad \tilde{\sigma}^{s} = \tilde{k}_{sM}\tilde{\varepsilon}^{s} + \tilde{k}_{s}\tilde{\varepsilon}^{s}$$  \hspace{1cm} (8)

where $\tilde{\mu}$ is a shear modulus while $\tilde{Z}$, $\tilde{k}_{sM}$ and $\tilde{k}_{s}$ are all volumetric moduli.

Introducing intrinsic solid pressure as $\tilde{p}^{(s)} = -\frac{\partial \tilde{\psi}^{(s)}}{\partial \varepsilon^{s}} = -\tilde{\sigma}^{s}$, with the aid of a customary volumetric-deviatoric split and of relationship $\tilde{Z} = \tilde{\lambda} + \frac{\tilde{k}_{sM}^{2}}{\tilde{k}_{s}}$ [16], the elastic constitutive law for the solid phase can thus be conveniently represented in the alternate uncoupled form in terms of Lamé coefficients $\tilde{\mu}$ and $\tilde{\lambda}$:

$$\tilde{\sigma}^{(s)}_{\text{dev}} = 2\tilde{\mu}\tilde{\varepsilon}^{(s)}_{\text{dev}}, \quad \begin{bmatrix} \tilde{p}^{(s)} \\ \tilde{p}^{(s)} \end{bmatrix} = -\begin{bmatrix} \tilde{K}^{(s)}_{\text{iso}} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}^{(s)} \\ \tilde{\varepsilon}^{s} \end{bmatrix}, \quad \text{with} \quad \begin{bmatrix} \tilde{K}^{(s)}_{\text{iso}} \end{bmatrix} = \begin{bmatrix} \left( \frac{3}{2}\tilde{\mu}\tilde{\lambda} + \frac{\tilde{\lambda}^{2}}{\tilde{k}_{s}} \right) & \tilde{k}_{sM} \\ \tilde{K}_{sM} & \tilde{k}_{s} \end{bmatrix}$$  \hspace{1cm} (9)

The macroscopic strain energy density of the fluid is:

$$\tilde{\psi}^{(f)} = \phi_{0}^{f}\tilde{\psi}^{(f)} = \phi_{0}^{f}\tilde{k}_{f}(\tilde{e}^{f})^{2}$$  \hspace{1cm} (11)

where $\tilde{\psi}^{(f)}$ is the \textit{interstitial} strain energy density and $\tilde{k}_{f}$ is the fluid bulk modulus.

Momentum equations are hereby recalled by directly neglecting inertia terms under the assumption of quasi-stationarity. The linear momentum balance of solid is reported assuming two simplificative hypotheses: the medium is homogeneous in space, so that the porosity gradient vanishes and, as a consequence, also the volume force term responsible for buoyancy forces as shown in [16]. Moreover, all external body forces, such as gravitational, are neglected. Consequently, the drag $\pi^{fs}$ exerted by the fluid phase over the solid is the only nonvanishing volume force applied on the solid phase:

$$\nabla \cdot \tilde{\sigma}^{(s)} + \pi^{fs} = 0$$  \hspace{1cm} (12)

In particular, a simple linear Darcy law $\pi^{fs} = K(v^{f} - \frac{\partial u^{(s)}}{\partial t})$, with $K$ being the friction coefficient, is considered.

Characteristic feature of TMCPM is the presence of an additional momentum balance equation which, on account of a least-action principle, stems from imposing stationarity of the Action functional with respect to the effective solid volumetric strain $\tilde{\varepsilon}^{s}$:

$$\frac{\partial \tilde{\psi}^{(s)}}{\partial \tilde{\varepsilon}^{s}} + \frac{\partial \tilde{\psi}^{(f)}}{\partial \tilde{\varepsilon}^{s}} + \rho_{0} \frac{\partial^{2} \tilde{\varepsilon}^{s}}{\partial t^{2}} = 0$$  \hspace{1cm} (13)
In (13), $\rho_0^*$ is an additional inertia term associated with the acceleration of the parameter $\dot{e}^s$. The previous equation comes alongside of the classical linear solid momentum balance (see equation (12)) which, instead, expresses stationarity with respect to infinitesimal variations of the displacement field of solid. The interested reader is referred to [15] for a complete account of these theoretical issues. Balance (13) accounts for volumetric coupling between phases. Rewritten in terms of pressure-like quantities with neglected inertia forces it reads:

$$-\ddot{p}^{(s)} - \ddot{p}^{(fs)} = 0$$

(14)

where $\ddot{p}^{(s)}$ has been previously defined, and $\ddot{p}^{(fs)} = -\frac{\partial \hat{\psi}^{(f)}}{\partial \hat{e}^s}$ is an interaction term between the two phases. On account of (6), (11), and of the chain rule, this last term is:

$$\ddot{p}^{(fs)} = -\phi_f^o \frac{\partial \hat{\psi}^{(f)}}{\partial \hat{e}^f} \frac{\partial \hat{e}^f}{\partial \hat{e}^s} = -\phi_s^o \hat{p}$$

(15)

Thus, due to the second of (10), (14) achieves the following form:

$$\bar{k}_s M \ddot{e}^s + \bar{k}_s \dot{e}^s - \phi_s^o \bar{k}_f \dot{e}^f = 0$$

(16)

Concerning the momentum balance of the fluid phase, this is expressed with respect to the reference frame of the solid phase. Consistent with the assumption of neglecting body forces other than drag (for the fluid $\pi^{sf} = -\pi^{fs}$), as well as inertia terms, this reads:

$$\phi_f^o \nabla p + \pi^{fs} = 0$$

(17)

Concerning mass balances, mass generation/depletion is neglected for simplicity. Notice that mass balance of solid phase is not included in the main set of governing equations since solid density variation $d\bar{\rho}^s$ is a secondary variable, whose value is related to $\dot{e}^s$ by $d\bar{\rho}^s = -\bar{\rho}_0^s \dot{e}^s$. Conversely, fluid mass balance is included among the governing equations and, as a consequence of linearization, admits the simple dimensionless form [16]:

$$\frac{\partial \dot{e}^f}{\partial t} = \nabla \cdot \mathbf{v}^f$$

(18)

The set of governing equations so far derived for the quasi-stationary system amounts to 8 independent scalar PDEs in the primary fields $\mathbf{u}^{(s)}$, $\dot{e}^s$, $\mathbf{v}^f$, $p$ (8 scalar fields overall).

However, a simpler form of the governing set, where all quantities are expressed in terms of solid displacement $\mathbf{u}^{(s)}$ and fluid pressure $p$ fields, can be conveniently achieved by means of few manipulations hereby synthesized. This equivalent formulation will be referred to as reduced quasi-stationary $u - p$ form. To this end, a dimensionless ratio of the solid moduli $\tilde{k}$, and a further modulus $\tilde{k}_s$ are introduced as $\tilde{k}_r = \frac{\phi^o k_{rs}}{k_s}$ and $\tilde{k}_s = \frac{k_s}{\phi^o}$.

A first equation, that may be viewed as a combined linear momentum balance of the mixture, is obtained adding (12) and (17), whereby drag forces cancel each other:

$$\nabla \cdot \tilde{\mathbf{\sigma}}^{(s)} - \phi_f^o \nabla p = 0$$

(19)
Moreover, upon performing a partial Legendre transform of the constitutive equation (10) with respect to \( \varepsilon^s \), and by also recalling relation \( Z = \lambda + \frac{k_s^2}{k_s} \) together with equations (8), a constitutive law is obtained, in which \( \varepsilon^s \) is replaced by \( p \):

\[
\mathbf{\sigma}^{(s)}(\varepsilon^{(s)}, \varepsilon^s) \rightarrow \mathbf{\sigma}^{(s)}(\varepsilon^{(s)}, p) = 2\mu\varepsilon^{(s)} + \bar{\lambda}\varepsilon^s \mathbf{I} - \bar{k}_r p \mathbf{I}
\]

(20)

Using (20), the following restated form of the linear momentum balance of the whole mixture is obtained, which is the first equation of the quasi-stationary reduced \( \mathbf{u} - p \) set:

\[
\nabla \cdot \mathbf{\sigma}^{(s)} - \left( \phi_o^e + \bar{k}_r \right) \nabla p = 0, \quad \text{where } \mathbf{\sigma}^{(s)} = \bar{\lambda}\varepsilon^s \mathbf{I} + 2\mu\varepsilon^{(s)}
\]

(21)

The second equation of the reduced \( \mathbf{u} - p \) set is obtained by substituting (18) and (16) into (6). The resulting equation is finally combined with the divergence of the fluid momentum balance (17) to yield an equation expressing, altogether, \( \varepsilon^e \)-associated momentum balance, \( \varepsilon^s \)-associated momentum balance and fluid linear momentum balance:

\[
(1 + \bar{k}_r) \frac{\partial}{\partial t} \nabla \cdot \mathbf{\bar{u}}^{(s)} + \left( \frac{\phi_o^e}{\bar{k}_s} + \frac{\phi_f^e}{\bar{k}_f} \right) \frac{\partial p}{\partial t} - \left( \phi_f^e \right)^2 K \nabla^2 p = 0
\]

(22)

The sought reduced \( \mathbf{u} - p \) form of equations governing the quasi-stationary behavior in linearized TMCPM is composed of (21) and (22).

3 ESTIMATES OF THE MODULI

For a meaningful assessment of the compressible biphasic model at hand, it is fundamental to dispose of estimates of the moduli \( \bar{k}_s, \bar{k}_sM \) in (10) and of \( \bar{k}_e \) in (21)-(22). Basic estimates are hereby obtained by introducing simple geometrical/topological assumptions on the microstructural geometry of the RVE. With this perspective, hypotheses of spherical symmetry of the RVE, and isotropy of the constituent material are now introduced.

A simple hollow spherical geometry, with exterior radius \( R_e \) and interior radius \( R_i \), is considered. Figure 1a shows the diametral section of the hollow sphere. The solid domain is subjected to an internal, \( p_i \), and an external pressure \( p_e \). The solution for this elastostatic problem is available in spherical coordinates [18]. Denoting by \( r \) the radial coordinate, the outward radial displacement \( u \) is:

\[
u = - \frac{(1 - 2\nu)}{2(1 + \nu)} \frac{p_e R_i^3 - p_i R_e^3}{R_e^3 - R_i^3} r - \frac{1}{4\mu} \frac{(p_e - p_i) R_e^3 R_i^3}{R_e^3 - R_i^3} \frac{1}{r^2}
\]

(23)

where \( \mu \) and \( \nu \) are the shear modulus and Poisson ration of the material that constitutes the solid phase. Setting \( r = R_e \) and \( r = R_i \) in (23), one respectively obtains the displacements on the exterior, \( u_e \), and interior, \( u_i \), boundaries. Since the relation between the pressure vector \( \mathbf{p} = [p_e, p_i]^t \) and the displacements on the external and internal surfaces of the sphere \( \mathbf{u} = [u_e, u_i]^t \) is linear, this can be written in matrix form as:

\[
\mathbf{u} = \mathbf{D}_{up} \mathbf{p}
\]

(24)
where, upon setting $\alpha = 4\mu(1+\nu)(R_e^3 - R_i^3)$, $D_{up}$ turns out to be the following matrix:

$$D_{up} = -\frac{1}{\alpha} \begin{bmatrix} R_e^2(1 + \nu) + 2R_i^2(1 - 2\nu) & 3(\nu - 1)R_eR_i^3 \\ -3(\nu - 1)R_iR_e^3 & R_i^2(1 + \nu) + 2R_e^2(1 - 2\nu) \end{bmatrix}$$ (25)

Volumetric strains are collected in the vector $\bar{e} = [\tilde{\bar{e}}^s, \tilde{\bar{e}}^a]$. A relation between $\bar{e}$ and $\mathbf{u}$ is now sought. Suitable expressions for $\tilde{\bar{e}}^a$ and $\tilde{\bar{e}}^s$ are inferred from definition (3), by considering that $\bar{V}$ is the volume of the external sphere while $\tilde{\bar{V}}(s)$ is the volume of the hollow sphere (i.e. of the subdomain containing only the solid phase), so that $\bar{V} = \frac{4}{3}\pi R_e^3$ and $\tilde{\bar{V}}(s) = \frac{4}{3}\pi (R_e^3 - R_i^3)$. The infinitesimal variations $d\bar{V}$ and $d\tilde{\bar{V}}(s)$ can be computed as function of $u_e$ and $u_i$ by retaining, as usual in infinitesimal displacements, only first-order terms. Accordingly (3) provide:

$$\tilde{\bar{e}}^s = \frac{(R_e + u_e)^3}{R_e^2} \simeq 3\frac{u_e}{R_e}, \quad \tilde{\bar{e}}^a = \frac{(R_e + u_e)^3 - (R_i + u_i)^3}{(R_e^3 - R_i^3)} \simeq 3\frac{R_e^2u_e - R_i^2u_i}{(R_e^3 - R_i^3)}$$ (26)

As a check of the formulas so far introduced, observe incidentally that when equal external and internal pressures are applied (i.e. setting $\mathbf{p} = [p, p]^t$ in (24)), the cell experiments a homothety. In this case, the external and internal displacements are:

$$u_e = -R_e\frac{1}{2\mu} \frac{1 - 2\nu}{1 + \nu} p, \quad u_i = -R_i\frac{1}{2\mu} \frac{1 - 2\nu}{1 + \nu} p$$ (27)

and, according to (26), the strains are $\tilde{\bar{e}}^a = \tilde{\bar{e}}^s = \frac{p}{k_s}$, where $k_s = \frac{2(1+\nu)}{1-\nu}\mu$ is the bulk modulus of the solid constituent material. It is convenient to express (26) in the matrix form $\mathbf{e} = A_{eu}\mathbf{u}$, where:

$$A_{eu} = \begin{bmatrix} 3\frac{1}{R_e} & 0 \\ 3 - \frac{R_e^2}{(R_e^3 - R_i^3)} & -3\frac{R_i^2}{(R_e^3 - R_i^3)} \end{bmatrix}$$ (28)

The vector of stress measures that is energy-conjugated with $\mathbf{e}$ is $\mathbf{p}^{(se)} = -[\tilde{\mathbf{p}}^{(s)}, \tilde{\mathbf{p}}^{(s)}]^t$ (recall (7) and (10)). Energy conjugation implies that strain energy of the finite spherical cell $\psi_{finite}^{(s)}$ can be expressed in the form:

$$\tilde{\psi}_{finite}^{(s)} = \frac{1}{2}\mathbf{p}^{(se)} \cdot \mathbf{e}$$ (29)

It is also convenient to use, in place of the displacements $\mathbf{u}$, a couple of displacement parameters $\tilde{\mathbf{u}} = [4R_e^2\pi u_e, 4R_i^2\pi u_i]^t$ that includes the relevant surface area. In this case, energy conjugation with the pressures $\mathbf{p}$ can be written in the form:

$$\tilde{\psi}_{finite}^{(s)} = \frac{1}{2}\tilde{\mathbf{p}} \cdot \tilde{\mathbf{u}}$$ (30)
where $\mathbf{I}$ is a matrix that accounts for the opposite orientation of tractions associated with positive pressures on the external and internal surfaces $\tilde{\mathbf{I}} = [\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}]$. As one may check, the relationship between $\mathbf{e}$ and $\tilde{\mathbf{u}}$ is

$$\mathbf{e} = \tilde{\mathbf{A}}_{eu} \tilde{\mathbf{u}}$$

with $\tilde{\mathbf{A}}_{eu} = \frac{3}{4\pi} \begin{bmatrix} \frac{1}{R^2_e} & 0 \\ \frac{1}{R^2_e - R^2_i} & \frac{1}{R^2_e - R^2_i} \end{bmatrix}$ (31)

Accordingly, the stress strain relation in terms of $\tilde{\mathbf{u}}$ and its inverse are:

$$\tilde{\mathbf{u}} = \tilde{\mathbf{D}}_{up} \mathbf{p}$$

$$\mathbf{p} = \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{u}}$$

where $\tilde{\mathbf{D}}_{up}$ is the following matrix:

$$\tilde{\mathbf{D}}_{up} = \frac{4\pi}{\alpha} \begin{bmatrix} R^2_e [R^3_e (1 + \nu) + 2R^3_i (1 - 2\nu)] & 3(\nu - 1) R^3_i R^3_e \\ -3(\nu - 1) R^3_i R^3_e & R^3_i [R^3_e (1 + \nu) + 2R^3_i (1 - 2\nu)] \end{bmatrix}$$ (33)

Thus, on account of (31), the pressure vector turns out to be:

$$\mathbf{p} = \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{A}}_{eu}^{-1} \mathbf{e}$$ (34)

Equating (29) to (30) provides $\frac{1}{2} \mathbf{p}^{(se)} \cdot \mathbf{e} = \frac{1}{2} \tilde{\mathbf{I}} \mathbf{p} \cdot \tilde{\mathbf{u}}$. A substitution of (34) in this last identity provides $\mathbf{p}^{(se)} \cdot \mathbf{e} = \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{A}}_{eu}^{-1} \mathbf{e} \cdot \tilde{\mathbf{u}}$. Upon transposing $\tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{A}}_{eu}^{-1}$, one has:

$$\mathbf{p}^{(se)} \cdot \mathbf{e} = \tilde{\mathbf{A}}_{eu}^{-1} \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{I}} \mathbf{e}$$ (35)

and since (35) must hold for any $\mathbf{e}$, one infers $\mathbf{p}^{(se)} = \tilde{\mathbf{A}}_{eu}^{-1} \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{I}} \tilde{\mathbf{u}}$. Finally, on account of (31), one has:

$$\mathbf{p}^{(se)} = \bar{\mathbf{K}}_{iso/finite}^{(s)} \mathbf{e}$$

with $\bar{\mathbf{K}}_{iso/finite}^{(s)} = \tilde{\mathbf{A}}_{eu}^{-1} \tilde{\mathbf{D}}_{up}^{-1} \tilde{\mathbf{I}} \tilde{\mathbf{A}}_{eu}^{-1}$ (36)

where $\bar{\mathbf{K}}_{iso/finite}^{(s)}$ is the estimate of $\bar{\mathbf{K}}_{iso}^{(s)}$ for a finite sphere, defined as:

$$\begin{bmatrix} \bar{\mathbf{K}}_{iso/finite}^{(s)} \end{bmatrix} = \frac{16\pi \mu}{9} \begin{bmatrix} \frac{R^3_e (R^3_e - R^3_i)}{R^3_i} & -\frac{R^3_i (R^3_e - R^3_i)}{R^3_i} \\ -\frac{R^3_i (R^3_e - R^3_i)}{R^3_i} & (R^3_e (1 + \nu) + 2R^3_i (1 - 2\nu)) \end{bmatrix}$$ (37)

This stiffness matrix has to be homogenized in space. Accordingly, it is divided by the volume of the external sphere $\frac{4}{3} \pi R^3_e$. Expressing the resulting matrix in terms of volumetric fractions, which for the spherical cell the volumetric fractions are trivially related to the radii by $\frac{R^3_e}{R^3_i} = \phi^s_e$ and $\frac{R^3_i}{R^3_e} = 1 - \phi^s_e$, one finally obtains:

$$\begin{bmatrix} \bar{\mathbf{K}}_{iso}^{(s)} \end{bmatrix} = \frac{4}{3} \mu \frac{\phi^s_e}{1 - \phi^s_e} \begin{bmatrix} 1 & -1 \\ -1 & 2(1 - 2\nu) \end{bmatrix} \begin{bmatrix} \phi^s_e \\ 1 - \phi^s_e \end{bmatrix} = \phi^s_e \begin{bmatrix} \frac{4}{3} \mu & -\frac{4}{3} \mu \\ -\frac{4}{3} \mu & \frac{4}{3} \mu \end{bmatrix} \begin{bmatrix} \phi^s_e \\ 1 - \phi^s_e \end{bmatrix}$$ (38)
where the last equality holds since, from the theory of elasticity, one has 
\[
\frac{1 + \nu}{2(1 - 2\nu)} = \frac{3}{4} \frac{k_r}{\mu}.
\]

The ratio \( \tilde{k}_r \) and modulus \( \tilde{k}_s \) turn out to be:
\[
\tilde{k}_r = -\frac{\phi^s s \mu}{\frac{3}{2} \mu + k_s (1 - \phi^s)};
\]
\[
\tilde{k}_s = \frac{1}{1 - \phi^s} \left[ \frac{4}{3} \mu + k_s (1 - \phi^s) \right].
\]

(39)

The contour plot of \( \tilde{k}_r \) as function of \( \phi^s \) and \( \nu \) is shown in Figure 1 whereby the bounds 
\(-1 \leq \tilde{k}_r \leq 0 \) can be also visualized. In particular, the upper bound \( 0 \) is attained in the limit of vanishing solidity \( \phi^s = 0 \), and when the solid constituent material has \( \nu = 0.5 \) (i.e. is volumetrically incompressible). In contrast, the lower bound is attained at \( \phi^s = 1 \).

![Figure 1](image)

Figure 1: (a) diametral section of the solid spherical cell; (b) contour plot of \( \tilde{k}_r \) as function of \( \nu \) and \( \phi^s \).

4 LIMIT BEHAVIOR AT VANISHING POROSITY

Estimates (39) for \( \tilde{k}_r \) and \( \tilde{k}_s \) are now deployed to analyze the behavior of the \( u - p \) set in the limit of vanishing porosity (LVP) (i.e. as \( \phi^s \rightarrow 1 \)). This situation amounts to achievement of the limit of dilute suspension of the fluid phase in the solid matrix, and needs to be handled with special attention since all elastic coefficients in matrix (38) tend to infinity and, furthermore, equation (22) becomes clearly an identity \( 0 = 0 \). However, as shown below, it can be deduced with few logical passages that the behavior of a single continuum made by the sole solid phase is duly recovered.

Actually, upon introducing the compliance coefficient
\[
\frac{1}{\tilde{k}_{sf}} = \left( \frac{\phi^s s \mu}{k_s} + \frac{\phi^f f \mu}{k_f} \right);
\]
the order of the infinitesimal factors in (22) can be compared by taking the ratio of (22) over \( \frac{1}{\tilde{k}_{sf}} \):
\[
\left( 1 + \tilde{k}_r \right) \frac{\partial}{\partial t} \nabla \cdot \bar{u}^{(s)} + \frac{\partial p}{\partial t} - \tilde{k}_{sf} \left( \phi^f \right)^2 \nabla^2 p = 0 \quad (40)
\]

Since, as \( \phi^f \rightarrow 0 \), one has \( \tilde{k}_r \rightarrow -1 \) and \( \tilde{k}_{sf} \rightarrow \infty \), the first and third coefficients in (40) have both undetermined limit of the form \( [\infty \cdot 0] \). However, with (39) one computes:
\[
1 + \tilde{k}_r = (1 - \phi^s) \frac{\frac{3}{2} \mu + k_s}{\frac{3}{2} \mu + (1 - \phi^s) k_s};
\]
\[
\tilde{k}_{sf} = \frac{1}{(1 - \phi^s) \frac{3}{2} \mu + (1 - \phi^s) k_s + \phi^s \hat{k}_f}
\]

(41)
so that the undetermined forms can be evaluated, and the following limits are computed:

\[
\hat{k}_{sf} (1 + \bar{k}_r) = \frac{(\frac{4}{3} \mu + k_s)^2 \bar{k}_f}{(\frac{4}{3} \mu + \bar{k}_f) \frac{4}{3} \mu}, \quad \hat{k}_{sf} (\phi_o^f)^2 = 0
\]

(42)

Hence \( \hat{k}_{sf} (1 + \bar{k}_r) \) holds a finite value, and (40) becomes

\[
\frac{\partial}{\partial t} \left[ \hat{k}_{sf} (1 + \bar{k}_r) \nabla \cdot \mathbf{u}^{(s)} + p \right] = 0.
\]

Assuming that, at initial time \( t_0 \), the body is in an undeformed state with zero pressure, the previous equation can be integrated and solved, providing:

\[
p = -\hat{k}_{sf} (1 + \bar{k}_r) \bar{e}^s
\]

(43)

Substitution of (43) into (19) yields the classical equilibrium equation for Cauchy continuum \( \nabla \cdot \mathbf{\sigma}^{(s)} = 0 \). Moreover, taking the divergence of (17), upon setting \( \phi_o^f = 0 \) in (17), one infers \( \bar{e}^o = \bar{e}^f \). Substitution of this last relation and of (43) into the dimensionless saturation constraint (6) yields:

\[
\bar{e}^o = \frac{1}{\phi_o^s} \left[ 1 - \frac{\phi_o^f}{\hat{k}_f} \hat{k}_{sf} (1 + \bar{k}_r) \right] \bar{e}^s
\]

(44)

Setting \( \phi_o^s = 1 \) in (44), it is thus recognized that in the limit of vanishing porosity the only admissible path for volumetric strains of the solid phase is \( \bar{e}^o = \bar{e}^s \).

On the other hand, from (38) one infers that at limit vanishing porosity \( \bar{k}_s = -\bar{k}_{sM} \) and, consequently, in such a limit the elastic law (8) specializes to

\[
\mathbf{\sigma}^{(s)} = 2\bar{\mu} \bar{e}^{(s)} + \left[ \left( \lambda - \bar{k}_{sM} \right) \bar{e}^s + \bar{k}_{sM} \bar{e}^o \right] \mathbf{I}
\]

(45)

On account of the property at LVP just found \( (\bar{e}^o = \bar{e}^s) \), the limit of (45) can be computed. It is then recognized that a traditional isotropic law is also recovered.

5 NUMERICAL AND ANALYTICAL EXAMPLES

The quasi-stationary behavior of the biphasic compressible medium was investigated as function of porosity also numerically. Solutions for uni-axial compressive stress relaxation tests problems were computed to obtain a family of solid stress curves parametric with \( \phi_o^s \). Figs. 2a and 2b show the experimental setup and the relevant boundary conditions.

Note that, for the specific simulated testing conditions, the reduced \( \mathbf{u} - p \) form simplifies to a system of two scalar PDEs in the unknown functions \( \tilde{u}_x^{(s)}(x, t), p_x(x, t) \) with \( x \in [0, L] \):

\[
\frac{\partial \sigma_x^{(s)}}{\partial x} - (\phi_o^s + \bar{k}_r) \frac{\partial p}{\partial x} = 0, \quad (1 + \bar{k}_r) \frac{\partial^2 \tilde{u}_x^{(s)}}{\partial t \partial x} + \left( \phi_o^s \phi_o^f \right) \frac{\partial p}{\hat{k}_s} + \left( \phi_o^s \phi_o^f \right) \frac{\partial^2 \tilde{u}_x^{(s)}}{\partial t^2} - \frac{(\phi_o^f)^2}{K} \frac{\partial^2 p}{\partial x^2} = 0
\]

(46)

where \( \sigma_x^{(s)} = (\lambda + 2\mu) \frac{\partial u_x^{(s)}}{\partial x} \) whereas \( \tilde{\sigma}_x^{(s)} \) and \( \tilde{u}_x^{(s)} \) are the components of the drained solid stress and the solid displacement in the \( x \)-direction, respectively. Analytical solutions
for the system of equations (46) were obtained in the complex Laplace space. Subsequently, solutions in the real space were achieved by numerical computation of Laplace anti-transforms via de Hoog et al’s algorithm [7]. The time dependent behavior of the solid stress is reported in Fig. 2, parametric with the solid volumetric fraction $\phi^s_0$. Note that, for the specific problem at hand, the term $\tilde{\sigma}^s_x(L, t)$ corresponds to the stress at the plug. Numerical results indicate that, in the LVP ($\phi^f_0 = 0$), the biphasic system recovers the ramp-and-hold behavior of a Cauchy solid. This is in agreement with the theoretical predictions of Section 4.

Figure 2: (a) Schematic of the simulated experimental setup for the uni-axial compressive stress-relaxation test: a biphasic sample is confined in an impermeable chamber and compressed axially by a porous plug allowing for fluid exudation. (b) Boundary conditions for the simulated experiment: compression of the biphasic sample is performed by applying a compression of 5% of the sample via a ramp-and-hold displacement of the plug. (c) Time dependent response of solid stress at the plug parametric with $\phi^f_0$.

6 CONCLUSIONS

In this contribution, an assessment of the consistency of TMCPM was provided showing that linearized TMCPM successfully recovers, in the limit of vanishing porosity, a traditional single phase continuum Cauchy model. This result was obtained with the aid of a homogenization technique that, although simple, provides insights on the coefficients $\bar{k}_r$ and $\hat{k}_s$ appearing in the $u - p$ form of the quasi-stationary linearized theory. Overall, this contribution shows that TMCPM is a viable theoretical framework for modeling the mechanical behavior of intrinsically compressible biphasic systems.

REFERENCES


HIGH PERFORMANCE CONJUGATE HEAT TRANSFER WITH THE OPENPALM COUPLER.

Florent Duchaine*, Stéphan Jauré*, Damien Poitou*, Eric Quémerais†, Gabriel Staffelbach*, Thierry Morel* and Laurent Gicquel*

*CERFACS, 42 avenue G. Coriolis, 31 057 Toulouse Cedex 01, France
e-mail: florent.duchaine@cerfacs.fr

†ONERA, DSNA, 29 Avenue de la Division Leclerc 92 322 Châtillon Cedex, France

Key words: Massively Parallel, Conjugate Heat Transfer, Large Eddy Simulation

Abstract. Optimizing gas turbines is a complex multi-physical and multi-component problem that has long been based on expensive experiments. Today, computer simulations can reduce design process costs and are acknowledged as a promising path for optimization. Although the simulations of specific components of gas turbines become accessible, these stand-alone simulations face a new challenge: to improve the quality of the results, new physics must be introduced. Based on the simulation of conjugate heat transfer within an industrial combustor to predict the temperature of its walls, the current work aims at studying the scalability of code coupling on HPC architectures. Coupling accurately solvers on massively parallel architectures while maintaining their scalability is challenging. The strategy investigated relies on recent developments made in a generic parallel coupler. Performance tests have been carried out until 12,288 cores on the CURIE supercomputer (TGCC / CEA). Results show a good behavior and advanced analyzes are carried out in order to draw new paths for future developments in coupled simulations.

1 Introduction

Determination of heat loads such as wall temperatures and heat fluxes, is a key issue in gas turbine design [1, 2]. With the constant increase of computing power, numerical simulations of the thermal interaction between fluid flows and solids offer new design paths to diminish development costs through important reductions of the number of experimental tests. To determine mean heat loads on structures, many authors use Conjugate Heat Transfer (CHT) where the fluid and solid equations are resolved simultaneously. CHT is a difficult field and most existing tools are developed for chained (rather than coupled),
steady (rather than transient) phenomena thanks to Reynolds Averaged Navier-Stokes (RANS) solvers [3, 4]. The accuracy of the CHT predictions largely relies on the Computational Fluid Dynamics (CFD) method. Recent contributions based on Large Eddy Simulation (LES) [5, 6] provide promising results especially for the prediction of heat transfer in complex geometries [7, 8, 9, 10].

Using an unsteady LES flow solver to resolve such problems raises several complexities to address for CHT. LES requires high mesh resolutions to accurately capture the flow physics. It is also more CPU consuming than RANS methods to converge spatial and temporal statistics. These specificities imply the use of specific strategies to accelerate the convergence toward steady heat transfer problems as well as efficient methods to use existing high performance architectures. Previous studies have proposed guidelines to reach stable convergence of CHT problems with LES based on time desynchronization of convection and conduction as well as the use of high frequency information exchanges between the physics [7, 11].

Note that, there are two basic approaches to numerically solve CHT problems. The first one is a direct coupling approach where the different physics are solved simultaneously in a large system of equations by a monolithic solver. The second approach consists in solving each set of equations separately with dedicated solvers that exchange interface conditions through a coupler. The last solution adopted here has the advantage of using existing state-of-the-art codes to solve fluid and solid equations.

In this context and with the convergence recommendations, the resolution of CHT problems puts a lot of pressure on the tool used to couple the solvers. Several communities have investigated the use of code coupler in many different area ranging from climate studies to industrial applications. These communities are now faced to the challenge of running the coupled applications with highly loaded codes on massively parallel machines where the solvers exchange a lot of amount of data at a high frequency. The strategy investigated in this work to address these issues relies on recent developments made in a generic parallel coupler [12]. The first section presents the fluid and solid solvers. Then, implementation details on the coupling library are provided. Finally performance tests carried out until 12,288 cores on the CURIE supercomputer (TGCC / CEA) are proposed.

2 Presentation of the flow and conduction solvers

This section aims at presenting the fluid and solid solvers used to construct the CHT tool. Both solvers have a mesh partitioning based parallelism. After a short description of their functionalities, parallel performance of each solver is given.

2.1 The fluid solver AVBP

Recent information about this flow solver can be found in [13]. AVBP solves the compressible Navier-Stokes equations and focuses on unsteady turbulent flows (with and without chemical reactions) for internal flow configurations on unstructured grids. The
prediction of turbulent flows is based on the LES sub-grid scale closure problem [14]. The numerical methods are either 2nd order or 3rd order in time and space [13]. AVBP library includes integrated parallel domain partitioning and data reordering tools, handles message passing and includes supporting routines for dynamic memory allocation, parallel Input/Output and iterative methods.

Typical strong speed-up obtained with AVBP are illustrated on Fig. 1-(a). This flow solver shows an excellent scalability until 8,192 cores. Results are still good for 24,000 cores (efficiency is around 85%). The decrease of performance for a very large number of cores underlines a physical limit often encountered on massively parallel applications. Indeed for very large numbers of cores the ratio between computation time and communication time is directly proportional to the problem size (number of grid points and unknowns). For this specific illustration, the configuration tested with 16,000 cores corresponds to a calculation without enough cells by core or a too low computational workload for each core compared to the amount of exchanges needed to proceed with the CFD computation. It shows that a given task is limited in terms of scalability and no increase of performance is expected beyond a given number of cores.

2.2 The solid solver AVTP

The AVTP solver has been written based on the data structure of AVBP. The solver inherits from the mesh capability and the computational performances of AVBP. AVTP solves the time dependent energy conservation equation. A second order centered scheme is used for spacial discretization. Time integration is done either with an explicit or an implicit first order forward Euler scheme.

A typical strong speed-up obtained with AVTP is illustrated on Fig. 1-(b) for explicit
and implicit schemes. Due to the small number of operations per iteration compared to AVBP, the scaling of the explicit integration becomes poor at about 500 cores and then saturates. For the same Fourier number \((F = 0.5)\), the explicit scheme is limited due to stability reasons, the implicit scheme shows better scaling performances due to the fact that it needs more operations per iteration than the explicit scheme. The price for one iteration on one processor is thus higher when using the implicit scheme than the explicit one. Increasing the Fourier number, the number of operations per iteration increases (the method needs more sub-iterations to converge) and thus scaling appears to be better \((F = 10)\). On the other hand, using larger Fourier numbers increase the time step leading to a global decrease of the CPU time consumed to simulate a given physical time, which is why implicit schemes are preferred for this type of solver. Finally, the more the domain is partitioned by increasing the number of core, the more the conjugate gradient algorithm needs sub-iterations to converge which increases the number of operation per iteration and thus participate to the decrease of the efficiency of the solver. After 750 cores, the efficiency of the implicit solver at \(F = 10\) is less than 75%.

3 Presentation of the coupler

The OpenPALM software is a code coupler, i.e. a library of functionalities that facilitate the scheduling of existing components execution as well as the exchange of data between these components [12]. This is achieve in part via a collection of primitives that are called in the codes as well as with more complex mechanisms for application scheduling. OpenPALM aims at implementing a general tool allowing to easily integrate high performance computing applications in a flexible and evolutive way proposing a solution to the balance among performance, software reuse and numerical accuracy.

OpenPALM is mainly composed of 3 complementary components, (1) the PALM library\(^1\), (2) the CWIPI library\(^2\) and (3) the graphical interface PrePALM. As the application programming interface is available in Fortran and C/C++, OpenPALM can couple codes written in different languages. The CWIPI library is the part of OpenPALM that is mostly used in this study and is thus explained in more details in the following. CWIPI aims at providing a fully parallel communication layer for mesh based coupling between several parallel codes with MPI communications. Like most existing coupling libraries for multi-executables paradigm [15, 16, 17], CWIPI is a static coupler in the sense that all the components of the simulations are started at the beginning, exchange data during the run phase and finish together at the end. Coupling is made through 1D, 2D or 3D exchange zones that can be discretized in different ways in the coupled codes. The library takes into account all types of geometrical elements (polygon, polyhedral) with an unstructured description. CWIPI functionalities involve the construction of the communication graph between distributed geometric interfaces through geometrical localization, interpolation

\(^1\)Projet d’Assimilation par Logiciel Multiméthodes
\(^2\)Coupling With Interpolation Parallel Interface
on non coincident meshes, exchange of coupling fields for massively parallel applications as well as visualization file building.

3.1 High performance mesh based coupling

Code coupling is an appealing method to develop multiphysics applications. However if it is done incorrectly it can become a performance pitfall and render useless the efforts invested to optimize each individual code. There are at least two important aspects to take into account to manage efficient code coupling in a HPC context: (1) reducing the overhead of data transfer between the solvers and (2) maintaining a global processor idle time low, unless both codes have perfectly equal CPU per iteration times, the fastest code will have to wait the other. Having a good load balancing is the key to maintain a low idle time and thus reduce CPU waste. The first point requires the most attention and a direct point to point communication between each solver’s processors is proposed. Also non matching grids being used, a parallel interpolation method is required. The algorithm consists of two parts: the initialization or setup phase, i.e. where the communication routes and the interpolation coefficients are computed, and the run-time phase, or how inter code synchronization is actually executed. The first phase is done just once per coupled simulation except if the geometries are mobile.

For the description of these phases, lets consider that solvers A and B are linked by a coupling through their respective discretized interface $I_A$ and $I_B$.

3.1.1 Inter-code communication scheme (ICCS) determination

The communication routes construction consists in projecting the discretized interface $I_A$ on $I_B$ and vice-versa in order to prepare the communication phase. To maintain full scalability, coupling massively parallel applications has to remain a distributed process not only during the run-time part, but also the initialization part. Furthermore distributing the workload in the initialization improves the capacity of the coupled application to handle large simulations (which is a key for future applications).

Connecting the interfaces $I_A$ and $I_B$ means being able to perform geometrical searches from a computational domain into the other in order to locate the degrees of freedom of $I_A$ in $I_B$ and vice-versa. To keep the data distribution, the geometrical searches are performed in a parallel way by avoiding data centralization and sequential treatment. This objective faces a clear difficulty: in massively parallel CFD applications or heat transfer solvers the meshes are partitioned into sub-domains each processed by different processors. This partitioning also applies to the coupling interfaces. As the partitioning algorithm is usually not aware of the coupling process, the different distributions have no reason to match, leading to complex associations between interface processors of both solvers. To address these specific difficulties the CWIPI algorithm is composed of an optimized three levels location method (Algo. 1): the first level is the partition number of the mesh, the second one is the cell number in the selected partition, and the next one
is the mean values computed in the selected cell.

As the process is fully symmetric, let’s consider that code A is the source code (where the data is localized for the interpolation) and code B is the target code. The corresponding interfaces $I_A$ and $I_B$ are partitioned on processes, leading to $n_A$ sub-interfaces noted $I^n_A$ and $n_B$ $I^n_B$ with $n \in [1, n_A]$ and $m \in [1, n_B]$. It is worth noting that the number of sub-interfaces $n_A$ (respectively $n_B$) is lower or equal to the total number of total partition of the code A (respectively B). Only the processes which contains a sub-interface are involved in the projection algorithm.

The location algorithm is parametrized to adjust the size of the bounding box around the source process sub-interface $I^n_A$ for step #2 as well as around the source cells for step #4. This parameter takes the form of a tolerance: increasing it leads to helping locating points when geometries are not exactly matching. Note however that increasing tolerance results in an increase of the time requested by the location algorithm to converge.

Algorithm 1: Inter-code communication scheme (ICCS) determination algorithm

**Step0:**
- Each partition $n$ of the source code defines its discretized source sub-interface $I^n_A$ to the coupler (nodes coordinates & connectivity of the cells).
- Each partition $m$ of the target code defines its discretized target sub-interface $I^m_B$ to the coupler (nodes coordinates & connectivity of the cells)

**Step1:** Each process of the source code defines a surrounding box of its partition $I^n_A$.

**Step2:** Each process of the source code checks for geometrical intersections of its surrounding box of sub-partition $I^n_A$ with target nodes of the different target sub-interface $I^m_B$.

**return** Determination of a reduce number of target nodes per source process $n$

**return** Construction a first communication graph between source and target processes

**Step3:** Each process of the source code classifies the previous target nodes in an octree structure in order to optimize the next research step

**return** Octree structure containing the target nodes

**Step4:** Each source process defines a sub-box per mesh element of its sub-interface $I^n_A$.

**Step5:** Each source process checks the intersection between each source cell sub-box of $I^n_A$ and the target nodes classified in the octree.

**return** Determination of a limited number of candidate target nodes per source cell

**Step6:** For each target node, the source process identifies the closest element of the source sub-interface $I^n_A$ and defines the final communication graph

**return** Final communication graph from the source processes to the target ones
3.1.2 Communication phase

The communication phase consists in the interpolation of the fields and the exchange of the data between the solvers. The data can be stored either at the center of the cells for cell-centered solvers or at the nodes for cell-vertex solvers. Interpolation is done directly by the source solver via linear methods. Note that user can customize the interpolation with call-back definition. To ensure communication scalability the communication scheme between each solver is based on direct point to point communications between the processors which share a common interface following the communication graph setup in the previous phase. Each processor generally has several counter parts, which have to provide a portion of its data field.

Two communication schemes exist in the library: (1) synchronous and (2) asynchronous. In the synchronous mode, each process of code A that treats a sub-interface $I_A^i$ is involved in a loop of communications with processes of code B that share partially this sub-interface. The bounds of this loop are obtained thanks to the inter-code communication scheme determination phase and follow the natural order of the process numbering. The exchanges are based on the MPI_Sendrecv primitive so that they can be mono or bi-directional. This primitive is a blocking one implying that an exchange between process $n$ of code A with process $m$ of code B has to be finished before starting the exchange between $n$ and $m + 1$. This method is not well optimized when a very large number of cores is involved in the coupling. Concerning the asynchronous mode, the exchanges are based on loops around the primitives MPI_Isend for the sending and MPI_Irecv for reception. The completion monitoring of the exchanges is achieve thanks to loops around primitives MPI_Wait. In this mode, the communication times can overlap in a fully transparent way which is prone to be more performant than the synchronous mode. The other advantage of the method is its potential to overlap the communication times with other treatments in the code that don’t affect the exchanged fields.

4 Application to an industrial combustion chamber

The configuration of interest is a sector of an annular helicopter combustion chamber (Fig. 2), including the secondary air flow (A), the flame tube (B) and the high pressure distributor (C). The flame tube is fed with air and kerosene through an injector (D). The flame stabilizes in the primary zone (E). The thermal problem studied here by conjugate heat transfer consists in the determination of the temperature of combustor wall as well as of the stator.

The fluid domain is discretized with 3.8 million nodes and 21.4 million tetrahedra. To describe the flame with sufficient accuracy, the mesh is refined in the primary zone where the flame stabilizes. The solid domain is discretized with 3.8 million nodes and 18.2 million tetrahedra. The high number of solid cells is linked to a resolution constraint that requires at least 4 elements in the wall thickness. The conjunction of very thin walls with the use of tetrahedra leads to a large number of elements.
Computations are done on the CURIE supercomputer, owned by GENCI and operated at the TGCC by CEA. CURIE offers different fractions of x86-64 computing resources. Thin nodes among the 5040 B510 bullx nodes of the architectures are used in the present study. Each node is composed of 2 eight-core Intel processors Sandy Bridge EP (E5-2680) with 2.7 GHz, 64 GB and 1 local SSD disk. A total of 80,640 cores are availables on the machine. The coupling has been tested on 768 processors, ie 12,288 cores.

The strong speed-up obtained with AVBP on the present configuration is shown on Fig. 1-(b). The performance of the solver are very good till 2,048 cores. After, a significant loss of performance is observed. This is directly linked to two main reasons: (1) the MPI_allreduce that are not optimized in the version of Bullxmpi used for these tests and (2) the size of the fluid mesh which doesn’t contain enough degrees of freedom to reach good scaling properties up to 8,000 cores. The strong speed-up obtained with AVTP based on 8 cores on the present configuration is also shown on Fig. 1-(b). The curve is the one already discussed during the AVTP description. A good scalability is observed till 650 cores with the implicit scheme at Fourier $F = 10$ used for the coupled computations.

5 Parallel efficiency results

This section presents the strong scaling analysis of the coupling between AVBP and AVTP from 128 to 12,288 cores. In a first part, the simulations conditions are detailed. Some important features resulting from the application that directly impact the efficiency results are underlined. Finally, the performance of the coupled application are analyzed.

5.1 Simulations conditions

The physics and the numerics of the coupling methodology to reach a steady thermal state of the solid are detailed in [7, 11]. The main idea relies on a desynchronization of
the temporal evolutions in the fluid and solid with a high frequency data exchange to reduce the CPU costs while ensuring stability of conjugate heat transfer computations. It results that for the present case, both solvers run in parallel and exchange data at a fixed frequency corresponding to 20 time steps for the fluid code and 1 time step for the solid one.

Table 1 presents the repartition of cores between AVBP and AVTP for the 9 cases tested. Knowing the performances of the codes on the target machine, these repartitions aim to not slow down the fluid code which is the more CPU greedy in case of not perfect synchronization at meeting points (i.e., the cores of the solid solver wait for the fluid ones for data exchange). As a result, the performances of the coupling are compared to AVBP performances in the following of the paper. Note that the driver of OpenPALM runs on one core and that AVBP always runs on a whole processor count in order to avoid processor sharing between AVBP and AVTP.

AVBP treats around 0.48 million cells on its discretized coupling interface and AVTP about 1.7 million elements. The core distribution among the solvers as well as the number of cores on which these coupling interfaces are partitioned for the 9 cases are shown on Fig. 3. It is worth noting that 100% of the solid cores are involved in the coupling process for all cases. On the other hand, the proportion of AVBP processes involved in the coupling interface decreases almost exponentially when the number of cores increases. This behavior is linked to the ratio between the volume of the configuration and its surface which is higher in the case of the fluid solver. It is important to note that mesh partitioning is managed independently in each solver resulting in non conformal patterns of the partitioned discretized surfaces.

It is thus of interest to analyze the evolution of the distribution of the number of mesh elements in the sub-interfaces as the number of cores increases. Indeed, a reduced amount of AVBP cores have more task to perform than the others in order to perform the exchanges with AVTP. Moreover, the way this additional load is distributed and its evolution when the number of processing cores increases is particularly important to ensure the scaling of the coupled application. To analyze these distributions, Fig. 4 and 5 show the probability density functions (PDF) of the number of cores treating $n_c$ cells (these PDFs are constructed based only on the cores with coupling cells). Abscissa scales are logarithmic in order to facilitate the interpretation. Concerning the fluid solver AVBP,
the PDFs at low number of cores are rather spread out with distinct peaks, reflecting an inhomogeneous partitioning of the coupled discretized interface among the computing cores. Increasing the number of cores narrows the range of the PDF around the small number of cells. A significant peak emerges indicating a homogenization of the distribution of coupling cells on the cores. Nevertheless, a peak at very low number of cells exists highlighting the participation of cores in the process of inter-code exchanges for a small number of information. From case #3 to 9, the peak has the same order of magnitude as the main peak of the distribution.

As far as AVTP is concerned, case #1 exhibits a perfectly homogenous distribution on the 3 cores. Then all the distributions show a multimodal profile with a marked peak in the range of small numbers of coupling cells per core.

The analyzes of this section underline phenomenological issues resulting from the coupler as well as the CHT application. Some of these results consist of potential weaknesses for the performance results: imbalance between the number of cores allocated to the
solvers, all the cores of the AVBP solver don’t participate to the coupling creating a difference of computing load, imbalance repartition of the sub-interface among the coupling cores on each solver, complex communication scheme with a lot of connection between processes of the solvers, and finally a combination of these features. Knowing these features, the next section analyzes the scalability of the application.

### 5.2 Performance of the coupled application

The time needed to construct the ICCS is presented on Fig. 6-(a). Except for case #1, this time is almost constant when the number of cores increases. For case #1, the ratio of cores involved in the coupling for AVBP over the one involved in the coupling for AVTP is the most important: this imbalance of cores that have to communicate together leads to a very poor efficiency of the algorithm. Increasing the number of cores for AVTP from case #1 to case #2 leads to a drastic reduction of the consumed time. From case #2 to 4, the number of cores allocated to the AVTP coupling interface is fixed while the number of AVBP cores increases. It results a slight increase of the time spent to construct the communication scheme. Then, the ratio of coupling cores between the solvers is almost constant resulting in a slight decrease of the time consumed by the ICCS algorithm until 8,000 cores. From this analyzes, one can note that the scaling of the ICCS algorithm largely depends on the number of processes involved in the coupling on each side of the coupled problem. Increasing the number of cores of just one code is not sufficient to ensure good scaling of this step.

Figure 6-(b) presents the evolution of the time taken by an AVBP iteration compared to the time of an exchange between the solvers as a function of the total number of cores of the coupled system. The exchanges are made in the synchronous mode. As previously mentioned, the fluid solver exhibits a rather good scaling until 4,000 cores. Above 8,000 cores, the time of an iteration reaches a plateau. The time needed for the data exchange is almost constant on the whole range of cores. As for the first phase, the time of the communication step exhibits an important decrease from case #1 to 2. This reveals that the balance of processes that are coupled continues to play an important part in this phase. Then, the communication time slightly increases as the number of cores grows. Until several thousand of cores, the time requested by an exchange is several order of
magnitude lower than the time of an AVBP iteration. Then both times are of the same order. As the coupling is done every 20 iterations of the fluid solver AVBP, the global scaling of the application is conserved as shown on Fig. 7. The communication phase is thus fully transparent in term of restitution time for the user.

Figure 6: (a) Time taken by the inter-code communication scheme determination algorithm, (b) comparison of the time for one iteration of AVBP and the time to do one exchange as a function of the total number of cores of the coupling.

Figure 7: Comparison of the strong speed-up curve obtained for AVBP / AVTP coupling with the AVBP standalone one.

6 Conclusion

The different choices made for the resolution of conjugate heat transfer problems on complex geometries put a lot of pressure on the tool used for coupling. As in several
communities, developers are now faced to the challenge of running coupled applications with highly loaded codes on massively parallel machines where the solvers exchange a lot of data at a high frequency. The strategy investigated in this work to address these issues relies on recent developments made in a generic parallel coupler. The characteristics of the coupled tool including the fluid and solid solvers as well as the coupler are detailed. Then, the tool is applied to an industrial combustion chamber. Performance tests are carried out until 12,288 cores on the CURIE supercomputer (TGCC / CEA) and relevant informations that influence the coupling scalability are detailed. The coupler exhibits a very good behavior up to 12,288 cores implying that the use of HPC can drastically reduce the restitution time of coupled applications for industrial design with high fidelity solvers. Analyzes of the scaling response underline the impact of imbalanced repartition of cores among the codes, imbalance repartition of the sub-interface among the coupling cores on each solver, as well as the complex communication scheme with a lot of connections between processes of the solvers on coupling overhead. These points are independent from the coupler and can be addressed by incorporating the knowledge of the coupling in the preprocessing step of the solvers (constraint and co-partitioning). Moreover, recent tests on asynchronous communication show an important improvement of the scalability of the coupler indicating development paths for the future.

The work was supported by the Fondation de Recherche en Aéronautique et Espace in the project STRASS. The TGCC is gratefully acknowledged for providing access of the CURIE machine during the Grand Challenge. Turbomeca is acknowledged for supporting code developments and permission for publishing results.

REFERENCES


INFLUENCE OF THE IONIC STRENGTH ON THE DEPOSIT PHENOMENON AND TRANSPORT DYNAMIC OF MICROPARTICLES THROUGH SATURATED POROUS MEDIUM

Z. MESTICOU, M. KACEM AND PH. DUBUJET

Université de Lyon, Ecole Nationale d'Ingénieurs de Saint-Etienne, Laboratoire de Tribologie et Dynamique des Systèmes UMR 5513 .58 rue Jean Parot, Saint Etienne 42023 France
E-mail: zyed.mesticou@enise.fr, mariem.kacem@enise.fr, philippe.dubujet@enise.fr

Key words: Saturated porous media, micro particles, ionic strength, deposition phenomenon, dynamic transport model.

Abstract. In this paper, the influence of ionic strength on the dynamics transport of silt micro particles through saturated sand texture is studied in the presence of repulsive interactions. The deposition phenomenon is investigated through column experimental trials. Four ionic strengths are applied by adjustment of suspension salinity (0, 5.13, 10.26 and 13.68 mM). Through this experimental study, the ionic strength influence on deposition phenomenon is shown at the micro particle scale. Ionic strength variation is the primal parameter which predicts attachment and detachment particles at constant flow. These experiments are simulated and reproduced through a numerical model based on an original deposit kinetic which is proposed in this study. This model is the coupling of two multiphasic problems describing conservative salt and micro particles transport. The proposed kinetic formulations are founded on experimental tests constitutions with respect to literature trends. They are based on functional relationships between model parameters and the suspension ionic strength. In addition, the transient blocking phenomena is taken into account through a retention function Langmuir type. The suggested model shows a good match to reproduce the experimental description of the suspended particles transport under the influence of ionic strength variation. It permits to predict deposition phenomenon.

1 INTRODUCTION

In natural subsurface and flow suspension, the micro particles can present a grateful factor to contaminant porous saturated media. Indeed, they can act as a vehicle of organic, inorganic contaminant and heavy metals or instead became a barrier to the migration of these pollutants in the case of deposition and porous medium clogging.

Chemical characteristics of the suspension flow are among the major factors involved in the process of suspended particles retention or release. Indeed, both porous texture and suspended particles are negatively charged under natural conditions. Therefore, electrostatics interactions significantly affect the physicochemical retention mechanisms. The authors ([1], [2], [3] [4], [5]) have confirmed through the investigation of experimental model system and numerical studies that increasing ionic strength enhances particles deposition rates under
constant hydrodynamic factors at typical pH values. This behavior leads to favorable attachment conditions. However, decreases in suspension ionic strength generate an increase of the electrostatic repulsion interactions and the retained particles mobilization ([6], [7]).

Nevertheless, colloidal scale transport is the major topic of these studies. The physicochemical mechanisms are dominant in the depth filtration with colloidal particles which the size is less than 1 μm [1]. Moreover, the hydrodynamic conditions influence is insignificant at this scale since the surface forces are dominant ([8], [9], [10]). In contrast, both mechanical and physicochemical mechanisms are present for particles transport with diameter between 1 and 30 μm ([11], [12]). In addition, the ionic strength influence on the micro particles suspension behavior depends also on hydrodynamic. Indeed, changes in hydro-dynamical conditions affect directly hydrodynamic forces and extension of the boundary layer diffusion mainly for micrometer particles ([13], [14]). Thus, it seems interesting to investigate the ionic strength influence on micro particles retention.

In this paper, the ionic strength influence on silt micro particles transport across saturated sand bed is studied. The micrometer silica diameter ensures the presence of both physicochemical and mechanical retention even as the size ratio of suspended particle to media grain is greater than 7.1 10^{-3} [15]. To characterize the deposition phenomenon, experimental trials are conducted using a laboratory column at soil natural conditions. In these conditions, both mobile and immobile materials present negative charge surfaces at neutral pH [16]. In the first part, the followed experimental procedure is described and the used materials are exposed. In this study, particle loading of sand texture are performed at different ionic strength. In the second part, a multi-phasic model is proposed to describe micro particles transport through continuous approach. This model is the coupling of two transport phenomena of micro particles and salt. It simulates experimental tests through an original deposit kinetics proposed herein with consideration of the ionic strength variation. The applied kinetic is drew upon experiments. The model parameters are determined by reverse optimization in accord to experimental results and sensitivity analysis.

2 MATERIALS AND METHODS

2.1 Porous media and suspended particles

The porous media is composed of fine Hostun sand grain HN 31 (Sibelco, France). The sand grains have a mean diameter of 0.375 mm. They are characterized with a uniformity coefficient $C_u=1.5$ and the curvature coefficient $C_c = 1.1$. These intrinsic properties reveal a uniform distribution size and well graded porous medium. The bed sand present an hydraulic conductivity of $9.8 \times 10^{-4}$ m.s^{-1} and a bulk density of 2675 kg.m^{-3}. The porosity $\Phi$ of the grain medium is equal to $0.43\pm0.05$ for each experimental trial. It is determined with an error of 0.005. The sand was thoroughly cleaned prior to use to remove organic matter and other impurities, leaving only a pure SiO$_2$ grain surface. The cleaning procedure consisted on the emersion of the sand in deionized water for long duration to have stabilized chemical proprieties and then it is put in oven for 24 h at 100°C.

Monodisperse suspensions of silica micro particles are used as model particles in the deposition column experiments. The particles have a mean diameter $d_p$ between 1.3 μm and 22 μm ($d_{50}=11$ μm; $C_u=10.8$; $C_c = 0.948$). The microscopic description shows a crystalline
form of the silica particles with no uniform sizes. Silica particles are composed mainly of quartz SiO₂ (> 98 %) and a small fraction of inorganic matter (Fe₂O₃, Al₂O₃, K₂O, CaO, TiO₂).

The suspended particle sizes and the ratio size of particle and sand grain suggest the presence of both retention physicochemical and mechanical (straining, wedging, interception) mechanisms ([11], [15]).

2.2 Transport particles experiments

Particle transport experiments are conducted in a plexiglas column packed with clean sand grains at dry conditions and vibrated regularly to minimize layering or air entrapment. The column has an inner diameter of 7 cm and a height of 60 cm. The column dimensions are justified as they allow a homogenous flow along the sand media and reducing preferential pathways. They ensure an important pore space that approximates particles transport under natural conditions.

A peristaltic pump (Verder AU.UPC.EZ) is connected to the sand bed. It ensures continuous injection of electrolytic solution or particles suspension through the vertical oriented column (Figure 1). The effluent properties are analyzed continuously using a turbidimeter (Turbimax CU21) and a conductimeter (Liquisys M CLM223). These measures describe respectively the particles and salt concentration of the effluent.
Four differential manometers are used to follow the evolution of the pressure at different points through experimental test in order to identify possible clogging phenomenon. The porous media has been purged with CO₂ to ensure a complete saturation. Then, it was initially equilibrated by injection of deionized water. Then, particle free background electrolytic solution is injected through the sand texture for 6 hours at the specific ionic strength used for each experimental test. This procedure eliminates traces and also impurities that can disrupt turbidity measure.

Transported particles tests are performed with a particles concentration $c_{ao}$ of 1 g.L⁻¹ at constant darcyan velocity ($u_d=5.19 \times 10^{-2}$ cm.s⁻¹). A particles suspension of 3.5 $V_p$ (pore volume) is injected for different ionic strength $I$ (0, 5.13, 10.26 and 13.68 mM) at neutral and constant pH (6-7). The sodium chloride (NaCl) is used to adjust the ionic strength. For the pore volume injected, there is not permeability reduction through sand texture since hydraulic gradient still constant according to realized pressure measures. The particles suspension is mixed continuously to ensure the mixture homogeneity and to have isotropic properties. All the experiments are run in duplicate to verify the criterion of repeatability.

4 MODELING AND VALIDATION

4.1 The transport particles model

A continuous approach is adopted to describe the micro particles transport through the sand texture ([17], [18]). The medium is considered stable, isotropic, isothermal and homogeneously saturated. It is constituted from two phases: A solid phase (The matrix ($g$) and the deposited particles ($p$)) and a fluid phase (The interstitial fluid ($w$), the solute ($s$) and the suspended particulates ($a$)). The fluid phase is considered homogeneous and incompressible. The sodium chloride salt ($s$) is used to adjust the ionic strength $I$. Therefore, it is taken into account as a fluid phase component. The salt precipitation absence hypothesis is assumed.

The soil representative volume components are characterized by an absolute velocity $v^\alpha$ ($\alpha \in [g, p, w, s, a]$). They have an intrinsic density $\rho_\alpha$ relative to the constituent material ($\alpha$) and a bulk density $\rho^\alpha$ which is defined as the ratio of the material mass $m_\alpha$ by the total volume $V$. The suspended particles and porous media grain have equal intrinsic density as assumption. In addition, the suspended particles and the salt have equal effective velocity as the intestinal fluid ($v^\alpha = v^s = v^w$) [19].

The transport model is the coupling of two multi-phasic problems: Particle and solute transport. It is based on the synthesis of the mass conservation equations which are applied to the different medium constituents ([20], [5]). Laminar and unidirectional flow assumptions are considered. The model is presented through the following equations,

$$\phi \frac{\partial c_s}{\partial t} - \phi \frac{\partial}{\partial z} \left( D_s \frac{\partial c_s}{\partial z} + u_d \frac{\partial c_s}{\partial z} \right) = 0$$

$$\phi \frac{\partial c_a}{\partial t} - \phi \frac{\partial}{\partial z} \left( D_a \frac{\partial c_a}{\partial z} + u_d \frac{\partial c_a}{\partial z} \right) = -R_a$$

$$\rho_a (1 - \phi) \frac{\partial c_p}{\partial t} = R_a$$

1171
is the conservative salt concentration (g.L\(^{-1}\)), \(c_s\) is the suspension particle concentration (g.L\(^{-1}\)), \(c_a\) is the deposed particle concentration, \(R_a\) deposit rate (g.L\(^{-1}.s^{-1}\)), \(\rho_g\) is the intrinsic density of porous medium (Kg.m\(^{-3}\)), \(u_d = \phi v^w\) is the darcyan velocity (m.s\(^{-1}\)), \(\phi\) is the porosity (m\(^3\).m\(^{-3}\)), \(D_s = \alpha_s u_d\) and \(D_a = \alpha_a u_d\) are the hydrodynamic dispersion coefficient respectively of salt and particles (m\(^2\).s\(^{-1}\)), \(\alpha_s\) and \(\alpha_a\) (m) are the dispersivity respectively of salt and particles.

Equation (1) represents the mass conservation and dispersion equation for salt. Equation (2) represents the mass conservation and dispersion equation for suspended particles. Then, equation (3) represents the mass conservation equation for the deposed particles. It characterizes micro particles exchange between solid and fluid phases. In the absence of clogging phenomena, the total porosity and darcyan velocity remain constant since the permeability reduction effects do not occur.

The deposition rate \(R_a\) can be expressed as a first order kinetic (4). It is expressed as a function of the deposit kinetics \(k_a\) (s\(^{-1}\)) and the retention function \(\psi\) (\(\psi(\cdots)([21], [22])\)).

\[
R_a = \phi \psi c k_a(c_s)
\]  

(4)

The deposition kinetic \(k_a\) describes the velocity at which the suspended particles are transported from the fluid phase to the solid phase. In accord to the literature trend, suspended particles retention increases with the ionic strength. Thus, the deposition kinetics \(k_a\) may be expressed as a function of the suspension salinity. It is based on experimental results and drew upon literature considerations ([1], [23]).

\[
k_a(c_s) = k_{a0} \left(1 + \frac{\kappa}{1 + \frac{CDC}{c_s}}\right)
\]  

(5)

\(\kappa\) is empirical model parameters to be adjusted with the experimental curves, \(k_{a0}\) is the deposition kinetic coefficient (s\(^{-1}\)) and CDC is the critical deposition concentration (g.L\(^{-1}\)) below which the retention phenomenon dominates. The deposition kinetic coefficient \(k_{a0}\) characterizes the suspended particles retention under favorable deposition conditions.

The retention function \(\psi\) describes the dynamic deposition phenomenon. Indeed, the deposition rate decreases progressively with the particles retention on the grains surfaces and the deposition sites saturation ([1], [24]). The retention function \(\psi\) (6) is expressed as Langmuir formulation as a function of deposed particles concentration. It involves the assumption that the deposition rate \(R_a\) depends on the particles concentration in the fluid phase even on the deposited particles fraction.

\[
\psi = 1 - \frac{c_p}{c_{p,max}}
\]  

(6)

\(c_{p,max}\) is the maximum concentration of deposed particles per mass unit (g.g\(^{-1}\)) at given chemical condition. \(c_{p,max}\) (7) is expected to decrease as ionic strength increase [24].

\[
c_{p,max} = \frac{\alpha_1}{\rho_g} \exp(c_s)
\]  

(7)

\((\alpha_1/\rho_g)\) corresponds to the maximum concentration of deposed particles per mass unit under unfavorable deposition conditions with the dominance of repulsive interactions.

Model is implemented using a finite element schema that allows the resolution of equations system (1-3) and determines the evolution of resituated and retained particle
concentration. The applied 1D mesh includes 240 quadratic elements. It ensures stable numerical solution.

The boundary conditions are presented in Figure 2. The porous medium is initially considered blank, devoid of particles.

![Figure 2: Model 1D of micro particle transport and boundary conditions](image)

The model parameters identification is conducted using the successive approximations principle and a parameters sensitivity analysis. It required executing several calculations before determining the parameters in accord to experimental results. The experiments previously presented are used to adjust the model parameters. Table 1 exposes the identified parameters and their values. The micro particles dispersivity $\alpha_a$ is estimated equal to that of the solute $\alpha_s$ as simplifying assumption.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\alpha_s$ (cm)</th>
<th>$\alpha_c$ (cm)</th>
<th>$k_{a0}$ ($s^{-1}$)</th>
<th>$\kappa$</th>
<th>CDC (g.L$^{-1}$)</th>
<th>$a_1$ (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>$10^{-2}$</td>
<td>$10^{-2}$</td>
<td>0.488</td>
<td>1.7</td>
<td>3</td>
<td>10.5</td>
</tr>
</tbody>
</table>

4.2 Sensitivity analysis

A parameter sensitivity analysis is made to analyze qualitatively the model behavior in response to a ± 20% variation of the identified parameters change (Figure 3). The reference simulated curve is associated to experiment with ionic strength of 13.68mM ($c_{s0}=0.8$ g L$^{-1}$).

The deposition kinetic coefficient $k_{a0}$ is the most sensitive model parameter (Figure 3.a). The $k_{a0}$ coefficient variation (± 20%) shows a constant and regular variation of the restitution rate ($c_a/c_{a0}$). However, the critical deposition concentration $CDC$ and the model parameter $\kappa$ have a moderate influence in comparison to $k_{a0}$ (Figure 3.a-b). These parameters increase the model results accuracy.
The empirical parameter $\alpha_1$ allows the increase of the model sensitivity to the salinity variations (Figure 3.c). Its influence increases over the time. Indeed, the parameter $\alpha_1$ reflects the dependence of the maximum retained particles concentration $c_{pmax}$ on the suspension salinity. It ensures the description of the particulate restitution rate increase over the time which results in the progressive retention sites saturation.

4.3 Results and discussion

The particles release curves are shown in figure 4. These curves represent a typical resituated particles rate variation over the time. The experimental and numerical data are displayed as normalized resituated particles concentration ($c_a/c_{a0}$) at the outlet of the column as a function of the time.
Z. Mesticou, M. Kacem and Ph. Dubujet.

Figure 4: Evolution of the restituted suspended particles rate \( \frac{c_i}{c_{a0}} \) as a function of time-monotonous tests

\[ \frac{c_i}{c_{a0}} = 1 \text{g L}^{-1}; \quad u_i = 5.19 \times 10^{-2} \text{ cm s}^{-1} \]

\( c_{a0} \) are the silica particles injected concentration and at the inlet of the column and \( I \) is the suspension ionic strength.

The simulated experiments allow the parameters identification associated to the deposition rate \( R_d \) (Table 2). These curves show good agreement with the experimental results for the different ionic strength. In particular, the adopted deposition kinetic reproduces correctly the increase in the particles retention rate with ionic strength augmentation.

Indeed, figure 4 shows an increase in the deposition rate with ionic strength under constant flow. This result may be clarified by the classical Derjaguin-Landau-Verwey-Overbeek (DLVO) theory ([25], [26]). It describes the surface interaction between particles and porous medium grains: The system stability results in the sum of the Van der Waals attractive forces and the repulsive electrostatic forces due to the double layer. The salinity augmentation results in a reduction in double layer thickness with ionic strength increase. It is shown that the influence of electrostatic interactions which are characterized by the Debye length is inversely proportional to the ionic strength [27]. Thus, double layer thickness reduction results in repulsive forces decrease and then increasing particles deposition on the grain surface.

In addition, the model can well reproduce the quasi-linearly increasing level of particulate restitution over the time (Figure 4). This growth can be assign to the gradual decrease in physicochemical and mechanical retention sites. Indeed, the suspended particles deposit and occupy the available retention sites over time. Then, restitution particles rate is increasing at the column outlet as the deposition sites is limited. Thus, complete deposition sites saturation is possible so that restitution rate may be equal to one at long-term. But, the clogging phenomenon should eventually disrupt this trend.

Experiments have been carried out with colloidal particles in a similar framework by other authors ([24], [22], [28], [5]). The colloids present similar behavior likewise silt particles. Nevertheless, these tests show in general a rapid increase in the restitution rate \( \frac{c_i}{c_{a0}} = 1 \) at low even high salinity. However, in accord to figure 4, obtaining an increasing level may reflects with no doubt the fact that the tests are carried out with a long column in this study. The column length implies a substantial amount of material and so more retention sites. Then saturation of all column deposition sites requires a long duration not reached in these
5 CONCLUSIONS

This study is devoted to apprehend the suspended micro particles deposition in saturated porous media with the presence of both physicochemical and mechanical retention mechanisms. The performed experiments investigate the influence of ionic strength on the dynamics transport of silica micro particles through sand texture. In accord to performed experiments, high ionic strength promotes retention of suspended particles and leads to favorable deposition phenomena.

The experimental tests are simulated through a numerical model based on a first order deposition kinetic. This model is the coupling of two multiphasic problems. It describes the conservative salt and micro particles transport. The deposition kinetics is founded on functional relationships between model parameters and the suspension ionic strength. The proposed formulations are based on experimental results. The model shows a good match to reproduce the experimental description of the suspended particles transport through saturated medium under the ionic strength disturbances. It presents a useful tool to predict deposition phenomenon with the absence of porous media permeability reduction.

REFERENCES

[9]. Cerda, C.M., “Mobilization of kaolinite fines in porous media”, Colloids and Surfaces,


NEW COUPLED THERMOELECTRIC LINK
FINITE ELEMENT FOR FGM MATERIALS

JURAJ PAULECH*, JUSTÍN MURÍN*, VLADIMÍR KUTIŠ*, JURAJ HRABOVSKÝ*

* Department of Applied Mechanics and Mechatronics of IPAE, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, Ilkovičova 3, 812 19 Bratislava, Slovakia
e-mail: juraj.paulech@stuba.sk, web page: http://aladin.elf.stuba.sk/Katedry/KMECH/

Key words: FGM Link Finite Element, Thermoelectric Analysis, ANSYS.

Summary. The paper deals with derivation process of new FEM equations for steady thermoelectric two-way coupled analysis of link conductor made of Functionally Graded Material (FGM). One example of coupled analysis will be introduced to demonstrate accuracy and effectiveness of our new approach in computer modelling of such systems.

1 INTRODUCTION

Nowadays, new materials are necessary for sophisticated structures like MEMS systems, advanced electronic devices, etc. Computer modelling of such complex systems, like structures with spatial variation of material properties (e.g. FGM) are, using commercial FEM code with classic elements, needs remarkable effort during preparation phase and sufficient computer equipment for solution phase because of necessity the numbers of elements and material models.

Finite elements for electric-thermal analyses of FGM materials considering Joule heat have been developed in [1]. This paper deals with derivation of new link finite element for two-way coupled static thermoelectric analyses considering Joule heat and also thermoelectric effects like Seebeck, Peltier and Thomson effects. These effects describe direct conversion of thermal energy into electric energy (Seebeck effect), conversion of electric energy into the temperature difference within the system (Peltier effect) and heat exchange on a current-carrying conductor with applied temperature gradient (Thomson effect).

Let us consider straight link conductor, the conductor is a slender construction. Let the conductor is made of mixture of two component materials – matrix (index m) and fibre (index f). Let the change of material properties is axial-symmetric so we can then consider a longitudinal and radial change of the material properties (2D change). Let the
thermal conductivities $\lambda_m$ and $\lambda_f$, and electric conductivities $\sigma_m$ and $\sigma_f$ are known values for matrix and fibre, respectively. Then, according to the homogenization process described in [1] we can calculate homogenized thermal and electric conductivities for whole conductor. These homogenized material properties need to be polynomials.

Homogenization process supposes that volume fractions of matrix and fibre are known and have also polynomial form, see Figure 1. Then, homogenization includes two steps:

- In the first step the real link is transformed to a multilayered link, real link is divided into chosen number of layers in radial direction ($N = 11$ for the case shown in Figure 1). The extended mixture rule [3] will be used for calculation of material properties of the layers. Then, each layer will have constant volume fraction and material properties of the constituents through its height, see Figure 2. Polynomial variation of these parameters will appear in the longitudinal direction of the layers.

- In the next step, these layers will be used for calculation of homogenized material properties for whole link (see Figure 2) according to laminate theory described in [4]. This homogenized material properties will interpret material changes in the whole FGM link and they will have polynomial form with just one independent variable ($x$).

![Figure 1: FGM conductor (left) and fibre volume fraction in the conductor – division into layers (right)](image1)

![Figure 2: Graphic form of the homogenization process](image2)
For thermoelectric coupling that results in Seebeck, Peltier and Thomson effects, Seebeck coefficient $\alpha(x, r)$ of the whole conductor needs to be known. This material property cannot be calculated according to the homogenization process based on extended mixture rule and laminate theory, because this material property is not given according to volume fractions of individual material components (expect the Joule heat, thermoelectric effects are significant especially for semiconductor materials and the behaviour of the semiconductors is given according to atomic structure, not according to volume fraction of individual admixtures). Moreover, the value of Seebeck coefficient can also be negative number. The determination of final Seebeck coefficient according to material properties of individual components is beyond the scope of this article. For our model case the final Seebeck coefficient will be chosen as a polynomial function $\alpha(x)$ for longitudinal direction of the conductor.

3 MATHEMATICAL BACKGROUND FOR TWO-WAY COUPLED THERMOELECTRIC ANALYSIS

Thermoelectric and electric-thermal effects like Joule heat, Seebeck, Peltier and Thomson effects are described by set of two thermoelectric constitutive equations (static analysis) [5]:

\[
\begin{align*}
\mathbf{q} &= [\Pi] \cdot \mathbf{J} - [\lambda] \cdot \nabla T \\
\mathbf{J} &= [\sigma] \cdot (\mathbf{E} - [\alpha] \cdot \nabla T)
\end{align*}
\]  

where $\mathbf{q}$ [Wm$^{-2}$] is heat flux vector, $\mathbf{J}$ [Am$^{-2}$] is electric current density vector, $[\Pi]$ [V] is Peltier coefficient matrix, $[\lambda]$ [Wm$^{-1}$K$^{-1}$] is thermal conductivity matrix, $T$ [K] is absolute temperature, $\mathbf{E}$ [Vm$^{-1}$] is electric field intensity vector, $[\sigma]$ [Sm$^{-1}$] is electric conductivity matrix and $[\alpha]$ [VK$^{-1}$] is Seebeck coefficient matrix.

These constitutive equations are coupled by set of governing equations for static thermal and electric fields:

\[
\begin{align*}
\nabla \cdot \mathbf{q} &= P \\
\nabla \cdot \mathbf{J} &= 0
\end{align*}
\]  

where $P$ [Wm$^{-3}$] is heat generation per volume unit.

In general, we can write for electric field intensity, Peltier coefficient and heat generation:

\[
\begin{align*}
\mathbf{E} &= -\nabla \varphi \\
[\Pi] &= T[\alpha] \\
P &= P_J + P_{aux} \\
P_{aux} &= [\sigma]^{-1} \mathbf{I}
\end{align*}
\]  

where $\varphi$ [V] is electric potential, $P_J$ [Wm$^{-3}$] is Joule heat per volume unit and $P_{aux}$ [Wm$^{-3}$] is auxiliary heat generation per volume unit (equals to zero for our case).

Applying (1) and (3) into (2) we can write for 1D system (longitudinal direction $x$):

\[
\begin{align*}
\frac{d}{dx} \left[ T(x) \alpha(x)f(x) \right] - \frac{d}{dx} \left[ \lambda(x) \frac{dT(x)}{dx} \right] &= \frac{j^2(x)}{\sigma(x)} + P_{aux} \\
\frac{d}{dx} \left[ \sigma(x) \frac{d\varphi(x)}{dx} \right] + \frac{d}{dx} \left[ \sigma(x) \alpha(x) \frac{dT(x)}{dx} \right] &= 0
\end{align*}
\]  

1181
4 DERIVATION OF NEW FGM EQUATIONS FOR FGM LINK CONDUCTOR

We can use the method for solving 1D differential equation with non-constant coefficients and with right-hand side described in [2] for our system of equations (4). But all non-constant coefficients on the left-hand side and right-hand side itself have to be in polynomial form (it is the condition of used approach). So general formulation of one-dimensional differential equation suitable for the method has the form:

\[ \sum_{u=0}^{m} \eta_u(x) y^{(u)}(x) = \sum_{j=0}^{g} \varepsilon_j a_j(x) \]  

where:

- \( m \) – degree of the differential equation
- \( y(x) \) – unknown function of independent variable \( x \)
- \( y^{(u)}(x) \) – \( u \)th derivation of the unknown function
- \( \eta_u(x) \) – polynomial variable coefficient for \( u \)th derivation on the left-hand side of the differential equation
- \( g \) – degree of a polynomial on the right-hand side of the differential equation
- \( \varepsilon_j \) – constant coefficient for \( j \)th power of the right-hand side polynomial
- \( a_j(x) = \frac{x^j}{j!} \) – auxiliary function for the right-hand side polynomial formulation

at which \( x \in (0; L) \), where \( L \) is the length of considered interval of unknown solution.

Considering function \( a_j(x) \) for formulation the polynomials, we can write general rules for derivation and integration of such function:

\[ a_j'(x) = a_{j+1}(x) \]

\[ \int_0^x a_j(x) \, dx = a_{j+1}(x) \]  

According to [2] the solution of the differential equation (5) has the form:

\[ y(x) = \sum_{i=0}^{m-1} y_0^{(i)}(x) + \sum_{j=0}^{g} \varepsilon_j b_{j+m}(x) \]  

where:

- \( y_0^{(i)} \) – \( i \)th derivation of the function \( y \) in \( x = 0 \), thus \( y_0^{(i)} = y^{(i)}(x) \big|_{x=0} \)
- \( c_i(x) \) – function for uniform solution of the differential equation
- \( b_{j+m}(x) \) – function for particular solution of the differential equation

And derivation of the solution (7) is in the form:

\[ y^{(u)}(x) = \sum_{i=0}^{m-1} y_0^{(i)} c_i^{(u)}(x) + \sum_{j=0}^{g} \varepsilon_j b_{j+m}^{(u)}(x) \quad u = \{0, m - 1\} \]
The solution of differential equation (7) lies in determining the functions generally labelled $c(x)^1$ and $b(x)$ that appear in the solution.

The result of calculation of the differential equation with variable coefficients and the right-hand side is the solution according to equation (7). It should be noted that this is a solution for selected point $x$ of the considered interval of independent variable, so the program is designed for calculation of values $c_i(x)$ in given point $x$ for $i = \{0, m - 1\}$ where $m$ is the degree of the differential equation. Then the values $b_{j+m}(x)$ for selected point $x$ are also calculated for $j = \{0, g\}$ where $g$ is degree of the right-hand side polynomial of the differential equation. It means that functions $c(x)$ and $b(x)$ cannot be calculated analytically but only at discrete points $x$, where $x$ is from interval $(0; L)$.

The program algorithm, also described in [2], efficiently calculates matrixes of discrete values $c_i^{(u)}(x)$ and $b_j^{(u)}(x)$ for user defined point $x$ where $u = \{0, m - 1\}$ represents derivation for calculations according to (8).

It should be noted that the polynomial function of the right-hand side of the differential equation itself does not enter into the calculation. For the calculation of $c(x)$ and $b(x)$ it is only necessary to know the degree of the right-hand side polynomial. The values of the coefficients $\varepsilon_j$ of the right-hand side only enter into the final equation – solution (7) or derivation (8), respectively.

Programming code offers the possibility of automatic uniform interval division calculation, and the values of $c(x)$ and $b(x)$ for interval dividing points are included in the output. So during single run of the program we can get all the necessary values to solve the differential equation also for selected internal points in the interval $x \in (0; L)$, in our case for conductor inner region.

This program for solving differential equations will be now used in process of deriving new FEM equations for two-way coupled thermoelectric analysis in FGM link conductor mentioned above. Figure 3 shows geometry and physical quantities used during the derivation process.

Figure 3: Two-nodal conductor for thermoelectric analysis

---

1 denoting is symbolical – general; correctly it should be mentioned that it is a set of values $c_i(x)$ within the range $i = \{0, m - 1\}$; it is also similar for $b_j(x)$ with the range $j = \{0, m - 1\}$
Let us write the governing function (4) for heat flux in the form suitable for calculation according [2]. Comparing the general form of the differential equation (5) with governing function (4) in expanded form:

\[
\sum_{u=0}^{m} \eta_u(x) y^{(u)}(x) = \sum_{j=0}^{g} \varepsilon_j a_j(x)
\]

\[-\lambda(x) T''(x) + [\alpha(x)f(x) - \lambda'(x)] T'(x) + [\alpha'(x)f(x) + \alpha(x)f'(x)] T(x) = \frac{j^2(x)}{\sigma(x)} + P_{aux}
\]

we get:

\[
m = 2 \quad \text{degree of the differential equation}
\]

\[
y(x) \equiv T(x) \quad \text{unknown function of independent variable } x \text{ is function of temperature}
\]

\[
y^{(u)}(x) \equiv T^{(u)}(x) \quad \text{ } ^{u}\text{th derivative of unknown temperature}
\]

\[
\eta_0(x) = \alpha'(x)f(x) + \alpha(x)f'(x) \quad \text{0th derivation of the temperature on the left-hand side of the differential equation}
\]

\[
\eta_1(x) = \alpha(x)f(x) - \lambda'(x) \quad \text{non-constant coefficient of the 1st derivation of the temperature}
\]

\[
\eta_2(x) = -\lambda(x) \quad \text{non-constant coefficient of the 2nd derivation}
\]

\[
g > 0 \quad \text{degree of the polynomial on the right-hand side}
\]

\[
\varepsilon_j \quad \text{constant coefficient of the } j^{th} \text{ power of the right-hand side polynomial}
\]

\[
a_j(x) = \frac{x^j}{j!} \quad \text{auxiliary function for formulation of polynomial}
\]

Now, we can rewrite the solution of the differential equation for our case and its derivation:

\[
T(x) = \sum_{i=0}^{1} T_0^{(i)} c_i(x) + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(x) = c_0(x)T_0 + c_1(x)T'_0 + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(x)
\] (10)

\[
T'(x) = c'_0(x)T_0 + c'_1(x)T'_0 + \sum_{j=0}^{g} \varepsilon_j b'_{j+2}(x)
\] (11)

Let us write again the constitutive equation (1) for 1D heat flux:

\[
q(x) = T(x)\alpha(x)f(x) - \lambda(x)T'(x)
\] (12)

Let the boundary conditions for thermal field are:

\[
T(L) = T_L \quad \text{ } q(0) = q_0
\] (13)

Then we can put together equation (12) expressed for position \(x = 0\) (position at node 0) and equation (10) expressed for \(x = L\) (node L), and using boundary conditions (13) after some mathematical operations we can write in matrix form (lower index “0” and “L” for used physical quantities means that concerned quantity is evaluated for position \(x = 0\) or \(x = L\), respectively):

\[
\begin{bmatrix}
c_0(L) + \frac{\alpha_0L}{\lambda_0} c_1(L) & -1
\end{bmatrix}
\begin{bmatrix}
T_0
\end{bmatrix}
= \begin{bmatrix}
\frac{c_1(L)}{\lambda_0} q_0 - \sum_{j=0}^{g} \varepsilon_j b_{j+2}(L)
\end{bmatrix}
\] (14)
Now, let the boundary conditions are changed:

\[ T(0) = T_0 \quad \text{and} \quad q(L) = q_L \quad (15) \]

Then we can put together equation (12) expressed for position \( x = L \), equation (11) expressed for node \( L \) and equation (10) expressed for node \( L \), and using boundary conditions (15) after some mathematical operations we can write in matrix form:

\[
\begin{bmatrix}
    c_0(L) - \frac{c_1(L)\alpha_i J_L}{c_1(L)} - \frac{c_1(L)\alpha_i J_L}{c_1(L)\lambda_L} - 1 & 1 \\
    c_0(L) - \frac{c_1(L)\alpha_i J_L}{c_1(L)} - \frac{c_1(L)\alpha_i J_L}{c_1(L)\lambda_L} - 1 & 1 
\end{bmatrix}
\begin{bmatrix}
    T_0 \\
    T_L
\end{bmatrix} =
\begin{bmatrix}
    c_1(L) q_L \\
    c_1(L) q_L
\end{bmatrix}
\begin{bmatrix}
    \sum_{j=0}^{g} \epsilon_j \beta_j^L + \sum_{j=0}^{g} \epsilon_j \beta_j^L \\
    -\sum_{j=0}^{g} \epsilon_j \beta_j^L
\end{bmatrix}
\]

The matrixes (14) and (16) can be put together. Comparing mathematical formulation and FEM formulation of our task we can find out, that there is sign difference in heat flux at node \( L \) (heat flux at node \( L \) for FEM formulation has opposite direction than it is in mathematical formulation). Considering these facts we get the system of FEM equations for thermal field in the conductor:

\[
\begin{bmatrix}
    c_0(L) + \frac{\alpha_i J_0}{\lambda_0} c_1(L) - 1 & 1 \\
    c_0(L) - \frac{c_1(L)\alpha_i J_L}{c_1(L)} - \frac{c_1(L)\alpha_i J_L}{c_1(L)\lambda_L} - 1 & 1 
\end{bmatrix}
\begin{bmatrix}
    T_0 \\
    T_L
\end{bmatrix} =
\begin{bmatrix}
    \frac{c_1(L) q_0}{\lambda_0} - \sum_{j=0}^{g} \epsilon_j \beta_j^L \\
    -\sum_{j=0}^{g} \epsilon_j \beta_j^L
\end{bmatrix}
\]

We can also evaluate the temperature within the range of the conductor. Using equation (12) expressed for node 0 and substituting it into equation (10) then we can write:

\[ T(x) = c_0(x)T_0 + c_1(x) \frac{T_0 \alpha_i J_0}{\lambda_0} - q_0 + \sum_{j=0}^{g} \epsilon_j \beta_j^L \quad (18) \]

Now, comparing the general form of the differential equation (5) with governing function (4) for electric current density in expanded form:

\[ \sum_{u=0}^{m} \eta_u(x)y^{(u)}(x) = \sum_{j=0}^{g} \epsilon_j a_j(x) \]

\[ \sigma(x)\varphi'(x) + \sigma'(x)\varphi'(x) = -\alpha'(x)\sigma(x)T'(x) - \alpha(x)\sigma'(x)T'(x) - \alpha(x)\sigma(x)T''(x) \]

we get:

- \( m = 2 \) – degree of the differential equation
- \( y(x) \equiv \varphi(x) \) – unknown function of independent variable \( x \) is function of el. potential
- \( y^{(u)}(x) \equiv \varphi^{(u)}(x) \) – \( u \)th derivation of unknown electric potential
- \( \eta_0(x) = 0 \) – 0th derivation of the potential is not present on the left-hand side
- \( \eta_1(x) = \sigma'(x) \) – non-constant coefficient of the 1st derivation of the potential
- \( \eta_2(x) = \sigma(x) \) – non-constant coefficient of the 2nd derivation of the potential
- \( g > 0 \) – degree of the polynomial on the right-hand side
- \( \epsilon_j \) – constant coefficient of the \( j \)th power of the right-hand side polynomial
- \( a_j(x) = \frac{x^j}{j!} \) – auxiliary function for formulation of polynomial of the right side

1185
Similar to the derivation process for thermal field we can derive also FEM equations for electric field:

\[
\begin{bmatrix}
-\frac{c_0(L)}{c_1(L)} + \frac{c_1(L)c_0'(L)}{c_1(L)} & 1
\end{bmatrix}
\begin{bmatrix}
\varphi_0 \\
\varphi_L
\end{bmatrix} =
\begin{bmatrix}
-\frac{c_1(L)}{\sigma_0} J_0 - c_1(L)\alpha_0 T_0' + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(L) \\
-\frac{c_1(L)}{\sigma_L} \left(\frac{-J_L}{\sigma_L} + \alpha_L T_L' + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(L)\right) + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(L)
\end{bmatrix}
\tag{20}
\]

And similar to the equation (18) for thermal field we can also evaluate the electric potential within the range of the conductor:

\[
\varphi(x) = c_0(x)\varphi_0 - c_1(x)\left(\frac{J_0}{\sigma_0} + \alpha_0 T_0'\right) + \sum_{j=0}^{g} \varepsilon_j b_{j+2}(x)
\tag{21}
\]

The temperature and electric potential are primary variables for thermal and electric field analyses, respectively. Calculating the secondary variables, like heat flux and electric current density, is possible using equation (1) expressed for one-dimensional task. But this is suitable only for 1D model where we calculate one longitudinal distribution of the heat flux or electric current density for homogenized model, respectively. However, our 1D model results from real 3D FGM conductor, so it is more realistic to calculate the secondary variables in layers that were considered during homogenization process (see [1] for details).

FEM equations for two-way coupled thermoelectric analysis are equations (17) and (20). They are solved using iterative algorithm. During iteration process it is necessary to find substitutional functions for results obtained from FEM equations (results of these FEM equations are not continuous functions but only sets of discrete values) and also it is necessary to convert non-polynomials into polynomials (e.g. see $P_j$ in equation (3)). Iteration process can be set ahead by evaluating equations (1) expressed for one-dimensional task also within individual iterative steps.

5 THERMOELECTRIC ANALYSIS OF FGM LINK CONDUCTOR - NUMERICAL EXPERIMENT

In this chapter there will be one academic example of thermoelectric analysis of given FGM link conductor presented. The task will be solved using our new approach, by commercial FEM code ANSYS and also by numerical solution of differential equations in software Mathematica due to comparison reasons.

Let us consider electric conductor with circular cross-section according to Figure 1. Its length is $L = 500$ [mm] and diameter $d = 10$ [mm]. Let the conductor consists of mixture of two component materials – matrix (index $m$) with constant electric conductivity $\sigma_m(x,r) = 1.429 \times 10^6$ [Sm$^{-1}$] and thermal conductivity $\lambda_m(x,r) = 1.333$ [Wm$^{-1}$K$^{-1}$], and fibre (index $f$) with electric conductivity $\sigma_f(x,r) = 1.111 \times 10^7$ [Sm$^{-1}$] and thermal conductivity $\lambda_m(x,r) = 450$ [Wm$^{-1}$K$^{-1}$]. Axial symmetry of final material properties is assumed. Volume fraction of individual components is functionally changed according to chosen polynomial, graphically shown in Figure 1:
\[ v_f(x,r) = 0.2148 + 1.6972x^2 - 2.2629x^3 - 92.316r + 7197.8x^2r - 9597.1x^3r + 7148.4r^2 - 1.3548 \times 10^6x^2r^2 + 1.8064 \times 10^6x^3r^2 + 944433r^3 - 1.1333 \times 10^7x^2r^3 + 1.5111 \times 10^7x^3r^3 \]
\[ v_m(x,r) = 1 - v_f(x,r) \]

Using extended mixture rule for chosen number of layers (\( N = 11 \)) we get longitudinal variation of effective electric and thermal conductivities for individual layers and using laminate theory we can calculate also homogenized electric and thermal conductivity of FGM conductor, see Figure 4.

The equations of homogenized electric and thermal conductivities are:
\[ \sigma^H(x) = 1.8793 \times 10^6 + 7.926 \times 10^7x^2 - 1.0568 \times 10^8x^3 \text{ [Sm}^{-1}] \]
\[ \lambda^H(x) = 22.217 + 3672.7x^2 - 4897x^3 \text{ [Wm}^{-1}K^{-1}] \]

**Figure 4:** Longitudinal distribution of the homogenized electric and thermal conductivity (red) and the effective electric and thermal conductivities in respective layers for \( N = 11 \) (blue)

Let us consider final Seebeck coefficient for whole conductor according to chosen polynomial function (academic example, without considering homogenization process based on mixture of the components):
\[ \alpha(x) = -2 \times 10^{-4} + 8 \times 10^{-4}x \text{ [VK}^{-1}] \]

We assume static state for thermoelectric analysis. In nodes 0 and L there are electric potentials and temperatures specified so boundary conditions (see Figure 5) are:
\[ \varphi(0) = 0 \text{ [V]}; \ T(0) = 273 \text{ [K]} \]
\[ \varphi(L) = -0.09 \text{ [V]}; \ T(L) = 283 \text{ [K]} \]

**Figure 5:** Boundary conditions of the model
We also created 2D model in code ANSYS [6], we used 53 000 PLANE223 elements (8 node quad-elements). We considered axial symmetry for the model. The task was also solved in software Mathematica [7], where the differential equations (4) with specified boundary conditions and homogenized material properties were numerically solved using iterative algorithm. Finally, the task was also solved by only one our new developed two-nodal link element using FEM equations (17) and (20) for nodal points of the link and with equations (18) and (21) for chosen points within the link. In Figure 6 and Figure 8 we can see calculated longitudinal distribution of the electric potential and temperature in the conductor, respectively. In Figure 7 and Figure 9 there are shown distributions of the electric current densities and heat fluxes for chosen layers (1st, 6th and 11th layer), respectively. Summary of calculated results is in Table 1.

**Figure 6**: Distribution of the electric potential through the length of conductor

**Figure 7**: Longitudinal distribution of the current densities in the chosen layers of conductor
Figure 8: Distribution of the temperature through the length of conductor

Figure 9: Longitudinal distribution of the heat fluxes in the chosen layers of conductor

Table 1: Comparison of calculated electric and thermal quantities for chosen layers and homogenized values in nodal points of the conductor

<table>
<thead>
<tr>
<th>$J_{\text{layer, node}} \times 10^8$ [Am$^{-2}$]</th>
<th>$J_{1,0}$</th>
<th>$J_{1,L}$</th>
<th>$J_{6,0}$</th>
<th>$J_{6,L}$</th>
<th>$J_{11,0}$</th>
<th>$J_{11,L}$</th>
<th>$J^H_0$</th>
<th>$J^H_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANSYS</td>
<td>13.4104</td>
<td>4.3245</td>
<td>7.2969</td>
<td>9.1219</td>
<td>7.0017</td>
<td>4.2744</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mathematica</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.2179</td>
<td>7.2179</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q_{\text{layer, node}} \times 10^4$ [Wm$^{-2}$]</th>
<th>$q_{1,0}$</th>
<th>$q_{1,L}$</th>
<th>$q_{6,0}$</th>
<th>$q_{6,L}$</th>
<th>$q_{11,0}$</th>
<th>$q_{11,L}$</th>
<th>$q_0^H$</th>
<th>$q_L^H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>new element</td>
<td>-10.4785</td>
<td>1.2150</td>
<td>-4.6806</td>
<td>2.1449</td>
<td>-4.3673</td>
<td>1.2435</td>
<td>-5.6199</td>
<td>1.7838</td>
</tr>
<tr>
<td>ANSYS</td>
<td>-10.1082</td>
<td>1.3250</td>
<td>-4.7843</td>
<td>2.3242</td>
<td>-4.2698</td>
<td>1.3931</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mathematica</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-5.6108</td>
<td>1.8706</td>
</tr>
</tbody>
</table>
There is small difference in the results (secondary variables) between ANSYS solution and calculation using the new approach in nodal points because of substitutional functions used for conversion non-polynomials into polynomials during iterative process. But we can see from Figure 6 – Figure 9 that obtained results correspond to ANSYS 2D axial symmetry simulation very well.

6 CONCLUSION

New finite link element for two-way coupled static thermoelectric analyses has been developed in this contribution. New FEM equations with consideration Joule heat, and thermoelectric effects, like Seebeck, Peltier and Thomson effects, were derived. Numerical example with good agreement between calculations with just only one new link element and commercial FEM code that uses numbers of classic elements have been presented. The new approach fully agrees with numerical solution for 1D differential equation of thermal and electric fields calculated using iterative algorithm. So, effectiveness and accuracy of the new developed link element for these analyses are excellent.

ACKNOWLEDGEMENT

This contribution is the result of the project implementation Centre of Competence in New Materials, Advanced Technologies and Energy, ITMS 26240220073 supported by the Research & Development Operational Programme funded by the ERDF.

This work was supported by Grant Agency VEGA, project No. 1/0534/12.

This work was supported by Grant Agency KEGA, project No. 015STU-4/2012.

REFERENCES


[6] ANSYS Swanson Analysis System, Inc., 201 Johnson Road, Houston, PA 15342/1300, USA.

NUMERICAL SIMULATION FOR EVAPORATION OF CHEMICAL AGENT DROPLET

H. HAYAKAWA*, M. YAMAMOTO†, M. KAWAI‡ AND K. SAKAMOTO§

* Graduate School of Mechanical Engineering
Kudan-Campus, Tokyo University of Science
1-14-6 Kudankita, Chiyoda-ku, Tokyo 102-0073, Japan
email: j4512650@ed.tus.ac.jp

† Department of Mechanical Engineering
Kudan-Campus, Tokyo University of Science
1-14-6 Kudankita, Chiyoda-ku, Tokyo 102-0073, Japan
email: yamamoto@rs.kagu.tus.ac.jp

‡ Numerical Engineering Dept, IHI
1, Shin-Nakahara-Cho, Isogo-ku, Yokohama 235-8501, Japan
email: masafumi_kawai@ihi.co.jp

§ Ordnance Engineering Department Defense Systems Division
Aero-Engine & Space Operations, IHI
229, Tonogaya, Mizuho-machi, Nishitama-gun, Tokyo 190-1297, Japan
email: kazuyuki_sakamoto@ihi.co.jp

Key words: Computational Fluid Dynamics, Chemical agent droplet, Evaporation.

Abstract. Recently, terrorism is one of most daunted dangers in the world. Various terrorism can come about. Biological and chemical terrors have particularly high mortality rate, and tend to wreak the secondary disaster. For these reasons, research and development of measures for decontaminations are emergent and of key issue on a world scale. Sulfur mustard (HD) is one of chemical agent used to commit terrorism. HD gas has severe toxicity and long-period damage because of persistent agent. In addition, as HD gas achieves the toxicity even if it only contacts on a skin, it is important that HD-contaminated objects are decontaminated quickly. In the present study, HD gas emission from a surface of HD droplet is numerically researched. HD gas emission is usually investigated with experiments, but the experiment is so dangerous because of the toxicity. Therefore, in this study, CFD (computational fluid dynamics) is used to reproduce the emission. We propose a volatilization model which can estimate the volatilization of a chemical agent with the saturated vapor pressure of chemical agent and physical quantities of the air. We apply our three dimensional volatilization prediction code to a HD droplet placed on a wind tunnel wall. Comparing the numerical results with the experimental data, it is confirmed that the numerical data are in good agreement with the experimental data. In addition, volatilization volume is coupled with droplet geometry to simulate the temporal change of droplet shape and residual mass. It is reasonably predicted that the droplet volume gradually decreases as volatilization proceeds.
until the droplet disappears.

1 INTRODUCTION

There have been various terror attacks around the world since 9.11 in 2001 [1]. Terrorists hold extremely wide-ranging means of terrors such as bombs and toxic gases. Above all, nuclear, biological and chemical weapons have particularly high mortality rate so that they are recognized as being the most dangerous in all other ways of terror. Among these weapons, biological and chemical weapons are relatively-easy to be produced so that they are at high risk of being used in terrorisms. In the past, there have been some cases by the use of chemical weapons. In 1978, a refugee from Bulgaria was killed by the pellet containing ricin. In Japan, sarin was used to commit indiscriminate murders in Matsumoto city in 1994 and in the Tokyo subway in 1995. Sarin is extremely dangerous nerve agent so that 5,510 people were injured and 12 people were killed in the Tokyo subway attack. In terror attacks, many civilians can be victimized. Furthermore, it is hard to predict when and where it happens so that research and development against unanticipated terror attacks are emergent. However, research and development of chemical weapons are of great difficulty since the experiments are highly risky and are legally-prohibited in many cases. Therefore, numerical simulation is very useful to research it. HD (Sulfur mustard) is a type of blister agent. It affects eyes, skins and respiratory organs. In addition, HD is a persistent agent so that the damage continues over a long period.

In the present study, computational fluid dynamics is used to simulate the evaporation of HD droplets. Fick’s law is used to calculate the droplet’s volatilization rate and volatilization volume is coupled with droplet geometry to simulate the temporal change of droplet shape and residual mass. In addition, the versatility of our simulation code is confirmed by applying it to various conditions of temperatures and wind speeds. It is reasonably predicted that the droplet volume gradually decreases as volatilization proceeds until the droplet disappears.

2 NUMERICAL PROCEDURE

In this study, volatilization of chemical agent droplet placed in a steady flow is investigated. The flow and concentration fields with chemical agent vapor in a wind tunnel shown in Figure 1 is computed. Since the flow is assumed to be multiphase flow consisting of air and chemical agent, we use a homogenous fluid model. As the treatment of multiphase flow, we apply the two-way coupling method, that is, the concentration and the flow strongly affects each other. The numerical procedures for the flow field, the concentration field, the volatilization model and the coupling method of volatilization volume and droplet geometry will be described below.

2.1 Flow field

It is assumed that the flow field is three-dimensional and incompressible. We use the conservation equations of mass and momentum expressed by Eqs.(1) and (2) as the governing equations. For pressure-velocity coupling, the Low-Mach-Number approximated MAC method (marker and cell method) is used. Since we treat two-phase flow consisting of air and chemical agent, the mixture density calculated by Eqs.(3) and (4) at each grid points are considered in the flow computations. No-slip condition between air and chemical agent is
imposed to calculate the mixture gas velocity, $\vec{u}_{mix}$.

$$\frac{\partial \rho_{mix}}{\partial t} + \nabla \cdot (\rho_{mix} \vec{u}_{mix}) = 0$$

(1)

$$\frac{\partial \rho_{mix} \vec{u}_{mix}}{\partial t} + \vec{u}_{mix} \cdot \nabla (\rho_{mix} \vec{u}_{mix}) = -\nabla p + \mu_{air} \nabla^2 \vec{u}_{mix} - (\rho_s - \rho_{mix})\overrightarrow{g}$$

(2)

$$\rho_{mix} = \alpha_{air} \rho_{air} + \alpha_{HD} \rho_{HD}$$

(3)

$$\alpha_{air} + \alpha_{HD} = 1$$

(4)

In the above equations, $\alpha$ is the volume fraction, $\rho$ is the density, $u$ is the flow velocity, $t$ is the time, $p$ is the pressure, $\mu$ is the viscosity coefficient, $g$ is the gravity acceleration, the subscript $air$ is for the air, $HD$ is for the chemical agent (HD), $mix$ is for the mixture gas and $s$ is for the average value of overall computed domain. Since the volume fraction of chemical agent is low enough, typically 0.03 percent, the viscosity of air is used in Eq.(2). The gravity term is calculated by the difference between the average density and the local density.

A finite difference method is applied to discretize the governing equations. The Kawamura-Kuwahara [2] third-order upwind scheme is used for the convection terms, the Euler explicit scheme for the time marching terms, and the second order central difference scheme for the other terms.

2.2 Concentration field

For the concentration field, the governing equation is given as

$$\frac{\partial C_{HD}}{\partial t} + \vec{u}_{HD} \cdot \nabla C_{HD} = D_{HD} \nabla^2 C_{HD}$$

(5)

where $C$ is mol concentration and $D$ is the diffusion coefficient. $D$ is calculated from the estimate equation by Fuller et al. [3]. The numerical difference schemes are employed in similar to the flow field.

2.3 Volatilization model

In the volatilization volume calculation, at first, the diffusion flux is calculated by Fick’s law expressed by Eq.(6). The volatilization volume is calculated by diffusion flux integrated by time and surface area. It is expressed by Eq.(7).

$$J = -D \frac{dC_{HD}}{dx}$$

(6)

$$m = \int \int J dx dt$$

(7)

In these equations, $J$ is the diffusion flux, $x$ is the coordinate normal to the droplet surface, and $m$ is the volatilization volume.

In the volatilization model, whether chemical agent volatilize or not is judged in the following way. At first, the partial pressure of chemical agent is calculated by the state
equation. Next, if the partial pressure is lower than the saturated vapor pressure, the chemical agent is volatilized. On the other hand, if the partial pressure is higher than the saturated vapor pressure, the chemical agent is not volatilized. This judgement is made for each grid points on the droplet surface. Finally, volatilization volume is calculated.

2.4 Coupling method

In this study, temporal change of droplet volume with the progress of evaporation is reproduced by coupling of volatilization volume and droplet geometry. In response to the deformation of droplet geometry, the flow and concentration fields change. Furthermore, the change in concentration field affects the volatilization rate. Thus, in this evaporation system, concentration field, volatilization rate and droplet geometry are affected among others.

HD evaporation is long-term phenomena due to its persistent property. On the other hand, the time scale of flow and concentration fields is short. Therefore, the time scales of the two are remarkably different. For this reason, it is difficult that long-term time integration is performed in accordance with the short time scale of flow and concentration fields. Therefore, in this study, droplet geometry is estimated by using the steady solution of flow and concentration fields as long as the change of droplet geometry is not large enough to remarkably affects the flow and concentration. In other words, their mutual effects are considered by the iteration of the cycle consisting of “the flow and concentration fields” and “the change of droplet geometry by evaporation”. This method enables us to simulate the evaporation phenomena without enormous computational cost.

Figure 1: Computational domain, grid and examples of flow and concentration field
3 COMPUTATIONAL CONDITION

Computations are performed for the flow in the half geometry of 5cm × 5cm wind tunnel as shown in Figure 1. We employed this geometry so as to match the condition of the experiment performed by Danberg [4]. To reduce the computational costs, only the half area of the wind tunnel is computed. The chemical agent is considered to be HD, and the droplet is placed on the wind tunnel wall as shown in Figure 1. The number of grid points is 240 × 90 × 38, clustered near the droplet, walls and inlet boundaries, where the gradients of physical quantities are expected to be steep. On the wall boundary, no-slip condition is imposed. On the upper surface, slip boundary condition is applied. On the inlet boundary, velocity is fixed and others are extrapolated from the computational domain.

In order to investigate the effect of temperature and flow speed, the temperature is varied within the range of 15 degrees to 50 degrees, and the inflow velocity is changed from 0.12 to 2.57 m/s.

Validity of this numerical code was carried out with the combination of three non-dimensional numbers defined by Eqs.(8) to (10).

\[ Sh = \frac{N \cdot L}{(c_w - c_\lambda) \cdot D} \]  
(8)

\[ Re = \frac{uL}{\nu} \]  
(9)

\[ Sc = \frac{\nu}{D} \]  
(10)

In these definitions, \( N \) is the volatilization rate, \( L \) is the characteristic length, \( c_w - c_\lambda \) is the concentration difference of the evaporating vapor between the surface and the edge of the evaporating vapor dispersion layer, \( u \) is the characteristic velocity and \( \nu \) is the kinetic viscosity. \( Sh \), \( Re \) and \( Sc \) are Sherwood number, Reynolds number and Schmidt number, respectively. The computational conditions are listed in Table 1. We set Case7 as the reference case.

4 NUMERICAL RESULT AND DISCUSSION

4.1 Validation of developed code

The temporal change of residual and volatilization rate of the reference case (i.e. Case7) is exhibited in Figure 2. It indicates that volatilization rate become constant after about one second.
Table 1: Computational conditions

<table>
<thead>
<tr>
<th>Case</th>
<th>Temperature [degrees]</th>
<th>Mainstream velocity [m/s]</th>
<th>Diffusivity [m²/s]</th>
<th>Saturated vapor pressure [N/m²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>0.12</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0.26</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.64</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>1.28</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>1.92</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>2.57</td>
<td>6.56×10⁻⁶</td>
<td>5.01</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>0.12</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>0.26</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>0.64</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>10</td>
<td>35</td>
<td>1.28</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>11</td>
<td>35</td>
<td>1.92</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>12</td>
<td>35</td>
<td>2.57</td>
<td>7.38×10⁻⁶</td>
<td>32.4</td>
</tr>
<tr>
<td>13</td>
<td>50</td>
<td>0.12</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
<tr>
<td>14</td>
<td>50</td>
<td>0.26</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>0.64</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>1.28</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>1.92</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>2.57</td>
<td>8.02×10⁻⁶</td>
<td>105.4</td>
</tr>
</tbody>
</table>

Figure 2: Temporal change of residual and volatilization rate of reference case
Figure 3 shows the relation among $Sh$, $Re$ and $Sc$. The line indicates the experiment and the symbols do the present computation.

$$Sh = G \cdot Re^{2/3} Sc^{1/3}$$

(11)

Based on the experimental data, the mean gradient $G$ is 0.828. Our computation gives $G$ is 0.836. Hence, it is confirmed that our numerical code is reasonably validated.

Figure 4(a) shows the steady distribution of HD concentration in the reference case (Case 7). The molecular mass of HD is much heavier than that of air so that HD settles near the bottom of the tunnel soon after the volatilization. Furthermore, HD concentration near the droplet surface increases remarkably. This trend affects the volatilization rate of the downstream. Figure 4(b) exhibits the steady distribution of volatilization rate over the droplet surface in the reference case. The volatilization rate of the downstream side is relatively lower than that of the upstream side. This is because that the concentration of HD near the downstream side is increased by volatilized HD vapor at the upstream, and thus the volatilization rate of the downstream side decreases.

**Figure 3:** Relation among non-dimensional parameters

**Figure 4:** Numerical Results of the reference case (Case 7)
4.2 Evaporation simulation

Figure 5 shows the distributions of HD concentration and volatilization rate. In these figures, the droplet becomes progressively smaller along with time. Computations were performed until the droplet volume becomes zero. The computational result suggests that it takes over 13 hours to dry off 9 micro liters droplet. It comes from the strong stability of HD. Figure 6 plots the temporal change of droplet volume. It indicates that the change of droplet volume is non-linear. The smaller droplet volume is, the smaller the surface area becomes. As a result, volatilization rate for the whole surface area decreases.

4.3 Effect of temperature and main stream speed

Figure 7(a) shows the temporal changes of droplet volume for the cases that the inflow velocity is fixed at 0.12 m/s and the temperature is varied from 15 to 50 degrees. It is clear that evaporation speed increases with increasing the temperature. This trend is caused by the dependence of saturated vapor pressure on temperature. The value of saturated vapor pressure affects the suppression of volatilization.

Figure 7(b) shows the temporal changes of droplet volume for the cases that temperature is fixed at 35 degrees and the inflow velocity is varied from 0.12 to 0.64 m/s. From this graph, evaporation speed increases with increasing the inflow velocity. This is because that HD vapor is effectively blown at high speed and it results in the low concentration of HD near the droplet surface.

Through these computations, we confirmed that our numerical code can simulate the evaporation of chemical agent droplet in line with the principle of evaporation within the applied range of temperature and wind speed.

![Figure 5: Numerical Results of evaporation simulation](image)
Figure 6: Temporal change of droplet volume

Figure 7: Temporal change of droplet volume under different conditions

(a) Effect of temperature

(b) Effect of inflow velocity
5 CONCLUSIONS

We conducted three-dimensional computations for evaporation of chemical agent droplets in the wind tunnel. Through this study, following insights were obtained.

- Our volatilization code can predict the volatilization of chemical agent droplets reasonably.
- In volatilization of a droplet, the volatilization rate tends to be large at upwind side of the droplet.
- Our simulation code can reproduce the evaporation of chemical agent droplet by coupling volatilization volume with droplet geometry.
- The increase of temperature and main stream speed enhance the volatilization.

REFERENCES
A MULTISCALE DAMAGE MODEL FOR COMPOSITE MATERIALS USING A FFT-BASED METHOD

JOHANNES SPAHN*, HEIKO ANDRÄ†, MATTHIAS KABEL* AND RALF MÜLLER†

*Fraunhofer Institute for Industrial Mathematics (ITWM)
Department Flow and Material Simulation
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: spahn@itwm.fraunhofer.de, web page: http://www.itwm.fraunhofer.de

†Institute of Applied Mechanics
Technische Universität Kaiserslautern
Gottlieb-Daimler-Straße, 67653 Kaiserslautern, Germany
web page: http://www.mechanik.mv.uni-kl.de

Key words: Multiscale Analysis, Computational Homogenization, Continuum Damage Mechanics, FFT, Composite Materials, Localization

Abstract. Modeling failure and progressive damage of composite materials presents a challenging task and is currently subject of many research activities in the field of computational mechanics. Conventional methods which assume constant material coefficients or global failure criteria, are in many cases not sufficient to predict the appropriate mechanical material response. Composite failure occurs as a result of complex meso-structural damage mechanisms and therefore it is preferable to capture these nonlinear material effects directly on a finer scale. Hence, recent multiscale modeling and simulation techniques were developed to consider the mesoscopic material behavior. In this contribution we propose an alternative multiscale approach similar to FE². Nonlinear material effects caused by progressive damage behavior are captured on a finer length scale. The constituents are modeled explicitly and simple isotropic damage laws are used to describe the constitutive behavior. Hence, the resulting material response is based on genuine physical effects and only a few material parameters are required which can be measured directly in physical experiments. The fine scale problem (material level) is reformulated into an integral equation of Lippmann-Schwinger type and solved efficiently using the fast Fourier transformation (FFT). The calculation is carried out on a regular voxel grid which can be obtained from 3D images like tomographies without using any complicated mesh generation. Furthermore, the fine scale problem is integrated in a standard Finite Element framework which is used to solve the macroscopic BVP (component level).
1 INTRODUCTION

Fiber-reinforced composites possess a complex material response, which is not sufficiently understood at present time. Especially modeling failure and progressive damage of composite materials, presents a major challenge in current research activities. Composite failure occurs as a result of a variety of complex mesostructural damage mechanisms, such as matrix damage, fiber pull out and fiber breakage.

Recently, phenomenological macroscale models are state of the art for failure investigations in many applications. These models assume homogeneous material behavior and are usually based on macroscopic failure criteria. Disadvantages are the extensive identification of material parameters in dependence of the material structure and the loading conditions. Physical phenomena occurring on a finer length scale are not considered.

A more accurate approach is to capture the nonlinear material effects directly on a finer scale. Therefore, the mesostructural constituents are modeled explicitly on the interesting scale. The resulting material response is based on genuine physical effects and consequently arbitrary complex non-proportional, multiaxial loading conditions can be captured. Moreover, simple (isotropic) constitutive laws are used to define the material behavior of the mesostructural constituents and the required parameters can be measured directly in physical experiments. However, the detailed resolution of the mesostructural constituents leads to a fine discretization of the computational model and thus to large algebraic systems with many degrees of freedom. Recent multiscale modeling and simulation techniques were developed to restrict the computational effort to an acceptable extent and nevertheless capture the mesomechanical material effects in a proper way [8].

One well-known method is the FE\(^2\) approach [4]. The scales are solved separately and each macroscopic point is equipped with a certain mesostructure. Therefore the constitutive equation for the coarse scale is replaced by an additional mesoscopic boundary value problem (BVP) which is solved on a representative volume element (RVE).

In this work an alternative method is presented which uses the Fast Fourier Transformation (FFT) to solve an equivalent elastic meso BVP (material level). A periodic BVP known from ordinary elasticity problems is reformulated in an integral equation of the so called Lippmann-Schwinger type [20, 9]. This method was introduced by Moulinec and Suquet [14, 15]. Advantages of this method are its efficiency in terms of memory consumption and computational time. Further the calculation is carried out on a regular voxel grid and could therefore directly be applied to calculate homogenized quantities on 3D images like tomographies without using any complicated mesh generation. The fine scale problem can easily be integrated in a standard Finite Element framework which is used to solve the macroscopic BVP (component level).

In the first part of this paper the constitutive equation and the numerical solution of the mesoscale model are introduced. The second part presents the coupling technique of the two geometrical scales and the principles of numerical homogenization are explained. Finally the paper closes with numerical examples of some simple scale coupling problems.
2 MESOSCALE MODEL

In this section the constitutive equation and the numerical solution of the mesoscopic BVP is introduced.

2.1 Isotropic Nonlocal Damage Model

In this work we use a nonlocal formulation of a strain based continuum damage model according to Simo & Ju [18], whereby any other constitutive model can be used to describe the material behavior of the mesoscopic constituents.

In the context of Continuum Damage Mechanics (CDM), Kachanov [7] introduced an internal variable \( d \) ranging from 0 to 1. While \( d = 0 \) represents the undamaged state, \( d = 1 \) describes the status of completely failed material. The constitutive equation is derived from a thermodynamic state potential, the Helmholtz free energy \( \psi \):

\[
\psi = \frac{1}{2} (1 - d) \varepsilon : C_{el} : \varepsilon ,
\]

(1)

\[
\sigma = \frac{\partial \psi}{\partial \varepsilon} = (1 - d) \ C_{el} : \varepsilon ,
\]

(2)

where \( C_{el} \) is the isotropic elasticity tensor. The specific strain energy \( \frac{1}{2} \varepsilon : C_{el} : \varepsilon \) is the thermodynamic force \( Y \) associated with the internal state variable \( d \):

\[
Y = -\frac{\partial \psi}{\partial d} = \frac{1}{2} \ varepsilon : C_{el} : \varepsilon .
\]

(3)

A damage criterion \( f \) controls the state of damage, which is in the case of growing damage described by a monotonic function \( \phi(Y) \):

\[
f = \phi(Y) - d \leq 0 \quad \left\{ \begin{array}{ll}
f & < 0 \quad \text{elastic} \\
f & = 0 \quad \text{damage} .
\end{array} \right.
\]

(4)

An analytical integration of the evolution law yields an exponential expression for the growing damage variable \( d \):

\[
d = \phi(Y) = 1 - e^{-H(\tilde{\varepsilon}(Y) - Y_0)} \quad d \in [0, 1) .
\]

(5)

In the equation above the material parameters \( H \) as the damage hardening modulus and \( Y_0 \) as the initial damage threshold are introduced. According to Simo & Ju [18] the strain energy \( Y \) is replaced by an equivalent strain measurement \( \tilde{\varepsilon}(Y) \) which is obtained by a small modification of \( Y \), namely the energy norm of the strain tensor:

\[
\tilde{\varepsilon}(Y) = \sqrt{2Y} = \sqrt{\varepsilon : C : \varepsilon} .
\]

(6)

There exist several proposals to calculate the equivalent strain [10, 12]. Choosing the energy norm of the strain tensor as an equivalent strain measurement in conjunction with
the thermodynamic consistent (associated) damage formulation, ensures the symmetry of
the tangential material moduli [18].

Increasing damage leads to local softening behavior, the tangential material stiffness
becomes negative and consequently strain localization effects occur. The strain localizes
in certain regions, while the surrounding area gets unloaded. The size of the localizing
area is related to the spatial resolution of the mesh and consequently the solution becomes
dependent on the existing discretization [3]. The reason is the loss of ellipticity of the
governing differential equation and consequently the loss of stability and uniqueness of
the solution [17].

In order to avoid mesh dependence of the computational model, regularization methods
are used. In this context enhanced models enriched by spatial gradients of the field
variables were developed (e.g. [2]). In this work a nonlocal approach according to [16] is
used. Hereby the damage variable $d$ is averaged over its surrounding area and integrated
in the constitutive equation:

$$\sigma = (1 - \bar{d}) C_{el} : \epsilon .$$  \hspace{1cm} (7)

The averaging is performed by applying a Gaussian weight function $W$ on the damage
variable $d$ on certain points $y$ in the surrounding area:

$$\bar{d}(x) = \frac{1}{\int V} \int W(x, y) dV(y) \int W(x, y) d\langle y \rangle dV(y) \hspace{1cm} (8)$$

$$W(x, y) = \frac{1}{(2\pi)^{2/3}l^3} \exp \left(-\frac{||x - y||^2}{2l^2}\right) \hspace{1cm} (9)$$

It is mentioned here, that by this procedure a length scale $l$ is introduced in the damage
model. To assure mesh independent solutions the discretization must be able to resolve
this length $l$.

2.2 Equivalent BVP and FFT-based Numerical Solution

For computing effective quantities of a periodic medium with local stiffness $C(x)$ in
the context of a numerical homogenization process, a cubic RVE $\omega \in \mathbb{R}^3$ with periodic
boundary conditions on $\partial \omega$ is considered. The local strain field is split into a prescribed
constant macroscopic strain $E$ and a fluctuation term $\epsilon(u^*(x))$.

$$\text{div} \sigma(x) = 0 \hspace{1cm} \text{on } \omega \hspace{2cm} (10a)$$

$$\sigma(x) = (1 - \bar{d}(x)) C(x) : \epsilon(x) \hspace{1cm} \text{on } \omega \hspace{2cm} (10b)$$

$$\epsilon(x) = E + \frac{1}{2} \left( \nabla u^*(x) + (\nabla u^*(x))^T \right) \hspace{1cm} \text{on } \omega \hspace{2cm} (10c)$$

$$u^*(x) \text{ periodic} \hspace{2cm} \text{x on } \partial \omega_D \hspace{2cm} (10d)$$

$$\sigma(x) \cdot n(x) \text{ anti-periodic} \hspace{2cm} \text{x on } \partial \omega_N \hspace{2cm} (10e)$$
According to Zeller & Dederichs [20] the differential equation (10a) can be reformulated in an integral equation, the so called Lippmann-Schwinger equation, which is attributed to Lippmann & Schwinger [11] in the field of quantum mechanics. By introducing a homogeneous reference material with the stiffness $C_0$, the polarization stress $\tau$ with respect to this reference material the resulting constitutive equation read as follows:

$$\begin{align*}
\tau(x) &= (1 - \bar{d}(x)) C(x) : \varepsilon(x) - C^0 : \varepsilon(x) \\
\sigma(x) &= C^0 : \varepsilon(x) + \tau(x).
\end{align*} \tag{11}$$

The solution of the local problem in equation (10) can now be obtained using a nonlocal Green operator $\Gamma^0$ which is applied on the stress polarization $\tau$:

$$\varepsilon(x) = E - (\Gamma^0 * \tau)(x). \tag{13}$$

The operator $\Gamma^0$ is only associated with the stiffness of the homogenous linear elastic reference material $C^0$ and the given boundary conditions, but does not depend on the fluctuating quantities [9]. The convolution (denoted by ‘*’) in equation (13) is defined by:

$$\left(\Gamma^0 * \tau\right)(x) = \int_\omega \Gamma^0(x - y) : \tau(y) \, dy. \tag{14}$$

In the Fourier space it transforms into a direct product and the corresponding relation reads:

$$\hat{\varepsilon}(\xi) = -\hat{\Gamma}^0(\xi) : \hat{\tau}(\xi), \quad \forall \xi \neq 0, \ \hat{\varepsilon}(0) = E, \tag{15}$$

where $\hat{f}(\cdot)$ denotes a function in the Fourier space and $\xi$ the Fourier space variable corresponding to the coordinates $x$.

Substituting the expression of $\tau$ (11) in (13) we arrive at a nonlinear Lippmann-Schwinger equation:

$$\varepsilon(x) = E - (\Gamma^0 * ((1 - \bar{d})C - C^0) : \varepsilon)) (x). \tag{16}$$

There exist different numerical schemes to solve the Lippmann-Schwinger integral equation iteratively. In this work the so called basic scheme is used, which was introduced by Moulinec & Suquet [15]. The benefit of the scheme is based on the fact that the constitutive equation is solved in the real space and the convolution integral in the Fourier space. The transformation of the calculated fields is performed by a discrete fast Fourier transformation (FFT). Each iteration consists of four steps:
1. Solve the constitutive equation in the real space:
\[ \tau^i = ((1 - \bar{d})C - C^0) : \varepsilon^i \]

2. Fourier transformation of the stress polarization field:
\[ \hat{\tau}^i = \text{FFT}(\tau^i) \]

3. Convolution with the Green operator in the Fourier space:
\[ \hat{\varepsilon}^{i+1} = -\hat{\Gamma}^0 : \hat{\tau}^i, \quad \hat{\varepsilon}^{i+1}(0) = E \]

4. Inverse Fourier transformation of the updated strain field:
\[ \varepsilon^{i+1} = \text{FFT}^{-1}(\hat{\varepsilon}^{i+1}) \]

According to [15] convergence is reached when the global stress field is in equilibrium:
\[ \frac{\|\text{div} \ \sigma^{i+1}\|^2}{\|\sigma^{i+1}\|^2} \leq \text{tol}_1, \quad (17) \]

which can easily be computed in the Fourier space:
\[ \frac{\|\xi \cdot \sigma^{i+1}(\xi)\|^2}{\|\sigma^{i+1}(0)\|^2} \leq \text{tol}_1. \quad (18) \]

Thereby \( \sigma(0) \) equals the average or macroscopic stress and \( \| \cdot \|^2 \) denotes the \( L^2 \) norm of the appropriate field variable.

In this work a second convergence criterion is applied, which controls the difference of the strain field between two iterations in the real space (see e.g. [1]):
\[ \frac{\|\varepsilon^{i+1} - \varepsilon^i\|^2}{\|E\|^2} \leq \text{tol}_2. \quad (19) \]

Typical values for the convergence tolerances are \( \text{tol}_1 = 10^{-4} \) and \( \text{tol}_2 = 10^{-10} \).

The elastic parameters of the reference material \( C^0 \), which have important influence on the convergence rate of the scheme, are chosen according to [15]:
\[ k^0 = \frac{1}{2} \left( \min_x k(x) + \max_x k(x) \right), \quad \mu^0 = \frac{1}{2} \left( \min_x \mu(x) + \max_x \mu(x) \right), \quad (20) \]

where \( \mu \) and \( k \) denote shear and bulk modulus of an isotropic material.

It should be mentioned that in the last decade certain numerical schemes were developed to improve the convergence behavior of the FFT method for materials with high stiffness contrasts. A review of different schemes and an analysis of the convergence behavior for the computation of precise bounds of effective properties in comparison with analytical estimates can be found in [6].
3 MULTISCALE APPROACH

For the investigation of the macroscopic damage behavior the mesoscopic model is extended to a multiscale framework. Therefore the two scales are related to each other by a procedure similar to the FE$^2$ approach. Thus the constitutive equation on the macro scale is replaced by a BVP on the meso scale (see figure 1).

Each macroscopic point is linked to a RVE which represents the particular meso-structure at this point. The variables which are transferred between both scales are defined by volume averages over the meso domain:

\[
\langle \cdot \rangle = \frac{1}{\omega} \int_{\omega} (\cdot) \, d\omega ,
\]

\[
\mathbf{E} = \langle \varepsilon \rangle , \quad \mathbf{\Sigma} = \langle \sigma \rangle , \quad \mathbf{D} = \langle d \rangle .
\]

On the macroscale mixed BCs could be used and hence the macroscopic BVP reads as follows:

\[
\operatorname{div} \mathbf{\Sigma}(\mathbf{x}) = 0 \quad \mathbf{x} \text{ on } \Omega \tag{23a}
\]

\[
\mathbf{E}(\mathbf{x}) = \frac{1}{2} \left( \nabla \mathbf{U}(\mathbf{x}) + \nabla^T \mathbf{U}(\mathbf{x}) \right) \quad \mathbf{x} \text{ on } \Omega \tag{23b}
\]

\[
\mathbf{U}(\mathbf{x}) = \mathbf{U}_0(\mathbf{x}) \quad \mathbf{x} \text{ on } \partial \Omega_D \tag{23c}
\]

\[
\mathbf{\Sigma}(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}) = \mathbf{T}_0(\mathbf{x}) \quad \mathbf{x} \text{ on } \partial \Omega_N . \tag{23d}
\]

The macroscopic BVP is solved using a standard Newton-Raphson scheme. Hence, in each macroscopic Newton iteration the averaged mesoscopic stresses $\langle \sigma \rangle$ (for setting up the macroscopic residual force vector) and additionally for an efficient macro computation, the macroscopic tangential stiffness $\mathbb{C}^M$ has to be calculated. The latter is calculated by applying six infinitesimal perturbation loadcases to the current equilibrated solution according to [19, 13].
4 NUMERICAL EXAMPLES

The mesoscopic BVP stated in section 2 was implemented in the FFT code FeelMath developed at the Fraunhofer ITWM. In some numerical examples the mesomechanical damage behavior and the scale coupling technique are illustrated.

4.1 Mesoscale Simulation of a Fiber Composite

The following example demonstrates the behavior of the mesoscopic damage model using a short fiber reinforced composite which is discretized by 384,000 voxel cells. The behavior of the glass fibers is assumed to be linear elastic while for the polymer matrix the nonlocal damage model described in section 2.1 is used. The structure is loaded with periodic deformation BCs.

Figure 2: Damage evolution on the mesoscale demonstrated on the example of a short fiber reinforced composite.

As illustrated in figure 2 the damage starts to grow on the sharp fiber edges and increases along the fiber matrix interface. With increasing load, shear bands occur between the fibers which connect the separated damage zones around the fibers. This results finally in a continuously meso crack and consequently in a complete failure of the meso structure.
4.2 Detection of the Onset of Macroscopic Localization

As already mentioned, at a certain point during the load history the macroscopic strain starts to localize and the problem becomes ill-posed. This point, which can be regarded as the onset of macroscopic failure, is detected by an acoustic analysis of the current homogenized tangent stiffness tensor \([5, 17]\). Therefore the acoustic tensor \(A(n)\) is calculated by applying the normal vectors \(n\) for all possible angles \((\varphi, \theta)\) on the macro tangent [3]:

\[
A_{jk}(n) = C_{ijkl}^M n_i n_l ,
\]

\[
n = n(\varphi, \theta) ,
\]

\[
\|n\| = 1 ,
\]

where the definition of the acoustic tensor \(A\) is given in index notation. Localization occurs if the following criterion is fulfilled for a certain direction \(n\):

\[
\det (A(n)) = 0 .
\]

We demonstrate this analysis on a cubic RVE containing a single linear elastic fiber inclusion. The structure is discretized by \(75^3\) (421,875) voxel cells and loaded during 10 time steps with a uniaxial macroscopic strain perpendicular to the fiber direction. In this example the nonlocal part of the damage model is suppressed to demonstrate the localization effect during the applied load history. In each time step we compute the homogenized macroscopic tangent and the acoustic tensor determinant for all directions.

![Figure 3: Detection of the onset of macroscopic failure with the acoustic tensor.](image)

In figure 3 the normalized determinant of the acoustic tensor is plotted using spherical coordinates. Thereby the value of the determinant illustrated as the spherical radius decreases with growing damage in the direction perpendicular to the crack direction.
4.3 Multiscale Simulation of a Fiber Composite

In figure 4 an example for a fully coupled multiscale approach is demonstrated. A strip with a hole discretized by 516 tetrahedral finite elements is coupled to a short fiber reinforced RVE structures which are modeled with $64^3$ (262144) voxel cells. The macroscopic linear elastic material behavior is replaced in some critical regions with a coupled multiscale approach. RVEs with 45° and 90° fiber orientation with respect to the macro loading direction are associated to particular macroscopic regions. The polymer matrix is modeled by the nonlocal damage model described in section 2.1 and the behavior of the fibers is assumed to be linearly elastic.

![Figure 4](image)

**Figure 4**: Fully coupled meso-macro computation of a strip with a hole and short fiber reinforced composite material structures.

Damage is captured on the finer scale which is illustrated in figure 5 for both fiber mesostructures. Growing damage on the fine scale with an increasing macroscopic load results in macroscopic reduction of stiffness. The maximal stresses move from the vicinity of the hole towards the center of the specimen. The meso- and macroscopic stress-strain response both show a nonlinear material behavior.

![Figure 5](image)

**Figure 5**: Stiffness reduction around the hole of the macroscopic specimen due to increasing damage on the mesoscale.
5 CONCLUSIONS

In this work we proposed an alternative multiscale approach similar to FE\(^2\) to investigate progressive damage behavior of composite materials. Nonlinear material effects caused by mesoscopic physical effects were modeled on a finer length scale and used to predict the macroscopic damage behavior.

In the first numerical example a meso scale problem was set up to demonstrate the behavior of the mesoscopic model. In this context we illustrated the mechanisms of growing damage which can be observed in short fiber reinforced composites.

Further we showed a simple way to predict the onset of macroscopic failure. We carried out an acoustic analysis of the current effective tangent stiffness tensor which was obtained by homogenization of a RVE around different macroscopic load states.

Finally, in the third numerical example a coupled multiscale simulation was performed. The elastic material behavior of a strip with a hole was replaced in some critical regions with a multiscale approach to model progressive damage on the coarse scale. Growing damage which occurred at the fine scale between the material constituents resulted in macroscopic reduction of stiffness and redistribution of macroscopic stresses.

REFERENCES


COUPLED THERMOELASTIC SIMULATION OF
NANOVOID CAVITATION BY DISLOCATION EMISSION
AT FINITE TEMPERATURE

M. PONGA†, M. ORTIZ‡ AND M.P. ARIZA*

*Escuela Técnica Superior de Ingeniería, Universidad de Sevilla
Camino de los descubrimientos, s.n. 41092-Sevilla, Spain
E-mail: mpariza@us.es, http://personal.us.es/mpariza/

† Mechanical and Civil Engineering, California Institute of Technology,
1200 E. California Blvd. Pasadena, 91125 CA, USA.
E-mail: mponga@caltech.edu

‡ Graduate Aeronautical Laboratories, California Institute of Technology,
1200 E. California Blvd. Pasadena, 91125 CA, USA.
E-mail: ortiz@aero.caltech.edu

Key words: Multiscale modeling, Crystal plasticity, Thermomechanical coupling, Nanovoids, Dislocations

Abstract. In this work we study the early onset of void growth by dislocation emission at finite temperature in single crystal of copper under uniaxial loading conditions using the HotQC method. The results provide a detailed characterization of the cavitation mechanism, including the geometry of the emitted dislocations, the dislocation reaction paths and attendant macroscopic quantities of interest such as the cavitation pressure. In addition, this work shows that as prismatic dislocation loops grow and move away from the void, the material surrounded by these loops is pushed away from the void surface, giving rise to a flux of material together with a heat flux through the crystal.

1 INTRODUCTION

Crystal defects play a critical role in determining macroscopic properties of solids, even when they are presented in small concentrations [1, 2]. Defects such as vacancies and nanovoids, modify the perfection of the crystal lattice and produce changes in different scales of length and time. The void nucleation mechanism is usually followed by plastic cavitation when the void attains the critical size for dislocation emission. Furthermore, the dissipation of this plastic work generates thermal fields around the nanovoid that fosters localization of deformation and mass transport by self-diffusion through dislocation cores
This coupling between mass transport and heat flux occurs on multiple time and length scales. As a result, the study of this type of problems must reproduce the coupled nature of the thermoelastic response of the material as well as their multiscale nature.

One multiscale approach which enables us to simulate thermomechanical coupled problems such as void cavitation and crack tip opening, among others, is the Quasicontinuum method (QC). This method was proposed fifteen years ago [4] as a multiscale modeling approach for the solution of mechanical problems where atomistic resolution is required in a localized subset of the domain under analysis. Initially, the QC was applied to the solution of mechanical problems in solids with defects at 0K. More recently, the method has been extended to encompass non-equilibrium coupled thermomechanical problems [5, 3], enabling the full analysis of continuum/atomistic domains at finite temperature and away from equilibrium. This extension is called HotQC method and has been applied to different plasticity problems such as nanoindentation [5] and nanovoid growth under triaxial load [3]. The development of the HotQC method for non-homogeneous temperature distributions relies on three main considerations. Firstly, the instantaneous vibrations of the atoms are eliminated through the use of the Jayne’s principle of maximum entropy [6], by performing phase averages and formulating a mean field free energy in terms of the macroscopic atomic variables, i.e., position, temperature and frequency. Secondly, a variational thermoelastic formulation proposed by Yang et al. [7] is required in order to describe equilibrium and non-equilibrium processes. Third, a coarse-graining model of the material allows bridging the continuum and atomistic scales. The method starts with a small and complete atomistic system around a core defect. One of the main advantages of the HotQC methodology, is that we can simulate systems in different thermodynamic states, varying from isothermal to adiabatic states. In this work, we study the evolution of void growth in copper single crystals at finite temperature under uniaxial load using the HotQC method.

2 SIMULATION DETAILS

The simulations performed in this work were carried out using the following set up. The representative volume $\Omega_{RV}$ is a cubic box of dimensions $(72a_0)^3$ of Cu simulated using the EAM-type potential proposed by Mishin et al. [8]. Since the publication of our preliminary results [9, 10] where we used a different EAM potential, we have observed that the results present a high dependency on the interatomic potential used to perform the simulation. Even though this discussion is beyond the scope of this paper, we would like to mention that it is mostly due to the stacking fault energies predicted by each interatomic potential. In the center of the volume a full atomistic zone, which is a cube of dimensions $(14a_0)^3$, is defined. A spherical void of $12a_0$ diameter is modeled in the center of $\Omega_{RV}$, by removing atoms from the initial atomistic zone. Away from the full representative zone we apply systematically coarsening of the sample $\Omega_{RV}$ using a set of representative atoms or nodes. Therefore, a Finite Element (FE) mesh is constructed using a Delaunay triangulation over the nodes. Previously to loading process the sample
is allowed to relaxed at an initial temperature $T_0 = 300K$, this is performed by the minimization of the mean field free energy with respect to nodal positions and frequencies. For the thermal expansion, the Boundary Conditions (BC) applied to the sample consist on fixing the normal displacement of the atoms in planes $x = 0$, $y = 0$ and $z = 0$. This BC ensures isothermal expansion of the crystal without distortion.

After the thermal expansion, the computational cell is deformed applying a homogeneous deformation gradient to simulate a uniaxial load in [001] direction, where the deformation increment is set as 0.1%. Hence, we impose dilatational displacements on the external boundaries, i.e., at $z = \pm 36a_0$ and find the equilibrium positions of the remaining ones by minimizing the mean field free energy. In every deformation step, an equilibrium configuration is found using a nonlinear conjugate gradient \cite{11}, or alternatively dynamic relaxation method is used \cite{12, 13, 14} when the computational minimization process during a certain loading step becomes very slow. In addition, the temperature is allowed to change between steps of deformation and its equilibrium value for each atom is reached by using a nonlinear conjugate gradient algorithm. Finally, in order to capture all the emitted dislocations surrounding the void, we implement a routine that automatically remeshes the sample using the second invariant of the deviatoric part of the Lagrangian strain tensor as adaptivity indicator. In this work, we focus on the detailed analysis of dislocations patterns emitted from the void surface at high strain rate load, $\dot{\epsilon} = 10^{10}s^{-1}$, typical of molecular dynamics calculations, and the evolution of the atomic temperature during the process.

3 RESULTS

In a previous publication \cite{3}, we have estimated the range of strain rates for which microinertia is expected to play a significant role during void growth. Our findings were consistent with foregoing continuum analysis and suggested that inertia is negligible for small enough voids. Based on this, we restrict attention for the present to quasistatic HotQC calculations.

Figure 1 shows the evolution of the virial stress vs strain, as well as the void fraction ($V_{\text{void}}/V_{\text{sample}}$) for an initial temperature of 300K and strain rate $\dot{\epsilon} = 10^{10}s^{-1}$. We distinguish four main stages in the virial stress evolution: i) an initial elastic regimen up to the cavitation of the void at $\epsilon = 6.5\%$, characterized by the elastic expansion of the void without the emission of dislocations, ii) a second and third plastic stages where dislocations are emitted from the void surface and which extend up to $\epsilon = 9.0\%$ and $\epsilon = 13.0\%$, respectively, and finally, iii) a fourth stage characterized by loss of stiffness up to ductile failure of the crystal. The deformation mechanisms which dominates the void growth under uniaxial load are revealed in details in the following section.
Figure 1: Virial stress vs strain and void fraction for single crystal copper predicted by HotQC using Mishin potential for a simulation sample of size $72a_0^3$ and void of radius 2.2 nm.

Figure 2: Dislocations emission from the void surface for uniaxial loading case. (a) Shockley partial dislocations with Burgers vector $1/6\langle 211 \rangle$ emitted from the void at $\varepsilon = 6.8\%$. (b) Trailing dislocations of the first ones are emitted at $\varepsilon = 7.0\%$. 

(a) Shockley partial dislocations at $\varepsilon = 6.8\%$. (b) Trailing edge dislocations at $\varepsilon = 7.0\%$. 

Mishin $10^{10}$ s$^{-1}$

V. F. Mishin
Dislocations Emission

The virial stress vs strain curve (Figure 1) presents a linear behaviour up to $\epsilon = 6.5\%$. Throughout this first stage, the void is mostly deformed along the loading direction and no significant transversal deformation is observed. However, as the strain attains the value of 6.6%, the critical stress for cavitation is reached, and therefore, we start observing dislocation emission from the equator of the void ($z = 0$). This first set of eight stacking fault dislocations grow on planes \{111\} with Burgers vector $1/6[11\bar{2}]$. In this sense, due to its shorter Burgers vector, Shockley partial dislocations are energetically more favourable than other type of dislocations [15, 16]. The Burgers vectors of the leading Shockley partial dislocations emitted are

Table 1: Leading dislocations emitted from void surface

<table>
<thead>
<tr>
<th></th>
<th>z positive axis</th>
<th>z negative axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_1$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
</tr>
<tr>
<td>$Z_2$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
</tr>
<tr>
<td>$Z_4$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
<td>$1/6[1\bar{1}\bar{2}]$</td>
</tr>
</tbody>
</table>

Figure 3: Lomer-Cottrell dislocations emitted from the void surface at $\epsilon = 8.5\%$. ($T_0 = 300K$ and $\dot{\epsilon} = 10^{10}s^{-1}$).
Figure 2a shows the set of leading dislocations emitted at $\varepsilon = 6.8\%$. As the energy of the Shockley partial dislocations is lower than any other type of dislocation, the stacking fault plane expands outward easily. In addition, during this plastic growth regime, a second set of Shockley partial dislocations is emitted from the void surface, which are the trailing edge dislocation of the first set, as we can see in Figure 2b. However, the shear stress in the crystal decreases proportionally to $1/r^2$ from the void surface [17] and consequently the driving force needed to move the set of stacking fault planes also decreases. Hence, these stacking fault planes decelerate giving rise to different dislocation structures. Therefore, additional shear loops with $1/6(211)$ Burgers vector are emitted from the void along the perpendicular direction to the loading direction. These shear loops join at the intersection of planes $x = \pm z$ with $x_0 = \pm R_0$ and $y = \pm z$ with $y_0 = \pm R_0$ to form a Lomer-Cottrell dislocation at $\varepsilon = 8.5\%$. Figure 3 shows the atoms comprising the void surface together with the Lomer-Cottrell dislocations that remain attached to its surface. For instance, dislocation reactions on the $x = R_0$ plane might be summarized as

\begin{align*}
X1 & : \frac{1}{6}[12\overline{1}] + \frac{1}{6}[12\overline{1}] \rightarrow \frac{1}{3}[10\overline{1}] \\
X2 & : \frac{1}{6}[12\overline{1}] + \frac{1}{6}[12\overline{1}] \rightarrow \frac{1}{3}[10\overline{1}] \\
X3 & : \frac{1}{6}[12\overline{1}] + \frac{1}{6}[12\overline{1}] \rightarrow \frac{1}{3}[10\overline{1}] \\
X4 & : \frac{1}{6}[12\overline{1}] + \frac{1}{6}[12\overline{1}] \rightarrow \frac{1}{3}[10\overline{1}] \\
\end{align*}

The Lomer-Cottrell dislocations form a rhombus shape and constitute the starting point for the emission of four Prismatic Dislocation Loops (PDL) along directions $\langle 110 \rangle$. The mechanism underlying the emission of the PDLs along the $\langle 110 \rangle$ directions is as follows: first, two shear loops are emitted form the void surface on two different planes, such as $(\overline{1}1\overline{1})$ and $(1\overline{1}1)$. Following these two shear loops, another two are also emitted on the same planes to form the rhomboidal shape which characterizes the PDLs (see Figure 4b). The complete reactions for the shear loops into prismatic ones are

\begin{align*}
1/6[2\overline{1}1] + 1/6[\overline{2}1\overline{1}] + 1/6[\overline{1}2\overline{1}] + 1/6[\overline{1}2\overline{1}] & \rightarrow \overline{1}\langle 10 \rangle \\
1/6[2\overline{1}1] + 1/6[\overline{2}1\overline{1}] + 1/6[\overline{1}2\overline{1}] + 1/6[\overline{1}2\overline{1}] & \rightarrow \overline{1}\langle 10 \rangle \\
1/6[\overline{1}2\overline{1}] + 1/6[\overline{1}2\overline{1}] + 1/6[\overline{1}2\overline{1}] + 1/6[\overline{1}2\overline{1}] & \rightarrow [1\overline{1}0] \\
1/6[211] + 1/6[12\overline{1}] + 1/6[21\overline{1}] + 1/6[12\overline{1}] & \rightarrow [1\overline{1}0] \\
\end{align*}

Figure 4 shows the complete sequence of dislocation patterns, including the emission of stacking faults and PDLs at different values of deformation. Finally, when the strain reaches $\varepsilon = 13\%$, the crystal gradually loss stiffness ending with the ductile failure at about $\varepsilon = 15\%$. 

1218
Figure 4: Dislocations emission from the void surface as the deformation increases. The snapshots are taken for 8.0% to 13.0% at steps of 1.0% of deformation. The load is applied in the vertical direction. All figures have the same scale. ($T_0 = 300K$ and $\dot{\varepsilon} = 10^{10}s^{-1}$).
Temperature Evolution

Figure 5a shows the time evolution of the temperature at three different points within the crystal: one point on the surface of the void, a second point near to the void and a third point far from the void. Before attaining the critical strain for dislocation emission $\epsilon = 6.5\%$, the crystal cools down due to the thermoelastic effect as expected for large strain rates. This variation in the temperature is clearly linear in consequence with the linear deformation of the crystal. However, when the deformation within the sample reaches the critical value for dislocations emission, the temperature in the vicinity of the void surface increases. The highest temperature is reached at the void surface and the increase is about $20K$. Figure 5b shows the temperature field of plane $z = 0$ for a deformation of $\epsilon = 7.0\%$. We clearly observe that the temperature field after dislocations emission is inhomogeneous and varies notably near to the void surface. Figure 5b also shows other important behaviour in the evolution of the atomic temperatures, showing how the atomic temperature increases when one PDL is passing through the atom labelled as ”Near to the void” (located around $5a_0$ from void surface in [110] direction). The increment of the atomic temperature due to dislocations emission indicates that as the PDLs move away the material surrounded by the dislocation loops is pushed away from the void surface, giving rise to a flux of material together with a heat flux through the crystal. Finally, when deformation reaches about 12.0% the temperature homogenizes within the crystal and continues cooling down up to the end of the simulation.

Void Growth

Figure 6a shows the initial configuration of the void at $T_0 = 300K$. At this point, the void is a discrete sphere of radius $r_0 = 6.0a_0$. When the sample is deformed the void starts growing elastically in the loading direction up to the first yield point at $\epsilon = 6.5\%$ and no transversal growth is observed, as we can see in Figure 6b. After this point, the critical strain for dislocations emission is reached and dislocations arise from the void surface. A detailed description of the evolution of the void growth as a function of the applied strain helps us to understand the mechanism of deformation underlying this loading condition. In fact, Figure 6c shows that the void grows in the direction perpendicular to the applied load in order to relax the internal energy of the crystal which causes the emission of dislocations. This mechanism has been predicted before based on continuum theory of spherical void growth and is known as prolate-to-oblate transition (see [18] and references therein). Finally, as deformation increases, the atoms comprising the surface of the void change their position and the void adopts an octahedral shape (see Figure 6f). This effect has also been observed in MD simulations performed by [19] where they attribute the octahedral shape to several factor such as: the low surface energy of the $\{111\}$ surfaces common in fcc metals, the high anisotropy of the copper elastic constants and the $\{111\}$ dislocation glide systems.
Figure 5: (a) Temperature evolution at different position within the crystal: void surface, near to the void surface and far from the void surface. Void fraction is also plotted. (b) Temperature field for $\epsilon = 7.0\%$ along the plane $z = 0$. Results obtained for $T_0 = 300K$ and $\dot{\epsilon} = 10^{10} \text{s}^{-1}$. 
Figure 6: Void growth under uniaxial loading conditions. $(T_0 = 300K$ and $\dot{\epsilon} = 10^{10}s^{-1})$. All figures have the same scale.
4 CONCLUSIONS

We have applied the HotQC method [5] to the study of quasistatic void growth in copper single crystals at finite temperature under uniaxial expansion using the EAM-type potential proposed by Mishin et al. [8]. We find that, upon the attainment of a critical or cavitation strain of the order of $\epsilon = 6.5\%$, dislocations are abruptly and profusely emitted from the void surface and the rate of growth of the void increases rapidly in directions perpendicular to the loading direction. The simulation carried out in this work shows the emission of four PDLs in $\langle 110 \rangle$ directions. This effect is called prolate to ablate and has also been observed in MD simulations [19] as well as for void growth using continuum mechanics [18]. Regarding the temperature field within the sample, prior to cavitation, the crystal cools down due to the thermoelastic effect. By contrast, dislocations emission after cavitation causes a rapid local heating in the vicinity of the void, which in turn sets up a temperature gradient and results in the conduction of heat away from the void. This result indicates that the problem of void growth at finite temperature need to be performed using a coupled computational tool, such as HotQC method, which allows us the study of both thermodynamic and plastic evolution of the crystal.

5 Acknowledgements

We gratefully acknowledge the support of the Ministerio de Ciencia e Innovación of Spain (DPI2009-14305-C02-01) and the support of the Consejería de Innovación of Junta de Andalucía (P09-TEP-4493). Support for this study was also provided by the Department of Energy National Nuclear Security Administration under Award Number DE-FC52-08NA28613 through Caltech’s ASC/PSAAP Center for the Predictive Modeling and Simulation of High Energy Density Dynamic Response of Materials.

REFERENCES


MULTISCALE ANALYSIS OF PIEZOELECTRIC MATERIAL BY USING EBSD-MEASURED REALISTIC MODEL

HIROYUKI KURAMAE*, HIDETOSHI SAKAMOTO† AND YASUTOMO UETSUJI‡

*Department of Technology Management, Faculty of Engineering
Osaka Institute of Technology
5-16-1 Omiya, Asahi-ku, Osaka 535-8585, Japan
e-mail: kuramae@dim.oit.ac.jp, www.oit.ac.jp

† Department of Mechanical System Engineering, Graduate School of Science and Technology
Kumamoto University
2-39-1 Kurokami, Kumamoto 860-8555, Japan
email: sakamoto@mech.kumamoto-u.ac.jp, www.kumamoto-u.ac.jp

‡ Department of Mechanical Engineering, Faculty of Engineering
Osaka Institute of Technology
5-16-1 Omiya, Asahi-ku, Osaka 535-8585, Japan
email: uetsuji@med.oit.ac.jp, www.oit.ac.jp

Key words: Multiscale Analysis, Piezoelectric Material, RVE Modeling, ODF Analysis, Coupled Problem.

Abstract. Material properties of a polycrystal piezoelectric ceramic, a barium titanate BaTiO$_3$, were analyzed by the two-scale crystallographic homogenization method. Three-dimensional (3-D) micro-finite element (FE) model was constructed based on the electron backscatter diffraction (EBSD) measured crystal orientation distribution images. The images are piled up to a 3-D voxel data of crystal orientation distribution by repeating mechanically and chemically polishing, and EBSD measurement of the ceramic. We obtained 13 EBSD images of 128×100 pixels, which measurement interval was 0.635 μm in-plane and the average amount of polishing was 1.66 μm in thickness (normal) direction of specimen. Each voxel of EBSD was assigned into 8-node solid FE in-plane with maintaining resolution of EBSD measurement, and was divided into three FEs along thickness direction with same crystal orientation, because of improvement of aspect ratio of FE. The total number of FEs was 499,200 (=128×100×13×3) which corresponded to over two millions degrees of freedom. In order to realize a large-scale micro-analysis using EBSD-measured voxel FE model, the coupled problem of the piezoelectric material was solved by parallel conjugate gradient (CG) method combined with the block Gauss-Seidel (BGS) method. The coupled micro-FE equation to obtain characteristic function vectors was separated into two linear equations, such as the elastic deformation and electrostatic analyses, by employing the BGS method, and then the equations were solved by the parallel CG solver while substituting coupling terms each other. Therefore, nested iterative scheme was constructed on a PC cluster. In addition, the representative volume element (RVE) size was determined based on the orientation
distribution function analyses of EBSD voxel data. The least RVE size was 25,000μm³, which corresponded to include 150 crystal grains.

1 INTRODUCTION

Piezoelectric ceramics, which have been used in actuators or sensors as a component of various electric and mechanical devices, consist of many crystal grains at a microscopic scale. Since each grain produces strongly anisotropic mechanical and electrical behaviour, the macroscopic properties of polycrystalline piezoelectric ceramics have large dependence of the microscopic crystal morphology. The multiscale analysis based on the homogenization method is needed for the piezoelectric materials to evaluate the effective macroscopic material properties [1]. Parallel processing technique is also required to the multiscale analysis because of large-scale coupling problem.

This paper presents a multiscale piezoelectric analysis based on the crystallographic homogenization method. In the finite element (FE) equation for the conventional piezoelectric analysis, coefficient matrix is not positive definite and strongly ill-condition because of coupling terms between mechanical and electrical fields [2-4]. In this study, a parallel computing technique for piezoelectric FE analysis is newly developed based on the iterative partitioned coupling method with the parallel conjugate gradient (CG) solver.

In order to construct a polycrystal microstructure model for the multi-scale FE analysis, a three-dimensional (3-D) representative volume element (RVE) which as crystal orientation distribution of a barium titanate (BaTiO₃) ceramic is measured by using the scanning electron microscope (SEM) with electron back scattered diffraction (EBSD) apparatus. Accuracy of solution is discussed through some numerical experiments. In addition, The optimum RVE size is determined based on the orientation distribution function (ODF) analysis of the EBSD-measured FE model.

2 METHOD OF THE ANALYSIS

2.1 Homogenization method for piezoelectric problem

A 3-D polycrystalline macro-continuum is formed by periodic microscopic structures of a RVE as shown in Fig. 1. The region of the RVE is made up of an aggregate of well-defined crystal grains and it is very small compared with the dimension of the overall macro-continuum region , defined by a scale factor , which represents the reciprocal order of the repetition.

Figure 1: Macroscopic, microscopic and crystal structures for the multiscale finite element analysis based on the homogenization method
The piezoelectric elasticity constitutive equation for the Cauchy stress tensor $\sigma_{ij}$ and the electric displacement vector $D_i$ is expressed as follows:

$$\begin{align*}
\sigma_{ij} &= C^{E}_{ijkl}e_{kl} - e_{ikj}E_k \\
&= C^{E}_{ijkl}\frac{\partial u_k}{\partial x_l} + e_{ikj}\frac{\partial \phi}{\partial x_l} ,
\end{align*}$$

(1)

$$D_i = e_{i\ell}e_{\ell\ell} + \epsilon_i^E E_k$$

(2)

where the displacement vector $u_k$ and the electric potential $\phi$ are selected as unknown variables in the piezoelectric problems, $e_{ij}$, $E_k$ are the strain tensor and the electric field, and $C^{E}_{ijkl}, e_{ikj}, \epsilon_i^E$ are the elastic compliance constant tensor, the piezoelectric strain constant, the dielectric constant tensor, respectively. The virtual work principle equation for the piezoelectric material is expressed as:

$$\int_{\Omega^l} \left( C^{E}_{ijkl} \frac{\partial u_l}{\partial x_k} + e_{ikj} \frac{\partial \phi}{\partial x_l} \right) \frac{\partial \bar{u}_i}{\partial x_j} d\Omega = \int_{\Gamma_t} t_i \delta u_i d\Gamma ,$$

(3)

$$\int_{\Omega^l} \left( e_{i\ell} \frac{\partial u_\ell}{\partial x_k} - \epsilon_i^E \frac{\partial \phi}{\partial x_l} \right) \frac{\partial \bar{u}_i}{\partial x_j} d\Omega = \int_{\Gamma_t} \rho \delta \phi d\Gamma .$$

(4)

In the homogenization procedure, macroscopic material properties $C^{EH}_{ijmn}$, $\epsilon_{ip}^H$ and $\epsilon_{ip}^S$ are obtained from the characteristic function by volume averaging in the RVE as follows:

$$C^{EH}_{ijmn} = \frac{1}{|V|} \int_Y \left( \text{micro} C^{E}_{ijmn} + \text{micro} e_{ikj} \frac{\partial \phi}{\partial x_l} \right) dY ,$$

(5)

$$\epsilon_{ip}^H = \frac{1}{|V|} \int_Y \left( \text{micro} e_{i\ell} + \text{micro} e_{ikj} \frac{\partial R_p}{\partial y_k} \right) dY ,$$

(6)

$$\epsilon_{ip}^S = \frac{1}{|V|} \int_Y \left( \text{micro} \epsilon_i^S + \text{micro} e_{ikj} \frac{\partial \Phi_p}{\partial y_k} \right) dY .$$

(7)

We assume that the piezoelectric material may have averaged properties in a macro continuum body. The characteristic function vectors $\{ \Phi^m \}$, $\{ \Phi^S \}$ and $\{ R^p \}$ are obtained by solving the microscopic FE equations for the RVE as follows:

$$\begin{bmatrix} K_{uu} & K_{ub} \\ K_{bu} & -K_{bb} \end{bmatrix} \begin{bmatrix} \chi^m \\ \Phi^m \end{bmatrix} = \begin{bmatrix} \mu^m \\ \Phi^S \end{bmatrix} , \quad (m, n) = (1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3) ,$$

(8)

$$\begin{bmatrix} K_{uu} & K_{ud} \\ K_{du} & -K_{dd} \end{bmatrix} \begin{bmatrix} \Phi^p \\ R^p \end{bmatrix} = \begin{bmatrix} t^p \\ q^p \end{bmatrix} , \quad p = 1, 2, 3 ,$$

(9)

where the characteristic functions have 9 components. The right-hand side vectors $\{ t^m \}$, $\{ q^m \}$, $\{ t^p \}$ and $\{ q^p \}$ are the constant values caused by the microscopic inhomogeneous structure. The coefficient matrices in both the linear equations are symmetric because of $[K_{ud}] = [K_{du}]^T$. The size of the linear equation depends on the number of nodal points in the RVE FE model, and the degrees of freedom may be a large number for the EBSD-measured FE model.

Since the diagonal terms in the sub-matrix $[-K_{dd}]$ are negative, the coefficient matrix of both the linear equations (8) and (9) is not positive definite. Since difference of numerical
order between \([K_uu]\) and \([K_{\phi\phi}]\) is very large more than \(10^{24}\) in actual problem, the coefficient matrix is strongly ill-condition. Therefore, the iterative solver such as the CG method, which is well-suited for distributed memory type parallel computing environment such as a PC cluster, is inapplicable to the equations.

2.2 Iterative coupling method based on Block Gauss-Seidel method

In order to apply the CG method to solve the system equations, they are rewritten to a partitioned form based on the Block Gauss-Seidel (BGS) method as follow:

\[
\begin{bmatrix}
K_{uu} & 0 \\
-K_{u\phi} & K_{\phi\phi}
\end{bmatrix}
\begin{bmatrix}
{u}^{(k+1)} \\
{\phi}
\end{bmatrix}
= 
\begin{bmatrix}
t - 0 \\
-q - 0
\end{bmatrix}
+ \begin{bmatrix}
K_{u\phi} \\
K_{\phi\phi}
\end{bmatrix}
\begin{bmatrix}
{u}^{(k)} \\
{\phi}
\end{bmatrix}
\]

(10)

where unknown vector \({u}\) corresponds to \({\chi}^{mn}\) and \({\Phi}^p\), and \({\phi}\) corresponds to \({\phi}^{mn}\) and \({R}^p\), respectively. According the coupling term \([K_{u\phi}]\) is moved to right hand side, displacement \({u}\) and electrostatic potential \({\phi}\) vectors are partitioned to maintain positive definite of both the coefficient matrices as follows:

\[
[K_{uu}]{u}^{(k+1)} = \{t\} - [K_{u\phi}]{\phi}^{(k)},
\]

(11)

\[
[K_{\phi\phi}]{\phi}^{(k+1)} = -\{q\} + [K_{u\phi}]{u}^{(k+1)}.
\]

(12)

The parallel CG solver [5], which is parallelization on the inner product of the coefficient matrix and the descent direction vector by row block distribution of the coefficient matrix, is applied to each partitioned equation until unknown vectors \({u}\) and \({\phi}\) are converged with mutually substituting the coupling terms based on the BGS method as shown in Fig. 2. A nested iteration procedure of the BGS method including the CG iteration is constructed in this scheme.

In this study, a hierarchical process distribution technique is introduced to reduce amount of data for communication. The CPUs of PC cluster are logically divided into 9 processor groups, and allocated to the micro FE equations using the partitioning form. In each processor group, the parallel CG solver based on the block partitioning is applied to the linear equations. The parallel analysis code is implemented by using the Fortran 90 language and the message passing interface (MPI) library MPICH 1.2 [6].

![Figure 2: Flowchart of the iterative partitioned coupling procedure based on the block Gauss-Seidel (BGS) method](image)
3 MULTISCALE ANALYSIS USING THREE-DIMENSIONAL EBSD-MEASURED MODEL

3.1 Three-dimensional EBSD measurement of barium titanate

We obtained distribution of crystal orientation in a 3-D parallelepiped box region of a BaTiO$_3$ ceramic (Murata Manufacturing Co. Ltd.) by using the EBSD (Oxford Instruments plc, Link ISIS C.7272) implemented in the SEM (JEOL Datum Ltd., JSM-5410) [5]. The specimen was a circular disk of 15mm diameter and 1mm thickness, and it was electrically poled along the thickness direction. The observed surface was mechanically polished using 3µm diamond particles (Marumoto Struers Co., DP-Spray) with a polishing sheet (DP-Mol). And then, it was chemically polished at pH3.5 using colloidal particles (OP-A) with a polishing sheet (DP-Chem). The surface of the isolative specimen was coated with the electrical conductive and amorphous osmium layer to defend the electrification due to electron beam. We employed an osmium coater (Meiwaforsis Co., Neoc-ST). In addition, a silver paste (Fujikura Kasei Co. Ltd., type D-500) was applied to the surface except for the SEM-EBSD measurement region to leak the accumulated charge.

Figure 3 shows crystal morphology images by the EBSD measurement. The scanning interval in plane direction was set to 0.635µm that is nearly smaller than one over ten of the average grain size. The interval of the polishing was set to 1.66µm in thickness direction. We obtained 13 images by repeating both the EBSD measurement and the polishing, and a 3-D EBSD model by the piled-up images as shown in Fig. 4. Therefore, the size of each voxel in the 3-D model is 0.635×0.635×1.660µm$^3$. The average grain size was 6.71µm which was obtained by the intercept method.

![Figure 3: Crystal morphology images of a barium titanate (BaTiO$_3$) ceramic by the SEM-EBSD measurement (measurement area: 81.3×63.5µm$^2$, number of measurement points: 128×100 pixels, interval of measurement: 0.635µm)](image-url)
3.2 Three-dimensional modeling and results

A 3-D RVE FE model is constructed by using the EBSD measured images as shown in Fig. 4. The 13 layers of EBSD images are stacked into 3-D parallelepiped box. Using the 8-node isoparametric solid element, one crystal orientation image pixel is assigned to a FE in-plane, and each image layer is divided into three FEs in the thickness direction. The total finite elements are over 500 thousand in the 3D model as summarized in Table 1. It corresponds to over 2 million degrees of freedom in the micro finite equation.

In addition, two types of 2-dimensional (2-D) RVE FE models are also constructed to compare with the 3D model. One is the same in-plane size with the 3D model, which uses only the first layer of EBSD image, the other is employed for a large in-plane region, 2D-L model, which is constructed from four neighboring regions by EBSD measurements.

![Figure 4: Three-dimensional RVE-FE modeling](image)

**Table 1: Properties of RVE-FE models**

<table>
<thead>
<tr>
<th>Model size [µm]</th>
<th>Number of finite elements</th>
<th>Number of nodes</th>
<th>Degrees of freedom</th>
<th>Element size [µm]</th>
<th>Aspect ratio of element</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D model 63.5×81.3</td>
<td>100×128×1</td>
<td>12,800</td>
<td>20,058</td>
<td>0.635×0.635×0.635</td>
<td>1.00</td>
</tr>
<tr>
<td>2D-L model 90.8×90.8</td>
<td>143×143×1</td>
<td>20,449</td>
<td>41,472</td>
<td>0.635×0.635×0.635</td>
<td>1.00</td>
</tr>
<tr>
<td>3D model 63.5×81.3×21.6</td>
<td>100×128×39</td>
<td>521,160</td>
<td>2,084,640</td>
<td>0.635×0.635×0.353</td>
<td>0.87</td>
</tr>
</tbody>
</table>

![Figure 5: Comparison of accuracy of the homogenized material properties between 2D-L and 3D models](image)

Figure 5 shows the homogenized material properties by using the 3-D RVE model compared with the 2D-L model. Comparing with the 2D-L model, the material properties of relation with thickness direction, such as $s_{33}$, $s_{13}$ and $s_{23}$, are smaller than that of the 3D model.
as shown in Fig. 5. In this study, Jaffe’s experimental data [7] are used for single crystal material properties. In the 2-D analysis, since microscopic heterogeneity along thickness direction is excluded, material properties for in-plane are over estimated.

In order to investigate relation between the number of layers of 3-D RVE model and accuracy of the homogenized material properties, the employing number of EBSD image layers in 3-D model is changed. Figure 6 (a) shows relative error of homogenized material properties with the 3D model which has 13 EBSD image layers. The one layer case corresponds to 2D model. The errors are less than 5% in three layers or more sampling cases. Figure 6 (b) shows parallel analysis time using 36 cores of the AMD Opteron HE275 2.2GHz CPUs connected by the Giga-bit Ethernet (1Gbps) network. The parallel analysis time is linear by the number of FEs.

![Figure 6: Relationship between number of sampling layers and accuracy, and analysis time](image)

(a) Relationship between accuracy of homogenized material properties and number of sampling layers of the 3-D RVE-FE model

(b) Relationship between analysis time and number of elements (number of layers) of 3-D RVE-FE model

**Figure 6:** Relationship between number of sampling layers and accuracy, and analysis time

![Figure 7: Image sampling of EBSD-measured crystal orientation distribution (4 types of plane size)](image)

**Figure 7:** Image sampling of EBSD-measured crystal orientation distribution (4 types of plane size)
To investigate the optimum (minimum) RVE size, various sizes of RVE models were sampled from full-size EBSD model, in which 4 types of plane size as shown in Fig. 7, and 7 types of number of layers from center layer. The total 28 models are sampled as shown in Fig. 8. The one layer models as shown in Fig. 8 (a-1), (b-1), (c-1) and (d-1) are correspond to 2-D model, and 13 layers with 100×128 pixels as shown in Fig 8 (d-7) corredponds to the full-size 3D model.

Figure 9 shows relation between volume of models and the accuracy. The error bars in this figure means the smallest, the average and the largest errors of 17 components of the homogenized material properties. It is found that the accuracy is increased by the volume of RVE, and more than 25,000µm³ volume of RVE is needed for less than 5% error of solution. The 25,000µm³ volume of RVE contains 150 crystal grains due to average grain size 6.71µm.

Figure 8: Sampling RVE models from full-size EBSD model (4 types of plane size and 7 types of number of layers)

Figure 9: Relationship between accuracy of homogenized material properties and volume size of RVE model
4 RVE DETERMINATION BASED ON THE ODF ANALYSIS

To determine the optimum (minimum) RVE size, in this study, the orientation distribution function (ODF) analysis [8] is used. The ODF \( f(\varphi_1, \Phi, \varphi_2) \) indicates intensity of the crystal orientation distribution in the Euler angle \((\varphi_1, \Phi, \varphi_2)\) space as follow:

\[
f(\varphi_1, \Phi, \varphi_2) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{n=-l}^{l} C_{lm} T_{lm}(\varphi_1, \Phi, \varphi_2),
\]

where \( T_{lm}(\varphi_1, \Phi, \varphi_2) \) and \( C_{lm} \) are spherical harmonic function and expansion coefficient, respectively. In this study, the ODF is analyzed for the crystal orientation distribution in RVE model every 5° in the Euler angle \((\varphi_1, \Phi, \varphi_2)\). Figure 10 shows results of the ODF intensity distribution in the 3-D Euler angle space for various sizes of RVE model. In the full-size 3D model which has 13 layers with 100×128 pixels as shown in Fig. 10 (d-7), some preferred orientations are appeared, but in the small model, the preferred orientation is disappeared. The larger model, the similar ODF distribution as full-size model is obtained. Therefore, the ODF results may be similar trend toward as accuracy of multiscale finite element results.

![Figure 10: ODF analysis results of various size of RVE models](image)
In order to quantitively evaluate similarity of the ODF results between full-size model \( f_{\text{full}}(\varphi_1, \Phi, \varphi_2) \) and sampled RVE model \( f_{\text{sample}}(\varphi_1, \Phi, \varphi_2) \), error sum of square \( S \) of ODF intensity distribution in the Euler angle space every \( 5^\circ \) is calculated as follow:

\[
S = \frac{1}{N} \sum_{\varphi_1=0}^{90^\circ} \sum_{\Phi=0}^{90^\circ} \sum_{\varphi_2=0}^{90^\circ} \left[ f_{\text{full}}(\varphi_1, \Phi, \varphi_2) - f_{\text{sample}}(\varphi_1, \Phi, \varphi_2) \right]^2,
\]

where \( N \) is number of evaluation points in the Euler angle space \( N = 19 \times 19 \times 19 = 6,859 \).

Figure 11 (a) shows comparison between error of homogenized material properties, which is multiscale FE results, and error sum of square of the ODF analyses. Similar decrease curve lines are indicated in this figure. Figuer 11 (b) shows scatter diagram of error of homogenized properties and the error sum of squares of ODF. The coefficient of correlation is 0.85. It is reveal correlation between accuracy of FE result and ODF errors. Therefore, estimation of error of homogenized material properties based on ODF analysis is effective to determine RVE size.

\[ S = 16.3 \times (\text{Error} \%) \]

(a) Comparison between accuracy of multiscale FE results and error sum of squares of ODF  
(b) Scatter diagram for correlation analysis between accuracy of multiscale FE results and error sum of squares of ODF

**Figure 11**: Determination of RVE size by ODF analysis

5 CONCLUSIONS

3-D realistic RVE model based on the EBSD measurement images of crystal orientation distribution was constructed. Homogenized material properties by using several sizes of RVE models were obtained. Determination method of the optimum RVE size based on ODF analysis is proposed. In this time, over 25,000µm\(^3\) volume size is required which contains 150 crystal grains due to average grain size.

REFERENCES


RULE-BASED EXPERT SYSTEM APPLICATION TO OPTIMIZING OF MULTISCALE MODEL OF HOT FORGING AND HEAT TREATMENT OF TI-6AL-4V

PIOTR MACIOŁ *, ALFRED KRUMPHALS*, STANISŁAW JĘDRUSIK†, ANDRZEJ MACIOŁ† AND CHRISTOF SOMMITSCH*

* Graz University of Technology, Institute for Materials Science and Welding
Kopernikusgasse 24, A-8010 Graz, Austria
e-mail: piotrmaciol@tugraz.edu.at, web page: http://iws.tugraz.at/
† AGH University of Science and Technology,
al. A. Mickiewicza 30,30-059 Krakow, Poland
Email: amaciol@zarz.agh.edu.pl - web page: http://www.agh.edu.pl

Key words: Multiscale modeling, Knowledge Based Systems, Expert Systems, Model Adaptation

1 INTRODUCTION

Nowadays, multiscale methods are increasingly used in modeling of processes and designing of materials. However, there are two serious obstacles to wider application of such methods. One is a demand of computing power, which is present in almost all applications. Increasing of available computing power is not a solution. Exponential growth of computational complexity with increase of accuracy will always lead exceeding of available resources. Therefore, adaptation of multiscale models is important point of interest. The second obstacle is a difficulty of multiscale model design. Model is an abstraction of reality. Therefore, some assumptions have to be done – which phenomena are important and which do not. There is an additional point of interest in multiscale modeling – in which scale the particular phenomena can be modeled with a good balance of accuracy and computing power demand. Usually, there is no single good answer. Designing of a model can be seen as a kind of an optimization process, where the goal is a function of efficiency and reliability. The need of such process is pointed out in many publications (e. g. [1]), but in the most of cases it have to be done by a researcher himself.

1.1 Adaptation of multiscale models

The term “adaptation” in computer science refers to a process, in which an interactive
system (an adaptive system) adapts its behavior to individual users (use cases) basing on information acquired about its user(s) and its environment. In numerical modeling, an adaptation refers to modifying of model properties to improve its quality. Usually, numerical error or its estimation is assumed to be a criterion of an adaptation. However, the question is how to estimate error in a multiscale manner. There are many error indicators for single scale models (e.g. [2] and many others). In a multiscale analysis, there are three main families of error indicators [3]. They are: multiscale reduction error estimators/indicators introduced by Fish and coworkers [4], localized modeling error estimates developed in the context of goal-oriented adaptive modeling by Oden and coworkers (e.g. [5]) and physically based error indicators introduced by Gosh and coworkers [6]. The first two families are based on an a’posteriori error estimator. The disadvantage of those methods is that a mathematical description of micro and macro scales phenomena must be consistent. Moreover, its application in regions with high gradients is difficult or even not possible. The physically based indicators are more flexible and can use various criteria. However, also in this case it is assumed, that error estimation can be directly computed.

For all of mentioned methods, a mathematical description of models in different scales must be known and analytically transformed before computational codes would be developed. It makes them not applicable if a more generous behavior is expected. Moreover they are not applicable when error estimation cannot be computed with mathematical equations or numerical procedures. This happens in two cases. First, a microscale model can be not compatible with a macroscale one in mathematical sense, than it will not be possible to project an error. FEM and Cellular Automata (CA) based models could be an example. Such models are quite common in modeling of a phase transformation [7], an microstructure evolution [8], thixoforming [9] and many others. There are no error estimators for CA simulations, furthermore there are no methods of projecting of error between FEM and CA. In the second case, phenomena which occur in a real material are not taking into account during modeling. While the numerical error can be very small, an absolute error, understood as a difference between a simulated and a real value can be extremely large.

1.2 Prediction of possible phenomena

Typical multiscale models consist of two models in two different scales, together with up- and downscaling (projection) procedures. Such models are able to predict usually only one microscale phenomenon, like a cracking, microscale heterogeneity or a microstructure evolution in well-defined conditions. An adaptation of such models could lead to results, which are very accurate from the numerical point of view, but completely different from a real material behavior. This happens, when a mathematical or a numerical description is not appropriate (the same as for example use of Galerkin method to a convection dominated problem in a single scale modeling), but also when the researcher did not take into account all possible phenomena. It enforces close cooperation of a numerical modeling and material science experts. Nevertheless, neglecting of important phenomena is still one of main reasons of modeling failures. On the other hand it is obviously not possible to take into account all possible phenomena. Nowadays, a prediction of possible phenomena and its significance must
be done by researchers themself.

Reassuming: (a) classical multiscale models are applicable only in well-defined conditions, (b) existing multiscale error indicators are not able to predict the difference between a numerical solution and a real behavior and (c) it is expected to simplify the use of multiscale models. In this paper, a concept of an adaptation of a multiscale model with a rule-based expert system is presented. With this methodology, the coupling between a multiscale models design and modeled phenomenological models is weakened. Moreover, capturing of phenomena which are not numerically modeled is feasible.

1.3 Knowledge Based Systems in numerical modeling

Knowledge Based Systems (KBSs) are widely used in a technologies design, but in the most of applications they are used instead of a numerical modeling, not for supporting it. KBS and a numerical model can cooperate in two ways: a numerical model can be a source of knowledge for KBS or KBS can be used for aiding of a numerical model development and controlling. Rec et al. [10] shown possibilities of a coupling of a rule-based system Rebit with a Finite Element Method (FEM) simulation of a rod rolling. In this paper, Rebit controls execution of FEM simulations, running computations and utilizing computations results in an inference process.

KBSs are also used to aid of a models development. Applications of KBSs to support of a finite mesh generation are relatively numerous. Sangiovanni [11] shown fuzzy logic rules used for a Finite Mesh design. Dolšak [12] presented the Prolog based FEMDES expert system, also for a finite mesh generation. Abd El-Ghany and Farag [13] presented an expert system that provides an intelligent interface between a testing engineer and a FEM software. Pinfold and Chapman [14] described DART system, aiding a FEM analysis design of automotive body structures with a KBS. Li and Qiao [15] shown a hybrid expert system for a FEM simulation of fuselage frame of an aircraft structure, which integrates an expert system with neural networks. Bellenger et al. [16] presented the framework for balancing between costs and quality of FEM analysis. They pointed out, that in industrial context not only accuracy, but also time of computations is crucial. Leitold et al. [17] and Németh et al. [18] described a problem of a models simplification. They did not use a KBS, but shown algorithm is in fact a set of rules for an expert system. It should be noticed, that almost all papers cited above refer to FEM. What is meaningful, there are no papers describing KBSs application in a multiscale modeling.

2 RULE-BASED MULTISCALE MODELLING

2.1 Agile Multiscale Modeling Method (AM3)

The most of presently developed multiscale models are designed as ”single author-single use” codes. It causes wasting of work, when the same problems are repeatable solved by different researchers. Also maintaining and improving of existing models are difficult. One of the possible solutions is designing of a unified multiscale modeling framework. The authors developed such a framework, combining a possibility of KBS based adaptation with a
unification of communication between single scale models.

Due to generalization of communication interfaces, it is not possible to merge models in different scales basing on mathematical transformations. Still, material properties can be calculated with microscale results (upscaling) and initial/boundary conditions for microscale models can be calculated with macroscale models (downscaling).

AM3 based multiscale model architecture consist of a hierarchy of models in many (minimum 2, maximum $n$) scales. The minimal configuration is one coarse scale and two fine scale models. It is assumed, that a “fine scale model” is any computational or mathematical model, able to calculate material properties with given initial and boundary conditions. Those properties could be e.g. viscosity or a stress-strain curve.

Respectively, a “coarse scale model” is assumed to be a model, which needs material properties from an external fine scale model. It is important, that the coarse scale model does not communicate with the fine scale model directly, but via so called Adapter. The Adapter communicates with a KBS, in turn. A KBS decides, basing on supplied information, which fine scale model should be used to calculate material properties. The Adapter stores at least two, alternative fine scale models. They are able to calculate the same properties, taking into account different phenomena, with different accuracy and different computation power demands. From the coarse scale model point of view, it makes no difference which fine scale model had been used.

Fine scale models could be a self-sufficient one. It means that all properties, needed for calculations are known and dependent only on initial and boundary conditions. However a fine scale model can expect providing some properties from outside. In such case, a model is concurrently a fine and a coarse scale model (with different reference points). Due to such a structure it is possible to create a hierarchy of models. Its depth is limited only by computer power demand. Simultaneously, models do not have any knowledge about a hierarchy, they know only their own upscale and downscale interfaces.

A fine scale model could be a primitive (e.g. Newtonian viscosity), a mathematical formula (a function of one or more variables), a simple numerical model (e.g. an interpolation of tabularized data) or more complex numerical model (CA, FEM, Molecular Dynamic, etc.).

The core of this process is a KBS. It decide, which fine scale model should be used. A KBS can access all data from a coarse scale model and previous results from fine scale models. Knowledge, stored in a KBS should apply to two issues families, which are a justification of using of the particular fine scale model and estimated computational costs of it. It has to be remembered that more detailed models have more computing power demands. During reasoning, a fine scale model’s reference is found. It is given to Adapter, who “transparently” manages a connection between fine and coarse scale models.

2.2 KBS

There are two main decisions to make during model setting-up. First, which phenomena should be modeled in particular scales? It is obvious, that increase of the number of phenomena taken into account rapidly increase computational complexity. On the other hand, neglecting of important behavior can make model useless. Second, location of fine scale
models (in spatial and temporal sense) have to be chosen. Nowadays, defining of modeled phenomena is done by designer, basing on his knowledge of process and material behavior. Microscale models positioning is also done by designer, partially supported by automatic adaptation methods. Increase of the number of modeled phenomena has to be limited due to an available computing power.

Pattern of design is determined by previously chosen strategy of multiscale modeling. Methods which are described in literature are based on two solutions. In the first one, computations are performed on coarse model, then results are analyzed, and eventually computations are repeated by means of fine model. Such solution requires a great amount of time and provides no opportunity to predict final time of calculations. In the remaining solution, model designer tries to predict correct multiscale simulation’s configuration. Usually it leads to acceptable, but far from optimal solution and requires highly skilled designers.

One of possible ways to improve quality of multiscale models is employing of Knowledge-Based Systems (KBS). KBS can be used as an alternative to precomputations, which are usually very resources consuming.

2.3 Rebit – Knowledge Based reasoning system

Rule-based system Rebit was developed at Management Faculty of AGH University of Science and Technology. From the point of view of functional approach, the basic task of the system is conducting the conclusion process upon the knowledge base chosen by the user. This task is executed by the rules engine in cooperation with the Rebit Client, which can be a human interface or other computer program. Knowledge representation language used in Rebit solution is enough expressive, as well as fully decidability. It allows writing in user-friendly form knowledge of material engineers. Likewise, correctness of knowledge in semantic point of view is possible. Due to application of SOA (service-oriented architecture), Rebit could be easily adapted to cooperate with multiscale modeling software.

An important feature of Rebit, allowing it to be used in multiscale modeling, is the possibility of suspending and resuming the process of inference at any time. It is possible to interrupt this process while saving the current state. In any moment, it is possible to resume the inference process, without any influence on a process. It gives an opportunity to acquire new knowledge when relevant information is not available in databases and starting the process from the beginning is impossible or not recommended. It is particularly important in case of multiscale simulation. Numerical models usually need long time to be solved. Thanks to suspending of inference process, the same engine can be used for other process or even turned off, if numerical computations are expected to take long time to finish. This ability could be used in parallelization of multiscale model. It is also useful in design time when some premises have to be obtained through in-depth studies.

Another important feature of Rebit System is its architecture. It is a service oriented system. The inference engine works as a web service according to Simple Object Access Protocol (SOAP) specification. This service can be consumed by SOAP clients running on any operating system. Rebit System is equipped with two generic clients, i.e. desktop client (in MS windows environment) and thin client (in any web browser). “Generic” in this case
means the ability to manage the inference process on any Knowledge Base (KB) selected by
the user. For AM3 based simulations, dedicated clients can be developed working with pre-
defined KB in automatic way. One of advantages of dedicated clients is the possibility of
extending a standard rule-based functionality. At the integration stage some workflow
capabilities and additional flow controls or translations can be introduced.

Rebit System can be deployed as a fully distributed system, where each component resides
on a different node. More integrated option, where engine, client and KBs repository are
available as a completely autonomous application, is also possible.

The inference process in Rebit System (in both deployment scenarios) proceeds according
to the following scheme. In the first step client has to obtain a KB. Next step depends on the
type of inference process. In forward chaining mode a set of input variables has to be chosen
and assigned. In backward chaining it is to formulate a hypothesis. Last type of inference, i.e.
mixed inference, requires choosing one variable, called final, from the entire set of variables.
At this point the inference engine starts interacting with the user (AM3 framework in this
case). The aim of this stage is obtaining values for all variables needed to finish the inference
process. As it was mentioned above, this stage can be suspended for an indefinite time and
resumed afterwards. In the case of forward chaining the inference process ends up after firing
all possible rules with input variable in their premises. Backward chaining is finished when
we are able to reject or accept the hypothesis. The inference process in mixed mode ends up
when it finds a value for final variable.

3 THERMO-MECHANICAL TREATMENT OF TI-6AL-4V

Components of the titanium alloy Ti-6Al-4V are often produced using hot forming
processes. During those processes the influence of processing parameters on microstructure
evolution is of high importance due to the effect on the final mechanical properties. In general
the thermo-mechanical treatment is divided into several processing steps which are heating
and soaking of the raw material on forming temperature, the forming process itself as well as
the heat treatment afterwards.

The success of thermo-mechanical treatment relies on the control of microstructure
evolution. Two of the most important microstructural mechanisms in hot processing of
titanium alloys are grain coarsening and phase transformation. Mechanisms of phase
transformation occur during heating and cooling whereas the effect of grain coarsening is
dominant during soaking on forming temperature.

In addition during forming at high strain rates the adiabatic heating has to be taken into
account due to its impact on microstructure. Those temperature changes caused by fast
deformation can lead to a significant change in phase amounts and grain morphology [19].

4 NUMERICAL MODELLING

4.1 Macroscale model

The 2D simulation of thermo-mechanical treatment of forged components is done using the
finite element (FE) software DEFORM2D™. For simulation of metal forming input data like
material flow curves, friction coefficients as well as thermo-physical parameters, are required. In general those input data are verified by experiments. In the preparation of FE-simulation process parameters have to be defined were the forming temperature is of high importance beside die speed and stroke.

After simulation the results are shown in the form of strain and temperature distributions for the simulated work piece. In addition the curves for temperature, strain and strain rate for every time step in each element are available. Those data are used as input for further modeling of microstructure evolution.

In this work, a simple process with an axisymmetric hot deformation of a billet has been modeled. The temperature on the axis of symmetry is given as a boundary condition, as well as the heat exchange with environment on outer surfaces of the billet. Between 200 and 220 seconds, the billet is deformed (Figure 1).

![Figure 1: Temperatures [°C] and shapes of the billet on the beginning (a), after deformation (b) and at the end of the process (c).](image)

### 4.2 Cellular Automata model

For a microstructure modeling of Ti-6Al-4V during a thermo-mechanical treatment the 2D Cellular Automata model (CA) had been created [8]. The CA model is based on a probabilistic approach which is capable to describe the grain coarsening behavior. To describe the microstructure the CA-model is using square shaped cells with different cell states. Furthermore, Moore’s neighborhood is used for identification of cells on grain boundaries and phase boundaries. The grain boundary movement, which affects grain coarsening, is determined by transformation probabilities. The value of those probabilities depends on process and material parameters like temperature and chemical composition. Furthermore fixed values of cell size and time step width are required and have to be defined previously.

The initial microstructure for CA-simulation can be imported from either artificial created...
microstructure, real micrographs or EBSD-maps. The model output includes mean diameter, grain size distribution and virtually simulated microstructures which can be easily compared with experimental micrographs.

4.3 AM3 based multiscale model

From the point of view of the multiscale modeling, the model presented above is relatively simple. There is only downscaling, because the macroscale model is not dependent on the microstructure. However, this problem could be successfully used as an example of AM3.

Numerical models presented above were developed to be used in a “batch mode”. The coarse model, which is not dependent of fine scale results, was computed at first. Results (temperature changes in time) were analyzed by the researcher and then microscale models were set up. The microscale models configuration, including its number, locations and properties were chosen manually.

This model was redesigned with AM3. While in the original problem information are passed in only one direction, from macro to microscale, in AM3 based application possibility of both direction communication is assumed. The macroscale model is similar to the original one. The only one difference is removing of stress-strain curve from Deform2D model. For all integration points, stresses are computed inside AM3, basing on material state (strain, strain rate, temperature, time) given by Deform2D model.

Stress computations are delegated from AM3 to one of two fine scale models. The first one is an interpolation of tabularized stress data in dependency of strain, strain rate and temperature. The second model is based on the CA model, described above. Because a stress-strain curve is not dependent on microstructure, interpolation of tabularized data is also used.

4.4 Knowledge design

The most important part of the AM3 based model is knowledge. In this case, the most important criterion for model choosing is an existence of a similar model. We expect that results of two fine scale computations in similar conditions will give negligibly different results. Efficiency of a whole model could be then improved by neglecting non necessary fine scale models. Criteria of a similarity have to be chosen on the beginning. The simplest condition is based on the difference of the temperature $\Delta T$ between the existing models and a probed point. If $\Delta T$ is smaller than given threshold, then a fine scale model is not necessary. Moreover, there is the temperatures range, where particular microscale models can be used. The CA model, described in 4.2 can be applied when the temperature $T$ is between 800°C and 1000°C. When $T$ is lower, no microstructure model can be started, while when $T$ is higher, a high temperature microscale model (which is not present here) should be used. The CA model, when started, will be active, even if the T in this point will fall below lower the temperature threshold. Entire knowledge definition is shown in Figure 2. The knowledge design in this case is relatively simple and should be treated rather as a concept description than a real case solution.

AM3 framework expects from the KBS an answer which model should be used. Then a mixed reasoning is only one possible schema.
4.5 Results

The KBS was requested to find the proper fine scale model for each time step, for each finite element. The first CA model had been started in the first time step in the first element (because there were no other CA models). For the following time step it was only one CA model. In the next step, the second fine scale model was started (because existing model was not similar). The third and in this case the last CA model was started at time $t = 110$ s. Temperature curves and alpha phase mean grain sizes are shown in Figure 3. Temperature peeks between $t = 200$ s and $t = 220$ s are caused by plastic deformation work. Final microstructures, simulated by CA models in all three steps are shown in Figure 4. Exemplary reasoning session can be seen in Figure 5.

```plaintext
TYPES
Models = "CAMicrostructEvolution", "noMicrostructEvolution", "HighTempMicrostructEvolution", "null"

VARIABLES
Temperature: Double; DeltaTemp: Double; isSimilar: Boolean; Model: Models; PreviousModel: Models;
LowerTreshold: Double = 800; UpperTreshold: Double = 1000;

RULE R01
IF PreviousModel = "CAMicrostructEvolution" THEN Model = "CAMicrostructEvolution"

RULE R02
IF Temperature < LowerTreshold AND PreviousModel = "null" THEN Model = "noMicrostructEvolution"

RULE R03
IF Temperature >= LowerTreshold AND isSimilar = false THEN Model = "HighTempMicrostructEvolution"

RULE R04
IF Temperature >= LowerTreshold AND Temperature < UpperTreshold AND isSimilar = false AND PreviousModel = "null" THEN Model = "CAMicrostructEvolution"

RULE R05
IF Temperature >= LowerTreshold AND Temperature < UpperTreshold AND isSimilar = true AND PreviousModel = "null" THEN Model = "noMicrostructEvolution"

RULE R06
IF Temperature > UpperTreshold AND isSimilar = true THEN Model = "noMicrostructEvolution"

RULE R07
IF Temperature < LowerTreshold AND PreviousModel = "noMicrostructEvolution" THEN Model = "noMicrostructEvolution"

RULE R08
IF Temperature >= LowerTreshold AND Temperature < UpperTreshold AND isSimilar = true AND PreviousModel = "noMicrostructEvolution" THEN Model = "noMicrostructEvolution"

RULE R09
IF Temperature >= LowerTreshold AND Temperature < UpperTreshold AND isSimilar = false AND PreviousModel = "noMicrostructEvolution" THEN Model = "CAMicrostructEvolution"

RULE R10
IF Temperature >= LowerTreshold AND Temperature < UpperTreshold AND PreviousModel = HighTempMicrostructEvolution THEN Model = "CAMicrostructEvolution"

RULE R11
IF DeltaTemp < 30 THEN isSimilar = true

RULE R12
IF DeltaTemp >= 30 THEN isSimilar = false
```

Figure 2: Rules for choosing of the fine scale model for thermo-mechanical treatment of Ti-6Al-4V
5 CONCLUSIONS

The methodology presented in this paper allows the adaptation of the multiscale model with use of knowledge. Due to unification of interfaces and some algorithms it also simplifies development of new multiscale models. Thanks to that, the researcher can work on higher level of abstraction. However, it has to be remembered that increasing of an abstraction level almost always leads to decreasing of efficiency. Therefore, described methodology is not competitive with mathematically consistent solutions cited in Introduction. Dedicated solution with strong mathematical background will be probably more efficient and accurate than general one. However, in many cases generality, ability of predicting additional phenomena and ease of use can be more important than numerical efficiency.

![Graphs showing alpha grain size and temperature over time](image)

**Figure 3**: Alpha phase grains sizes and temperatures for CA simulations

![Images of simulated microstructures](image)

**Figure 4**: Simulated microstructures at time $t = 2000s$ for $a,b$ and $c$ CA simulations
Presented example is relatively simple. It can be improved with adding more fine scale models types, making similarity more precise (e.g. dependent on temperature history, not only actual value) etc.

Rules for additional phenomena detection can be added to KB. It is possible to have rules also for phenomena, which cannot be modeled (there is no suitable fine scale model). In such case, while of course effects of such phenomena cannot be taken into consideration, neglecting of them will be signalized by KBS.

```
start
final_variable_set [Model]
source_variable_found [PreviousModel]
source_variable_value_set [PreviousModel]; [“null”]
rule_found_by_premise [PreviousModel]; [R01]
rule_evaluation_result [R01]; false
rule_found_by_premise [PreviousModel]; [R02]
variable_value_needed [Temperature]
rule_found_by_conclusion [Temperature]; [null]
source_variable_value_set [Temperature]; [800]
rule_evaluation_result [R02]; false
rule_found_by_premise [Temperature]; [R07]
rule_evaluation_result [R07]; false
rule_found_by_premise [Temperature]; [R10]
rule_evaluation_result [R10]; false
rule_found_by_premise [Temperature]; [R04]
rule_found_by_premise [Temperature]; [R05]
variable_value_needed [isSimilar]
rule_found_by_conclusion [isSimilar]; [R11]
rule_found_by_conclusion [isSimilar]; [R12]
variable_value_needed [DeltaTemp]
rule_found_by_conclusion [DeltaTemp]; [null]
source_variable_value_set [DeltaTemp]; [2000]
rule_evaluation_result [R12]; true
conclusions_value_set [isSimilar]; [false]
final_variable_identified [Model]; false
rule_evaluation_result [R05]; false
rule_evaluation_result [R04]; true
conclusions_value_set [Model]; [“MicrostructureEvolution”]
final_variable_identified [Model]; true
end
```

**Figure 5:** A trace of exemplary reasoning

**ACKNOWLEDGMENTS**

This work is supported by the NCN, project No. N N508 593739

**REFERENCES**


SPATIAL AND TEMPORAL INSTABILITY OF SLIGHTLY CURVED PARTICLE-LADEN SHALLOW MIXING LAYERS

IRINA EGLITE*, AND ANDREI A. KOLYSHKIN†

* Department of Engineering Mathematics
Riga Technical University
Riga, Latvia, LV 1048
e-mail: irina.eglite@gmail.com

† Department of Engineering Mathematics
Riga Technical University
Riga, Latvia, LV 1048
e-mail: andrejs.koliskins@rbs.lv

Key words: Linear Stability, Weakly Nonlinear Theory, Amplitude Evolution Equation, Computational methods.

Abstract. In the present paper we present linear and weakly nonlinear models for the analysis of stability of particle-laden slightly curved shallow mixing layers. The corresponding linear stability problem is solved using spatial stability analysis. Growth rates of the most unstable mode are calculated for different values of the parameters of the problem. The accuracy of Gaster’s transformation away from the marginal stability curve is analyzed. Two weakly nonlinear methods are suggested in order to analyze the development of instability analytically above the threshold. One method uses parallel flow assumption. If a bed-friction number is slightly smaller than the critical value then it is shown that the evolution of the most unstable mode is governed by the complex Ginzburg-Landau equation. The second method assumes that the base flow is slightly changing downstream. Applying the WKB method we derive the first-order amplitude evolution equation for the amplitude.

1 INTRODUCTION

Shallow mixing layers are widespread in nature. Typical examples include flows in compound and composite channels and flows at river junctions. Three widely used methods of analysis of shallow flows include experimental investigation, numerical modelling and stability analysis [1]. Experimental analyses in [2]-[5] showed that (a) bottom friction suppresses the growth of perturbations and (b) shallow mixing layer grows at a smaller rate than a free shear layer. Several papers are devoted to linear stability analysis of shallow mixing layers [6]-[9]. It is shown in [8] that rigid-lid assumption can be used for stability analysis of shallow flows for small Froude numbers. The effect of Froude number of stability characteristics of shallow mixing layers is analysed in [9]. Theoretical calculations in [6]-[9] support experimental observations: bed friction stabilizes the flow and reduces the growth of a mixing layer width.
The effect of small curvature on the stability characteristics of a free mixing layer is analysed in [10] where it is shown that curvature has a destabilizing effect on an unstably curved mixing layer and stabilizing effect on a stably curved mixing layer. Linear stability of two-phase flows where a fluid contains solid particles is investigated in [11] under some simplifying assumptions. It is shown in [11] that a particle loading parameter has a stabilizing influence on the flow.

In the present paper linear stability analysis of slightly curved shallow mixing layers for the case where a fluid contains solid particles is performed. Two basic methods are usually used in practice for linear stability analysis: (a) spatial stability analysis and (b) temporal stability analysis. The second approach is more convenient from a computational point of view since the corresponding generalized eigenvalue problem is linear and can easily be solved by standard software packages. The first approach is more convenient for the purpose of comparison with experiments but requires more computational efforts since the problem is nonlinear with respect to unknown eigenvalues. Well-known Gaster’s transformation [12] is often used to simplify stability calculations. However, Gaster’s transformation is valid only in the vicinity of a marginal stability curve where the growth rates are small. In other regions of the parameter space the difference between spatial and temporal growth rates can be quite large as is illustrated in the paper.

Linear stability gives only conditions of instability but it cannot describe the evolution of the unstable mode above the threshold. Weakly nonlinear theories are used in such cases in order to analyse the development of instability analytically. Two such methods are briefly described in the paper. The first approach is based on a parallel flow assumption and can be applied for the case where the bed-friction number is slightly smaller than the critical value (see, for example, [13]). Using the method of multiple scales an amplitude evolution equation for the most unstable mode is derived. It is shown that for particle-laden slightly curved shallow mixing layers the amplitude equation is the complex Ginzburg-Landau equation. The coefficients of the equation are calculated explicitly in terms of integrals containing linear stability characteristics of the flow.

The second approach takes into account slow longitudinal variation of the base flow. The analysis is based on weakly non parallel WKBJ approximation [14]. A first-order amplitude evolution equation is derived. The solution of the amplitude equation is then used to obtain the first-order approximation to the perturbation field.

2 LINEAR STABILITY ANALYSIS

The two-dimensional shallow water equations under the rigid-lid assumption have the form

\[ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \]

\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \frac{\partial p}{\partial x} + \frac{c_f}{2h} u \sqrt{u^2 + v^2} = B(u^* - u), \]

\[ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \frac{\partial p}{\partial y} + \frac{c_f}{2h} v \sqrt{u^2 + v^2} + \frac{1}{R} u^2 = B(v^* - v), \]
where \( u \) and \( v \) are the depth-averaged velocity components in the \( x \) and \( y \) directions, respectively, \( u' \) and \( v' \) are the components of the particle velocities, \( B \) is the particle loading parameter \([11]\), \( R \) is the radius of curvature, \( p \) is the pressure, \( h \) is water depth and \( c_j \) is the friction coefficient. The following simplifying assumptions are used in order to derive (1)-(3): (a) rigid-lid assumption is used (in other words, free surface acts as a rigid lid so that water depth is assumed to be constant); (b) Chezy formula \([13]\) is used to model bottom friction; (c) curvature is assumed to be small \((1/R << 1)\); (d) the distribution of particles in a carrier fluid is assumed to be uniform; (e) no dynamic interaction between carrier fluid and particles is assumed. Assumption (a) is verified in \([8]\) where it is shown that from a linear stability point of view rigid-lid assumption works well for small Froude numbers. Assumptions (d) and (e) are discussed in \([11]\) where it is shown that these assumptions are reasonable for the case of large Stokes number of the flow.

Introducing the stream function by the relations

\[
\psi = \frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}
\]

and eliminating the pressure the following equation is obtained from (1)-(3):

\[
(\Delta \psi)_x + \psi_x (\Delta \psi)_y - \psi_y (\Delta \psi)_x + \frac{2}{R^2} \psi_y \psi_{xy} + \frac{c_j}{2h} \Delta \psi \sqrt{\psi_x^2 + \psi_y^2} + \frac{c_j}{2h} (\psi_y^2 \psi_{yy} + 2\psi_x \psi_y \psi_{xy} + \psi_x^2 \psi_{xx}) + B \Delta \psi = 0
\]

Consider a perturbed solution to (5) of the form

\[
\psi(x, y, t) = \psi_0(y) + \varepsilon \psi_1(x, y, t) + \varepsilon^2 \psi_2(x, y, t) + \varepsilon^3 \psi_3(x, y, t) + \ldots
\]

Here \( \psi_0(y) \) is the base flow solution and \( \psi_1(x, y, t) \) is a small unsteady perturbation. Using the method of normal modes we represent \( \psi_1(x, y, t) \) in the form

\[
\psi_1(x, y, t) = \phi(y) \exp[i(\alpha x - \beta t)],
\]

where both \( \alpha \) and \( \beta \) can be complex. However, two widely used approaches in the theory of linear stability are (a) temporal stability analysis and (b) spatial stability analysis. In case (a) the wave number \( \alpha \) is assumed to be real while \( \beta \), in general, is complex: \( \beta = \beta_r + i\beta_i \). For spatial stability analysis the usual assumptions are as follows: \( \beta = \beta_r \) is real while the parameter \( \alpha \) is complex: \( \alpha = \alpha_r + i\alpha_i \).

Substituting (6) and (7) into (5) and linearizing the resulting equation in the neighborhood of the base flow we obtain the following ordinary differential equation for the amplitude \( \phi(y) \) of the normal perturbation:

\[
\phi_{yy} (ca_0 - \beta - iSu_0 - iB) - Su_{0y} \phi_y + 2u_0 \alpha / R \phi_y + \phi (\alpha^2 \beta - \alpha^2 u_0 - ca_{0y} + i\alpha^2 u_0 S / 2 + i\alpha^2 B) = 0
\]

with the boundary conditions

\[
\phi(\pm \infty) = 0
\]

where \( S = c_j b / h \) is the bed-friction number \([6]\) and \( b \) is a characteristic length scale (for
mixing layers $b$ is usually the width of a mixing layer). Note that (9), (10) is an eigenvalue problem where for temporal stability problem $\beta$ is an eigenvalue (it is seen from (9), (10) that the problem is linear in $\beta$) and for spatial stability problem $\alpha$ is an eigenvalue (problem (9), (10) is nonlinear in $\alpha$).

3 NUMERICAL RESULTS FOR SPATIAL AND TEMPORAL INSTABILITY

Problem (9), (10) is solved numerically by means of a pseudospectral collocation method based on Chebyshev polynomials (the details of the method are given, for example, in [13]). The corresponding discretized linear generalized eigenvalue problem in $\beta$ (temporal stability analysis) can be solved using one of subroutines in IMSL package. The base flow is said to be linearly stable if all eigenvalues $\beta$ have negative imaginary parts and linearly unstable if at least one of $\beta_i$ is positive.

The following procedure is suggested to solve spatial stability problem. Assuming that both parameters $\alpha$ and $\beta$ are complex of the form $\alpha = \alpha_r + i\alpha_i,$ $\beta = \beta_r + i\beta_i$, for each set of the parameters $R, S, B, \alpha_r,$ and $\beta_r$ we find $\alpha_i$ such that $\beta_i = 0$ (using a bisection method). The condition of instability is $\alpha_i < 0$.

Calculated spatial growth rates for the case $1/R = 0.025$ are shown in Fig. 1 for the case of a base velocity profile of the form

$$u_0(y) = \frac{1}{2}(1 + \tanh(y)). \quad (11)$$

The bed-friction number is $S = 0.1$. The three curves in Fig. 1 correspond to the following values of the particle loading parameter $B$ (from top to bottom): 0, 0.01, and 0.02. As can be seen from the graph, the increase in $B$ leads to a more stable flow (the growth rates are getting smaller as the parameter $B$ increases).

$-\alpha_i$
Spatial growth rates for the case $1/R = 0.05$ are shown in Fig. 2. The other values of the parameters are the same as in Fig. 1. It is seen from the comparison of Figs. 1 and 2 that both parameters ($B$ and $1/R$) have a stabilizing influence on the base flow. The growth rates for the case $1/R = 0.05$ are smaller than for the case $1/R = 0.025$.

Following Gaster [12] we denote by $(T)$ and $(Sp)$ the solutions to (9), (10) corresponding to temporal and spatial problems, respectively. It is shown in [12] that near the marginal stability curve

$$
\alpha_{r}(T) = \alpha_{r}(Sp), \quad \beta_{r}(T) = \beta_{r}(Sp), \quad \alpha_{r}(Sp) = -\beta_{r}(T)/c(T),
$$

where $c(T) = \beta_{r}(T)/\alpha_{r}(T)$. It follows from the Gaster’s transformation that on the stability boundary either spatial or temporal stability analyses can be used since in this case $\alpha_{r}(Sp) = \beta_{r}(T) = 0$. If the objective of the analysis is to construct a marginal stability curve then it is recommended to use temporal stability analysis (which is a simpler method from a computational point of view than spatial stability analysis). However, the use of the Gaster’s transformation away from the marginal stability curve can result in relatively large errors. We have computed temporal and spatial growth rates for the case $S = 0.05$, $B = 0$ and $1/R = 0$. The relative percentage errors $\delta$ in using Gaster’s transformation are shown in the Table 1.
It is seen from Table 1 that errors in using Gaster’s transformation for the calculation of growth rates away from the marginal stability curve can be quite large.

4 WEAKLY NONLINEAR ANALYSIS

In this section we briefly describe applications of weakly nonlinear theory to the analysis of development of instability above the threshold when the base flow loses stability.

The first approach is based on a parallel flow assumption. Using the method of multiple scales with the “slow” variables \( \xi = \varepsilon (x - c_g t) \) and \( \tau = \varepsilon^2 t \), where \( c_g \) is the group velocity, we assume that the evolution of the most unstable mode (in a small neighborhood of the critical value of the parameter \( S \)) can be described by the formula

\[
\psi_1(x, y, t) = A(\xi, \tau) \phi(y) \exp[i(\alpha x - \beta t)],
\]

(13)

where \( A(\xi, \tau) \) is a slowly varying amplitude. Applying the method of multiple scales to (5), (6) and (13) and using solvability conditions at order two we obtain the group velocity \( c_g \).

Using the solvability condition at order three we obtain an amplitude evolution equation of the form

\[
\frac{\partial A}{\partial \tau} = \sigma A + \kappa \frac{\partial^2 A}{\partial \xi^2} - \mu A |A|,
\]

(14)

where \( \sigma, \kappa \) and \( \mu \) are complex coefficients which are calculated in terms on integrals containing the characteristics of linear stability problems. Equation (14) is known as the Ginzburg-Landau equation in the hydrodynamic stability literature. It is shown (see, for example, [15]) that it has a rich variety of solutions from deterministic to chaotic depending on the values of the coefficients. In fact, (14) is used in two ways in the literature: first, as a phenomenological equation (that is, it is assumed that a certain phenomenon can be modeled by (14) where the coefficients are usually determined from experimental data), and second, it can be derived (in some cases) from the equations of motion. We have shown that (14) is derived from (5).

The second approach uses a slow longitudinal variation of the base flow under the assumption that the wave length of the most unstable mode \( \lambda \) is much smaller than the typical length scale \( L \) of the longitudinal variation of the base flow. In this case a small parameter \( \varepsilon = \lambda / L \) is used to measure non-parallelism of the base flow. Using the WKB method we assume that a perturbation stream function can be represented in the form
where \( \phi(y, X) \) is the amplitude which is represented in the form
\[
\phi(y, X) = \phi_1(y, X) + \varepsilon \phi_2(y, X) + \ldots
\]
and \( \phi_1(y, X) = A(X) \Phi(y, X, \omega) \), where \( A(X) \) is an amplitude function and \( \Phi(y, X, \omega) \) is the eigenfunction of the linear stability problem. It is shown that the amplitude evolution equation for this case has the form
\[
M(X) \frac{dA}{dX} + N(X) A = 0,
\]
where the functions \( M(X) \) and \( N(X) \) are calculated in terms of integrals containing the characteristics of linear stability problem.

5 CONCLUSIONS

Linear and weakly nonlinear analyses of particle-laden slightly curved shallow mixing layers are presented in the paper. Spatial stability problem under the rigid-lid assumption is solved numerically by pseudospectral collocation method based on Chebyshev polynomials. Spatial growth rates are calculated for different values of the parameters of the problem. Two weakly nonlinear approaches leading to amplitude evolution equations for the most unstable mode are described. Experimental and/or numerical data are needed in order to assess the applicability of weakly nonlinear models to the analysis of instability of particle-laden shallow mixing layers.

REFERENCES


CONSOLIDATION PROBLEM SOLUTION WITH A COUPLED HYDRO-MECHANICAL FORMULATION CONSIDERING FLUID COMpressIBILITY

SYLVIA R. C. BRANT *, MANOEL P. CORDÃO NETO * AND IGOR F. GOMES †

* Programa de Pós-graduação em Geotecnia (PPG – UnB)
Departamento de Engenharia Civil e Ambiental/FT
Universidade de Brasília
Campus Universitário Darcy Ribeiro, 70910-900 – Brasilia, Brazil
e-mail: sylvia@unb.br; porfirio@unb.br - Web page: http://www.geotecnia.unb.br

† Departamento de Engenharia Civil – Centro de Tecnologia e Geociências
Universidade Federal de Pernambuco
Rua Acadêmico Hélio Ramos, s/n, 50740-530 – Recife-PE, Brasil
e-mail: gomes@ufpe.br - Web page: http://www.ufpe.br/ctg

Key words: Hydro-mechanical Coupling, Fully Coupled Approach, Fluid Compressibility, Solids Compressibility, Finite Element Method.

Abstract. There are two principles which may be referred to as essentials to describing soil and rock behavior. The mechanical behavior is associated to the law of conservation of linear momentum, allowing forces balance analysis and the hydraulic behavior is characterized by mass conservation. These phenomena are related: stress-strain state is affected by fluid pressures and vice-versa. Therewith, it is intuitive the understanding of the importance of coupled analyses, which are certainly a more precise manner of describing how mechanical and hydraulic behavior are connected. Given certain difficulties related to the modeling process, porous media numerical model representation is usually simplified. In certain cases, simplifications do not imply on losses in results and behavior prediction. However, some situations require more comprehensive approaches, with development of previously neglected conditions. The main objective of this paper is to present a formulation for fully coupled hydro-mechanical analyses considering fluid and solids compressibility. This formulation, implemented in Finite Element program ALLFINE [1,2,3], was tested for a one-dimensional consolidation case. A sensitivity analyses for the fluid compressibility parameter using modified Cam-clay constitutive model showed that this consideration affects fluid pressure responses significantly, with a delay in fluid pressure dissipation during consolidation process. The simulations results were compared to Terzaghi’s analytical solution for the one-dimensional consolidation problem. Also, the comparison of the simulation results to the analytical responses clearly shows the differences between using linear elastic and elastoplastic models. In simulations for different stress levels with the modified Cam-clay model, it is possible to capture a flow induction effect due to high stress levels.
1 INTRODUCTION

The study of geomechanics relates the concepts traditionally employed in geotechnics to reservoir engineering. Hence, it is possible to conduct numerical analyses in order to predict mechanical and hydraulic behaviour of porous media such as oil reservoirs.

Altogether numerical modelling of porous media requires complex relations between the basic variables of the problem. Therefore, coupled analyses should be performed in order to accurately reproduce the effects on porous medium.

This type of analysis has much computational cost and in order to make it more feasible, some simplifying hypotheses are occasionally made. However, some of those simplifications may interfere in model responses, with simulation results not correspondent to the observed behaviour of the studied case.

It is known that the solid matrix and the fluids within the porous medium do not have its compressibility taken into account in traditional models in geotechnics or related areas, such as reservoir geomechanics. This kind of consideration could enhance modelling, contributing to more accurate results in behaviour prediction.

In this paper, simulation results are presented for a one-dimensional consolidation test using modified Cam-clay model. The hydraulic behaviour of a one-meter high soil column is evaluated, with verification of the effect of permeability variation during incompressible and compressible fluid flow. Therewith, it was possible to study the combined effect of compressibility changes in the fluid and permeability of the medium for each type of fluid.

The numerical model here employed for the simulations is a fully coupled hydro-mechanical formulation proposed by Jesus [3], here briefly presented. This formulation was fully described by the author, followed by its implementation, validation and calibration in ALLFINE [1,2,3].

2 PROBLEM FORMULATION

The presented formulation was organized and fully described by Jesus [3]. Here, a brief summary of the equilibrium and the mass conservation equations of the model is shown.

2.1 Spatial solution of the equilibrium equation

The equilibrium equation for a porous medium can be described as:

\[ \frac{\partial \sigma_{ij}}{\partial x_j} + b_i = 0 \]  

(1)

where: \( \sigma_{ij} \) is the stress tensor, \( b_i \) is the body forces vector and \( x_j \) corresponds to the coordinate system.

This formulation is previewed for a fully saturated system. Hence, the stress-strain relation is based on the definition of the generalized effective stress with influence of solids compressibility on solid matrix stress state, defined as [4]:

..
where: $\{\sigma'\}$ is the effective stress vector, $\{\sigma\}$ is the total stress vector, $\alpha_b$ is the Biot parameter, $\{m\}^T = \{1 1 1 0 0 0\}$ for a 3D analysis and $p$ is the fluid pressure (for a saturated analysis).

The Biot parameter ($\alpha_b$) is defined by the following expression:

$$\alpha_b = 1 - \frac{\{m\}^T \left[D^p\right]\{m\}}{9k_s}$$

where: $[D^p]$ is the constitutive matrix and $k_s$ is the bulk modulus of solids.

The Principle of Virtual Work is employed to solve the equilibrium equation in space using FEM [5]. The matrixes that represent the solved equilibrium equation are assembled and may be written as:

$$[K]\{\dot{u}\} + [C]\{\dot{p}\} = \{\dot{F}\}$$

where:

$$[K] = \int_{\Omega} \left[B\right]^T [D] \left[B\right] d\Omega$$

$$[C] = \int_{\Omega} \left[B\right]^T \alpha_b \{m\} \left[N\right]^T d\Omega$$

$$\{\dot{F}\} = \int_{\Omega} \left[N\right]^T \{\dot{b}\} d\Omega + \int_{\Gamma} \left[N\right]^T \{\dot{\tau}\} d\Gamma$$

$[K]$ is the stiffness matrix, $\{\dot{u}\}$ is the nodal displacements rate vector, $[C]$ is the solids-fluid coupling, $\{\dot{p}\}$ is the nodal fluid pressure rate vector and $\{\dot{F}\}$ is the external forces rate vector.

2.2 Spatial solution of the mass conservation equation

The mass conservation equation considered in this paper involves both liquid and solid phases. In order to represent the mass conservation of the entire porous medium, volume of solids and fluid within a finite element should be balanced in a total element volume approach [4].

The mass conservation equation for the liquid phase is defined by the following expression:

$$\frac{\partial}{\partial t} \left( \theta \rho' \right) + \frac{\partial}{\partial \vec{x}_i} \left( \theta \rho' \dot{U}_i \right) = 0$$

where: $\theta$ is the volumetric water content, $\rho'$ is the fluid density and $\dot{U}_i$ is the real fluid
The real fluid velocity can be expressed as:

$$\dot{U}_i = \dot{u}_i + \frac{w^f_i}{\theta}$$  \hspace{1cm} (6)

where: $\dot{U}_i$ is the real fluid velocity vector, $\dot{u}_i$ is the velocity of the solids vector, $w^f_i$ is the fluid velocity due to percolation vector and $\theta$ is the volumetric fluid content of the medium.

The real fluid velocity assumption implies on taking into account the effects of porosity changes in the porous medium. The velocity of the solids $\dot{u}_i$ is related to the displacements of the medium.

The evolution law for liquid fluids density is expressed as [6,7]:

$$\frac{d \rho^f}{\rho^f} = \frac{1}{k_f} dp$$  \hspace{1cm} (7)

where: $\rho^f$ is the fluid density, $k_f$ is the bulk modulus of the fluid and $p$ is the fluid pressure.

The mass conservation equation for the solid phase may be presented as:

$$\frac{\partial}{\partial t} \left[ (1-\phi) \rho^s \right] + \frac{\partial}{\partial x_i} \left[ (1-\phi) \rho^s \dot{u}_i \right] = 0$$  \hspace{1cm} (8)

where: $\phi$ is the material porosity, $\rho^s$ is the solids density and $\dot{u}_i$ is the solids velocity vector.

The evolution law for solids density is expressed as [4]:

$$\frac{(1-\phi) D \rho^s}{\rho^s} = \left[ \frac{(\alpha_b - \phi) D p^R}{k_s} - (1-\alpha_b) \dot{u}_{i,i} \right]$$  \hspace{1cm} (9)

where: $\rho^s$ is the solids density, $k_s$ is the bulk modulus of rock crystals, $\alpha_b$ is the Biot parameter, $\phi$ is the porosity of the rock and $p^R$ is the pressure of the fluids (for saturated medium $p^R=p$).

The spatial solution for the mass conservation equation via FEM can be made with Galerkin method, a weighted residual method [5].

The matrices that represent the solved mass conservation equation are assembled and written as:

$$[M]\{\dot{p}\}+[L]\{\dot{u}\}+[R]\{p\}=[Q]$$  \hspace{1cm} (10)

where:
\[ [M] = \int \alpha \left[ N^r \right]^\alpha \left[ \frac{(\alpha_s - \phi)}{k_r} + \frac{\phi}{k_f} \right] d\Omega \]

\[ [L] = \int \alpha \{ m \}^\alpha [B] d\Omega \]

\[ [R] = \int \alpha \left[ N^s \right]^\alpha \left[ \left\{ \frac{\alpha}{k_r} \right\} \left\{ \frac{\phi}{w} \right\} \right] d\Omega \]

\[ \{ Q \} = \int [B^r] \frac{\mu}{\rho} \left\{ g \right\} (\nabla \cdot \mathbf{u}) d\Omega - \int [N^r] \overline{q} d\Gamma \]

\{M\} is the mass matrix, \{p\} is the nodal fluid pressure rate vector, \{L\} is the fluid-solids coupling matrix, \{u\} is the nodal displacements rate vector, \{R\} is the flow matrix, \{p\} is the nodal fluid pressure vector and \{Q\} is the external discharges vector.

2.3 Coupling of equations and time solution of the problem

The spatial solution of equilibrium and mass conservation equations may be assembled as a system of equations. This procedure is required for the coupling of the equations. Considering both Equations 4 and 10, we have:

\[
\begin{bmatrix}
0 & 0 \\
0 & [R]
\end{bmatrix} \begin{bmatrix}
\{u\} \\
\{p\}
\end{bmatrix} + \begin{bmatrix}
[K] & [C] \\
[L] & [M]
\end{bmatrix} \begin{bmatrix}
\{\dot{u}\} \\
\{\dot{p}\}
\end{bmatrix} = \begin{bmatrix}
\{\dot{F}\} \\
\{Q\}
\end{bmatrix}
\]

(11)

Therefore, the system of equations may be expressed by:

\[
[W] \{x\} + [Y] \{\dot{x}\} = \{Z\}
\]

(12)

Considering the studied phenomenon is transient, this system of equations should be solved in time. Finite Difference (FDM) was the chosen method for the time discretization of the system of equations, already solved in space. This choice was based on the ease of application of FDM for that matter. Evaluating Equation 12 in time stage \( t + \alpha \Delta t \), the solution of the equation system is:

\[
\Delta t \alpha [W]_{t+0} + [Y]_{t+0} \Delta t = \Delta t [Z]_{t+0} - \Delta t [W]_{t+0} \{x\}_t
\]

(13)

3 METHODOLOGY

3.1 Description of the problem

The studied case consists of a consolidation analysis for a one meter high soil column. This
soil sample is assumed to be laterally confined and fully saturated. Fluid flow is restricted in the laterals of the column and free at the top surface, as illustrated in Figure 1. Also, this soil column is discretized in a non-uniform mesh of 10 elements and 44 nodes. 3-D 8-noded elements are shown in Figure 2.

An uniform load of a 10000 kPa induces the consolidation process in the sample. The constitutive model employed for these simulations was the modified Cam-clay. Permeability is defined in this paper as function of porous medium void ratio. Herein, the simulations are performed for three different configurations of this function.

In order to evaluate the combined effect of compressibility and permeability changes, fluid bulk modulus value is also varied, with $k_f = 1 \times 10^{12}$ kPa for incompressible fluid and $k_f = 1 \times 10^5$ kPa for compressible fluids. The values of fluid bulk modulus and solids bulk modulus used in these simulations do not necessarily correspond to real values of these parameters. They have been established to facilitate the visualization of the effects of compressibility in the model results, as a sensitivity analysis of the evaluated parameters. It has been shown [3] that depending on the stress state the porous media is subjected to, it is possible to visualize the same effects here presented with real values of these parameters.

The evaluation of the simulation results is made with the values of fluid pressure, which are monitored over time during the consolidation process. The time factor here presented for each time stage is the same predicted by Terzaghi in its analytical solution for the one-dimensional consolidation problem.

3.2 Permeability function

The pore volume reduction during consolidation influences permeability. Therewith, these features are associated and a permeability function may be expressed in order to relate them.
In this study, a permeability function is suggested:

\[ k = A \frac{\exp(Be)}{\exp(BC)} \]  

where: \( k \) is the permeability and \( e \) is the void ratio.

Parameter A corresponds to the initial permeability \( (k_0) \) and parameter C is equivalent to a reference void ratio value, in this case, the initial void ratio \( (e_0) \) for the simulations.

Parameter B was used to calibrate the void ratio variation range of the function. With its variation, it was possible to define three different formats for the permeability function. The function reaches constant values of permeability (function \( k_1 \)), values 10 times lower than the initial (function \( k_2 \)) or values 100 times lower than the initial (function \( k_3 \)) for the final void ratio value. The values for each parameter are presented in Table 1 and its corresponding permeability functions are shown in Figure 3.

<table>
<thead>
<tr>
<th>A</th>
<th>1.0x10^{-6}</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>For function k₁, 0,0</td>
</tr>
<tr>
<td></td>
<td>For function k₂, 95,8</td>
</tr>
<tr>
<td></td>
<td>For function k₃, 191,9</td>
</tr>
<tr>
<td>C</td>
<td>0,90</td>
</tr>
</tbody>
</table>

![Figure 3. Permeability functions.](image)

### 4 RESULTS AND DISCUSSION

The performed analyses use the parameters shown in Table 2. The values of permeability vary according to the permeability functions already presented.

| \( M \) | 1 |
| \( p_0 \) | 5000 KPa |
| Overconsolidation ratio (OCR) | 0,20 |
| \( \lambda \) | 0,0030 |
The values of final void ratio for the modified Cam-clay model analyses are presented in Table 3. There were no differences between final void ratio values for incompressible and compressible fluids.

Table 3. Final void ratio values for modified Cam-clay model simulations.

<table>
<thead>
<tr>
<th>Permeability</th>
<th>Final void ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ function</td>
<td>0.876</td>
</tr>
<tr>
<td>$k_2$ function</td>
<td>0.877</td>
</tr>
<tr>
<td>$k_3$ function</td>
<td>0.877</td>
</tr>
</tbody>
</table>

As noticed, volume change is small, but the effects of permeability variation can be noticed in fluid pressure results due to the chosen permeability functions. The evolution of the fluid pressure during the simulation of the consolidation process is monitored for specific time factors ($T=0; 0.2; 0.5$ e 0.8). The results for incompressible fluid analysis are presented in Figures 4 to 7.
At the first time stage (T=0), values of fluid pressure are equal for all permeability functions simulations. Then, in the following time stages, fluid pressure dissipates over time faster for constant permeability ($k_1$ function) and slower for the function with which soil reaches a permeability value 100 times lower than the initial value ($k_3$ function).

It can be inferred from these results that fluid pressure values are influenced by permeability even for low void ratio changes. The straining the porous medium suffers is sufficient to make the permeability function influence noticeable.

Analyzing the shape of the curves, it can be noticed that for permeability function $k_3$, fluid pressure is high for the lower layers of the sample and it decreases abruptly for the superficial layers (surface at 1 m).

During the process, the superficial layers consolidate first, having their pores closed, forming a low-permeability zone. This causes fluid flow decrease, making the values of fluid pressure higher in other layers of the sample.

It was possible to reproduce this pore closing effect due to the use of modified Cam-clay model. This model adequately represented straining of the sample during the consolidation test. Thus, soil void ratio was appropriately calculated and updated during the simulations.

In Figure 8, two curves are plotted, one for void ratio and other for permeability. These curves complement the understanding of the pore closing process during consolidation when performing simulations with modified Cam-clay model. It can be observed how void ratio gradually decreases until it reaches initial stress (5000 kPa), when it starts to decrease.
abruptly. As expected, being the permeability a function of void ratio, it follows the same tendency.

Figure 8. Void ratio and permeability changes with stress increase.

One can deduce the need for an appropriate permeability function, with adequate parameters to represent the pore volume changes over time. Porous medium permeability influences directly fluid pressure and flow, representing an important aspect of numerical modeling.

The following simulation was performed for a compressible fluid. The analysis results are presented in Figures 9 to 12.

Figure 9. Results of fluid pressure for compressible fluid simulation (T=0).

Figure 10. Results of fluid pressure for compressible fluid simulation (T=0,2).

Figure 11. Results of fluid pressure for compressible fluid simulation (T=0,4).

Figure 12. Results of fluid pressure for compressible fluid simulation (T=0,6).

Figure 13. Results of fluid pressure for compressible fluid simulation (T=0,8).

Figure 14. Results of fluid pressure for compressible fluid simulation (T=1,0).
The results for simulations with the compressible fluid follow the same behavior tendency of the incompressible fluid. The compressibility of the fluid causes an effect of delay in fluid pressure dissipation, given by volume changes the fluid suffers. The effect of fluid compressibility in a consolidation process in porous media is detailed in Jesus [3].

It is fair to state that the only difference among the values of pressure for both types of fluids is related to its compressibility, with no influence from the permeability functions. Thus, by analyzing the format of the fluid pressure curves over time, it can be inferred that fluid flow tendency is the same (the shape of the curves is maintained) regardless the fluid compressibility. Soil consolidation induces pore closing first on the superficial layers of the sample, forming a low-permeability zone at its surface. Fluid pressure within this zone increases, as observed for the results of all time stages for permeability function $k_3$ curves.

5 CONCLUSIONS

In this paper, the effects of permeability variation in fluid pressure dissipation during a consolidation process were investigated. Pore volume changes influence directly fluid percolation through the porous medium, evidencing that the different permeability functions may alter significantly the responses achieved with the proposed numerical model.

In addition, the combined effect of permeability changes and fluid compressibility is analyzed. One can deduce that fluid compressibility does not alter percolation through the
porous medium. Fluid flow takes place with the same tendency observed in incompressible fluid simulations. However, an effect of fluid pressure dissipation delay is registered due to fluid volume changes.

Interesting results were achieved with modified Cam-clay model. This model captures more accurately the effect of pore closing during the consolidation process. The applied load triggered significant changes in void ratio for the superficial layers of the sample. Thereby, a low-permeability zone was formed at the surface of the sample and it prevented fluid pressure dissipation for lower layers (fluid flow out of the sample). This could be easily observed in the results for permeability function $k_3$.

These analyses highlighted the importance of studying permeability functions for porous medium, in order to better represent its hydraulic behaviour and how an interesting mechanical behaviour of porous medium could be captured with the use of modified Cam-clay model.

REFERENCES


NUMERICAL SOLUTION OF P-SV WAVES WITH FREE SURFACE BOUNDARY CONDITIONS BY NETWORK METHOD

JOSÉ L. MORALES*, IVÁN ALHAMA† AND MANUEL ALCARAZ‡

* Department of Structures and Construction
Universidad Politécnica de Cartagena
Campus Muralla del Mar, 30202 Cartagena, Spain
e-mail: joseluis.moraes@upct.es, www.upct.es/~deyc

† Network Simulation Research Group. Applied Physics Department
Universidad Politécnica de Cartagena
Campus Muralla del Mar, 30202 Cartagena, Spain
email: simulacion.redes@upct.es, fisica.upct.es

Key words: P-SV seismic waves, Network Simulation, Numerical Simulation.

Abstract. A numerical model based on network simulation method is presented for the numerical solution of P-SV waves in elastic media subjected to mixed boundary conditions: free surface and fixed displacements. As excitation, the standard form of Ricker pulse is used. After a detailed explanation of the model, it is run in suitable circuit simulation software. Boundary conditions are implemented by simple electrical components contained in the library of the software. An application is presented in order to show the efficiency of the proposed method. Tabulates result are processed by Matlab in order to present pictures of displacement for times along the transitory; displacements for sections of the medium, as a function of time, are also presented.

1 INTRODUCTION

Despite the electric analogy widely treated in many books for educational proposes (as an alternative representation of a problem in the context of a general analogy between equations), particularly in heat transfer problems [1-2], it goes far beyond the scope of this subject and can be used as a real numerical tool, making good use of the powerful mathematical algorithms implemented in the circuit simulation codes. Recently, it has been applied to the solution of well known problems in engineering: flow and transport problems [3], tribology [4], chemical corrosion [5], magnetohydrodynamic flows [1], inverse problems [6], heat transfer [6] and others; processes whose mathematical model is formed by a set of coupled, non-linear partial differential equations. Despite the disadvantage of having to be familiar with the basic principles of electric circuit theory, in all these problems, the method has demonstrated itself to be an accurate and reliable tool for many researchers.

In this context we apply this method to obtain the solution of the P-SV wave equation: the emergence and propagation of elastic waves in a plane domain from the point source.

The design of the model starts from the finite-difference differential equation resulting from the spatial discretization of the mathematical model, the 2-D wave equation [7],
José L. Morales, Iván Alhama and Manuel Alcaraz.

retaining time as a continuous variable. The terms of this equation are implemented as electrical currents by using a special kind of device, named ‘controlled source’. These currents, in turn, are balanced according to their signs at a common node, forcing to find the solution for the dependent variable. The controlled sources are capable of implementing any kind of nonlinearities or couplings between variables thanks to the possibility of specify their outputs by software, as a function of currents and voltages at other elements or nodes of the network, while lineal terms are directly implemented by simple devices (resistors, capacitors and coils). As regards derivate or higher order terms, auxiliary networks are required [8]. To complete the model, boundary conditions are also implemented by suitable components whose constitutive equation fits these requirements.

Once designed, the network is run in a standard code of circuit simulation such as PSpice [9]. The widespread use of this code and its new versions ORCAD attests to the applicability of the program to a large variety of circuit simulation problems and provides a valuable base of experience that demonstrates the advantages of the powerful, efficient and reliable numerical algorithms that are implemented therein. Instantaneous pictures of the perturbation along the domain are depicted by processing the tabulated output data with MATLAB.

To show the application of the proposed method we describe a scenario of 2-D, P-SV waves, in seismic problems subjected to a standard Ricker pulse [7], as excitation, applied to the center of the boundary side. Mixed boundary conditions are imposed: free surface at the upper side of the domain and fix displacement at the other sides.

2 MATHEMATICAL MODEL

The governing equations of P-SV waves, in absence of body forces, are [7]:

\[
\begin{align*}
\rho \frac{\partial^2 u_x}{\partial t^2} &= (\lambda + 2\mu) \frac{\partial^2 u_x}{\partial x^2} + \mu \frac{\partial^2 u_j}{\partial z^2} + (\lambda + \mu) \frac{\partial^2 u_z}{\partial x \partial z} \\
\rho \frac{\partial^2 u_z}{\partial t^2} &= \mu \frac{\partial^2 u_x}{\partial x^2} + (\lambda + 2\mu) \frac{\partial^2 u_j}{\partial z^2} + (\lambda + \mu) \frac{\partial^2 u_z}{\partial x \partial z}
\end{align*}
\]

(1a) (1b)

being \(u_x\) and \(u_z\) the displacement components, \(\lambda\) the Lamé’s constant, \(\mu\) the shear modulus and \(\rho\) the density. The relation between these parameters and elastic constants, \(E\) (Young’s modulus) and \(\nu\) (Poisson’s ratio), is:

\[
\begin{align*}
\mu &= \frac{E}{2(1+\nu)} \\
\lambda &= \frac{E\nu}{(1+\nu)(1-2\nu)}
\end{align*}
\]

(2)

For the completed mathematic model, the mixed boundary conditions are:

\[
\begin{align*}
u_j &= u_j^b \quad \text{on } S_u \\
\sigma_{ij} n_j &= t_i^b \quad \text{on } S_t
\end{align*}
\]

(3a) (3b)

where \(S_u\) denotes boundary surface points where the displacement values are prescribed, while \(S_t\) refers to points where the traction values are also given. Note that \(S = S_t + S_u\) represents the complete boundary surface, Figure 1.
Figure 1: Mixed boundary conditions in a 2D domain

For the P-SV waves described by Equation (1), the characteristic velocity are
\[ \alpha = \left[ \frac{\lambda + 2\mu}{\rho} \right]^{0.5} \] and \[ \beta = \left( \frac{\mu}{\rho} \right)^{0.5} \] for the P and S waves, respectively.

To apply the conditions (3b), in terms of the displacements, it is necessary to consider the Hooke’s law in term of displacements \[ \sigma_{ij} = \lambda \varepsilon_{ij} + \mu (\varepsilon_{ii} + \varepsilon_{jj}) \]. For 2D problems, the relation reduces to:

\[ \sigma_{xx} = (\lambda + 2\mu) \frac{\partial u_x}{\partial x} + \lambda \frac{\partial u_z}{\partial z} \]  \hspace{1cm} (4a)

\[ \sigma_{zz} = \lambda \frac{\partial u_x}{\partial x} + (\lambda + 2\mu) \frac{\partial u_z}{\partial z} \]  \hspace{1cm} (4b)

\[ \sigma_{xz} = \mu \left( \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \right) \]  \hspace{1cm} (4c)

Finally, tractions conditions are:

\[ t_x = \left[ (\lambda + 2\mu) \frac{\partial u_x}{\partial x} + \lambda \frac{\partial u_z}{\partial z} \right] n_x + \mu \left( \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \right) n_z \]  \hspace{1cm} (5a)

\[ t_z = \mu \left( \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \right) n_x + \left[ \frac{\lambda}{\partial z} \frac{\partial u_z}{\partial x} + (\lambda + 2\mu) \frac{\partial u_z}{\partial z} \right] n_z \]  \hspace{1cm} (5b)

As the excitation point source of the seismic waves, the classical Ricker pulse is considered, Figure 2. The amplitude of the Ricker’s pulse is defined by the function:

\[ A(t) = \left( a^2 - \frac{1}{2} \right) e^{-a^2} \]  \hspace{1cm} (6)

with \( a = \pi (t - t_p) / t_p \), being \( t_s \) and \( t_p \) the delay and width times, respectively.
3 NETWORK MODEL

Following the rules of the Network Method, for a suitable electrical analogy, the first step is to establish the analogy between the displacements of the elastic problem and the electric voltages of the network model. After that, partial differential of governing equation (1), as well as boundary conditions (5), are written in finite-differences form discretizing the spatial variables. Times remains as continuous variable, as in the lines method. For example, using the nomenclature of Figure 3, the second partial derivative for one of the displacement components reduces to:

\[
\frac{\partial^2 u}{\partial x^2} = \frac{u_{k,2} - u_{k,0} - u_{k,0} + u_{k,4}}{2 \Delta x} = \frac{2u_{k,2} - 2u_{k,0} + u_{k,4}}{\Delta x^2} = \left( \frac{u_{k,0} - u_{k,2}}{\Delta x^2} + \frac{u_{k,0} - u_{k,4}}{\Delta x^2} \right)
\]  

(7)

Figure 3: Nomenclature of a volume element

The resulting network model for each volume element that implements the governing equation is shown in Figure 4. Note that there are two similar circuits for each displacement components $u_x$ and $u_y$, respectively.
José L. Morales, Iván Alhama and Manuel Alcaraz.

Figure 4: Network model for a volume element related with one of the displacement components of the P-SV wave equation

The network model for displacement boundary condition is a simple voltage source with the value of the imposed displacement, Figure 5(a).

For traction boundary conditions, controlled voltage source (defined by software) must be used. These devices are capable to implement the coupled relations between both displacement components. For example, Equation (5) applied to the free boundary condition, \( t^b_x = t^b_z = 0 \), at the upper side of a domain reduces to:

\[
\mu \left( \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \right) = 0 \tag{8a}
\]

\[
\lambda \left( \frac{\partial u_x}{\partial x} + \frac{\partial u_z}{\partial z} \right) + 2\mu \frac{\partial u_z}{\partial z} = 0 \tag{8b}
\]

Using the same nomenclature, the value of the controlled voltage source is:

\[
u_{x,k,3} = u_{x,k,0} - \frac{\Delta z}{4\Delta x} (u_{z,kr,3} - u_{z,kl,3}) \tag{9a}
\]

\[
u_{z,k,3} = u_{z,k,0} - \frac{\Delta z}{\lambda + 2\mu} \frac{u_{x,kr,1} - u_{x,kl,1}}{2\Delta x} \tag{9b}
\]

For the circuit corresponding to the \( x \) displacement component, Equation (9a) defines the value of the controlled voltage that implements the horizontal component of the free boundary condition, Figure 5(b). The other part of the condition is defined by the Equation (9b) and implemented by a similar component in the circuit related with the \( z \) displacement component.
The network model for the source of the seismic perturbation, that reproduces the standard Ricker pulse, is implemented by a generator with output signal obeying the Equation (6). In order to release the domain to excitation source, an electrical switch is imposing between the Ricker pulse generator and the node where it is applied. The network model for this condition is shown in Figure 6. The switch opens when pulse is finished, releasing the node allowing to move freely.

The connection between $N_x \times N_z$ cells, by ideal electrical contacts, for the $x$ component of the displacement (Figure 4), and $N_x \times N_z$ cells for the other $z$ component, defines the whole network of the physical domain to which mixed boundary conditions and Ricker pulse must be added, Figures 5 and 6.

Since very few components are required in most cases, including boundary conditions, very few rules are required for the design of the complete network model, which is run without other mathematical manipulations in a suitable code such as PSpice. The powerful computational algorithms implemented in such codes makes that the simulation reproduces the exact solution of the model so that the errors are only due to the chosen grid size.
4 THE CODE PSPICE

Once the complete network model is designed, its simulation is carried out using a standard circuit simulation analysis code such as PSpice [9] without any other mathematical requirements. This provides the exact solution of the model. PSpice was initially developed by the Integrated Circuit group of the Electronic Research Laboratory at the University of California, and the first version was finished in 1992. Since that time, thousands of copies of the new version of the program have been given to universities and companies in the electronic industry. The widespread use of PSpice and its new versions attests to the applicability of the program to a large variety of circuit simulation problems and provide a valuable base of experience that demonstrated the advantages and disadvantages of the powerful, efficient and reliable numerical algorithms that are employed in the program.

The system of equations that describes the complete circuit is determined by the model equations for each element and topological constraint, which are determined by interconnecting the elements, reflecting Kirchhoff’s Current and Voltages Laws (Kirchhoff’s Current law is a conservation principle while Voltages Kirchhoff’s law ensures the uniqueness of the related variable).

For transient analysis, PSpice maintain an internal time-step, which is continuously adjusted to maintain accuracy, while not performing unnecessary steps. The time-step is reduced by the code during the simulation so that the integrated charges and currents are sufficiently accurate. During periods of inactivity, the internal time-step is increased, while during activity periods, it is decreased. The maximum internal step size can be controlled by using software to specify it. The minimum time-step is the overall run time divided by $1E12$, while the initial time-step is fixed as a function of both the total time required for the simulation and the value of the relative accuracy, a data (parameter) also specified by the user. If convergence is not reached, the initial time-step is successively reduced in the simulation until the solution converges.

The circuit equations are a system of algebraic-differential equations of the form $F(x, x', t) = 0$, where $x$ is the vector of the unknown circuit variables, $x'$ is the time derivative of $x$ and $F$ is, in general, a nonlinear operator. The equations formulation algorithm based on a combination of Cutset and Loop Analysis is a modified version of the classical Nodal Analysis. This method provides the same generality as other formulation methods, but produces a near-symmetric system of equations that are solved with an amount of computational effort that is comparable to that needed for the simplest Nodal Methods. The Markowitz algorithm is used for reordering the system of equations. Once the equations $F(x, x', t) = 0$ are reordered, direct eliminations methods, such as $L>U$ factorization and sparse-matrix, are used to obtain the solution.

If the circuit contains elements that are modeled by nonlinear equations, then the solution is obtained by an iterative sequence of linearized solutions. The modified Newton-Raphson algorithm, which approximates each non-linearity by a Taylor series that is truncated after the first order term, is the most common method of linearization. The modified Newton-Raphson method contains the reliable and efficient simple-limiting of Colon (algorithm) that requires the lowest number of iterations for convergence.

For the implicit transient analysis method of numerical integration, stiffly-stable
algorithms are used, such as Trapezoidal integration. The local truncation error of the integration is proportional to the time-step, which is successively reduced to make the error negligible.

5 APPLICATION

A square domain has been used to simulate P-SV seismic waves subjected to mixed boundary conditions. The length of the medium is \( L_x = L_z = 5000 \) m. Physical properties are: density \( \rho = 1000 \) kg/m\(^3\), Poisson’s ratio \( \nu = 0.25 \) and shear modulus \( \mu = 1.33 \) GPa. The corresponding propagations velocity of the P and S waves are \( \alpha = 2000 \) m/s and \( \beta = 1155 \) m/s, respectively. For the upper side, free boundary condition is prescribed. Zero displacements are imposed at the rest of the boundary. The perturbation Ricker pulse, Equation (6), is applied in the middle of the bottom side of the domain using \( t_p = 1 \) s and \( t_s = 2 \) s values. The chosen grid is 31x31 and the simulation time window 10 s. Figures 7 to 9 show the results using Matlab.
Figure 7: Deformed domain at $t = 3, 4, 5, 6, 7, 8, 9$ and $10$ s.

Figure 8: Vertical displacements at $x=2500$ m
6 CONCLUSIONS

- Application of the network method to solving the P-SV wave equation has been demonstrated to be an efficient tool for the numerical simulation of this problem. The design of the network as well the implementation of the mixed boundary conditions, whatever they are, has been made in an easy way thanks to: i) the rules needed for the design are very few and ii) the existence of the especial controlled sources defined in the libraries of the circuit simulation codes; these sources are capable of implementing by software any kind of non-linearity and coupled condition involved in the problem.

- No mathematical manipulation other than the spatial discretization of the partial differential equation is required since this work is made by the software PSpice and the data treatment routines.

- The application of the free boundary condition is also direct by using the above mentioned controlled source that is suitable for solving the coupling between the two equations.

REFERENCES


International (2007) 40, 139-145.
SCHWARZ ALTERNATING DOMAIN DECOMPOSITION APPROACH FOR THE SOLUTION OF MIXED HEAT CONVECTION FLOW PROBLEMS BASED ON THE METHOD OF APPROXIMATE PARTICULAR SOLUTIONS

CARLOS A. BUSTAMANTE*, HENRY POWER† AND WHADY F. FLOREZ†

*Institute of Energy, Materials and Environment
Universidad Pontificia Bolivariana
Circular 1 No 73-34, Medellín, Colombia
e-mail: carlos.bustamante@upb.edu.co, whady.florez@upb.edu.co web page: http://www.upb.edu.co/

†School of Mechanical, Materials and Manufacturing Engineering
University of Nottingham
University Park, Nottingham NG7 2RD, UK
e-mail: henry.power@nottingham.ac.uk - Web page: http://www.nottingham.ac.uk

Key words: Instructions, Coupled Problems, Multiphysics Problems, Applications, Computing Methods

Abstract. The incompressible two-dimensional Navier-Stokes equations including thermal energy balance equation are solved by the recently developed Method of Approximate Particular Solutions (MAPS). In a previous authors’ work this method was implemented to solve the two-dimensional Stokes equations by employing the pressure and velocity particular solutions obtained by Oseen’s decomposition with the Multiquadric (MQ) RBF as non-homogeneous term. A pressure-velocity linkage strategy is not required since the pressure particular solutions are obtained from the velocity ones. In the present contribution, the Navier-Stokes equations with Boussinesq approximation are solved by linearizing the convective term in a Picard iterative scheme. With the velocity values obtained at each of the Picard iterations, the energy conservation equation is solved by the MAPS by approximating temperature with the particular solutions of a Poisson problem with the MQ as a forcing term. With the aim of improving the computational efficiency of the global strategy, the two-dimensional domain is split into overlapped rectangular subdomains where the Schwarz Alternating Algorithm is employed to find a solution by using velocity and temperatures values from neighbouring zones as boundary conditions. The mixed convection lid-driven cavity flow problem is solved for moderate Reynolds and low Richardson numbers with the aim of validating the proposed method.
1 INTRODUCTION

Mixed natural and forced convection problems are frequently found in industrial applications [1, 2]. Therefore, a detailed understanding of the transport phenomena involved is a key aspect when designing new devices or improving old designs. In this sense, numerical methods for solving partial differential equations (PDEs) have become an interesting tool because they allow obtaining a deterministic description of temperature, pressure and velocity field by solving the momentum and energy conservation equations in their differential form which in case of incompressible fluid problems are known as the Navier-Stokes equations. The accuracy of new numerical methods for solving PDEs ought to be verified by comparing their results to analytical solutions or other tested numerical results. When dealing with non-isothermal flows with mixed convection, the lid-driven cavity flow problem with differentially heated top and bottom walls has widely employed for code verification. Besides validation, several authors analysed the influence of the characteristic dimensionless numbers on the heat transferred through the cavity walls quantified by the Nusselt (\(Nu\)) number. For instance, Torrance et al. [3] studied the buoyancy effect on the flow structure by changing the Grashof (\(Gr\)) number, while Moallemi and Jang [1] found that the buoyancy effect are more notorious and the heat transfer is higher as the Prandlt (\(Pr\)) number is increased. Iwatsu et al. [4] analysed the influence of the Richardson (\(Ri = Gr/Re^2\)) number in the cavity flow problem based on numerical results. They concluded that for \(Ri < 1\) the buoyancy effect is almost neglected and the flow structure is similar to the one found for isothermal flows, while for \(Ri > 1\) the buoyancy effect sharply modified the flow structure. More recently, T.S. Cheng [5] studied the relationship between \(Nu\) and \(Pr, Re\) and \(Ri\) in the cavity flow problem. Based on numerical results, the author corrected the correlation proposed by Moallemi and Jang [1] to take into account the sudden decrease in \(Nu\) value when increasing \(Re\) for \(Ri > 1\) due to the change in the flow structure. In the present work some of the results found by the aforementioned authors are employed with the aim of validating a novel meshless strategy based on the Method of Approximated Particular Solutions (MAPS) for the solution of mixed convection problems with the Boussinessq approximation.

Meshless methods have been intensively developed during the last two decades due to its potential characteristics to deal with complex geometry domains without spending too much CPU time in the pre-processing phase. Among the meshless methods, collocation schemes have offered high accuracy as well as versatility to enforce boundary conditions in complex geometries. The Radial Basis Function (RBF) collocation method, originally suggested by Kansa [6], has been successfully used for the solution of several boundary value problems governed by different PDEs. However, it is well known that global RBF collocation methods suffer of a fundamental problem described by Robert Schaback [7] as the uncertainty relation: Better conditioning is associated with worse accuracy, and worse conditioning is associated with improved accuracy. This problem can be mitigated by using integrated RBF approaches such as the indirect RBF (IRBF) collocation method.
Figure 1: Solution domain and boundary conditions

proposed by Mai-Duy and Tran-Cong [8] and the Method of Approximate Particular Solutions (MAPS) developed by Chen et al. [9]. In both schemes the RBFs are used to approximate the highest order derivative in the PDE (IRBF) or the complete PDE (MAPS), thus the solution is approximated by an integration process, which unlike derivation does not contain inherent inaccuracy of the approximation.

In previous authors’ work [10] a new meshless method for solving the Navier-Stokes equations was developed and used to solve some benchmark flow problems such as the square cavity up to \(Re = 3200\) and the backward facing step at \(Re = 800\). This approach is based on the Method of Approximate Particular Solutions (MAPS) proposed by Chen et al. [9]. In order to achieve a more efficient strategy without affecting the accuracy obtained with the global method, we employ the Schwartz alternating algorithm as it was originally proposed by Schwarz [11]. This is a suitable option since the MAPS can be employed in its global version to solve the problem in relatively small overlapped subdomains without losing accuracy and preserving its stability in terms of the shape parameter value. The present work is sorted as follows. In the first section the non-isothermal lid-driven cavity flow problem is detailed as well as the governing equations. Then, a brief description of the MAPS for solving scalar and the two-dimensional Navier-Stokes equations is made. Following, the Schwarz Alternating scheme and the decoupling algorithm for solving in sequential way momentum and energy conservation equations are presented. Finally, the numerical results are shown and discussed.

2 PROBLEM DESCRIPTION AND GOVERNING EQUATIONS

The square cavity domain of side \(L\) and the problem boundary conditions are shown in Figure 1. The domain is two-dimensional and it is filled with an incompressible fluid. As can be observed the vertical walls are isolated while the horizontal one at the bottom is at temperature \(T_c\) and the upper wall is at temperature \(T_h\) and moves with horizontal velocity \(U\).

The incompressible steady Navier-Stokes equations in its primitive variable formulation
with Boussinesq approximation to take into account the effect of temperature on density, is given by the following equations

\[
\frac{\partial u_j}{\partial x_j} = 0 \quad (1)
\]

\[
\rho u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \mu \frac{\partial^2 u_i}{\partial x_j \partial x_j} + g_i \beta (T - T_c) \quad (2)
\]

\[
u_j \frac{\partial T}{\partial x_j} = \alpha \frac{\partial^2 T}{\partial x_j \partial x_j}, \quad (3)
\]

which are the mass, momentum and energy conservation equations, with \( i = 1, 2 \) for two-dimensional problems. The fluid properties \( \rho, \mu, \beta \), and \( \alpha \) are, respectively, density, absolute viscosity, thermal compressibility coefficient and thermal diffusivity, while \( g \) refers to the gravity vector and \( T_c \) to the lowest temperature in the system. Considering that \( U \) and \( L \) are the characteristic velocity and length and \( T_H \) is the highest temperature, the dimensionless numbers that characterise the situation are 

\[
Re = \frac{\rho U L}{\mu}, \quad Pr = \frac{\mu}{\rho} \frac{1}{\alpha}, \quad Gr = |\vec{g}| \rho^2 \beta (T_h - T_c) \frac{L^3}{\mu^2} \quad \text{and} \quad Ri = \frac{Gr}{Re^2}.
\]

Heat transfer through the top \( (b = L) \) or bottom \( (b = 0) \) wall is quantified by the local \( Nu \) number, which is defined as

\[
Nu(x_1, b) = \frac{1}{T_H - T_c} \left. \frac{\partial T}{\partial x_2} \right|_{x=(x_1,b)} \quad (4)
\]

and by its average given by 

\[
\overline{Nu} = \int_0^1 Nu(x_1, b) dx_1
\]

### 3 METHOD OF APPROXIMATED PARTICULAR SOLUTIONS

The first MAPS version, proposed as a global scheme by Chen et al. [9], is briefly presented in the first part of this section. This version is employed here in order to solve the energy conservation equation (3). In the second part, the proposed MAPS version for solving the isothermal Navier-Stokes equation is presented.

#### 3.1 MAPS for scalar problems

Let us consider the case of a linear boundary value problem whose partial differential operator \( L(\vec{x}) \), or only part of it, is in terms of the radial component of a polar or spherical coordinate system (i.e. axisymmetric), as:

\[
L(u(\vec{x})) = L_r(u(r)) + L_\vec{x}(u(\vec{x})) = f(\vec{x}), \quad (5)
\]

and

\[
B(u(\vec{x})) = g(\vec{x}) \quad \forall \vec{x} \in \Gamma \quad (6)
\]

with \( L_r(u) \) as the axisymmetric part of the PDE and \( B \) as the boundary operator.
The axisymmetric part of the PDE, is approximated by RBFs, as:

\[ L_r(u(r)) = \sum_{k=1}^{N} \alpha_k \phi(r_k) \]  

(7)

where the non-homogeneous term in the momentum equation, \( \phi \), is defined as the Multiquadric (MQ) RBF, \( \phi(r) = (r^2 + c^2)^{1/2} \), which only depends on the Euclidean distance \( r \) between a field point \( \vec{x} \) and a trial point \( \vec{\xi} \) and the shape parameter \( c \). In consequence, the field variable can be expressed as

\[ u(\vec{x}) = \sum_{k=1}^{N} \alpha_k \hat{u}(r_k), \]  

(8)

with \( \hat{u}(r) \), as the corresponding particular solution of the following non-homogeneous ordinary differential equation:

\[ L_r(\hat{u}(r)) = \phi(r). \]  

(9)

Thus the complete linear PDE operator can be expressed as a function of the RBFs and the particular solutions in the following way:

\[ L(u(\vec{x})) = \sum_{k=1}^{N} \alpha_k [\phi(r_k) + L_{\vec{x}}(\hat{u}(r_k))] = f(\vec{x}), \]  

(10)

By substituting the above approximation for \( u(\vec{x}) \), equation (8), into the boundary conditions of the problem, and into the full expression of PDE, the following linear system of algebraic equations is obtained:

\[
\begin{pmatrix}
B[\hat{u}_{1,1}] & \cdots & B[\hat{u}_{1,N}] \\
\vdots & \ddots & \vdots \\
B[\hat{u}_{N_b,1}] & \cdots & B[\hat{u}_{N_b,N}] \\
\phi_{N_b+1,1} + L_{\vec{x}}[\hat{u}_{N_b+1,1}] & \cdots & \phi_{N_b+1,N} + L_{\vec{x}}[\hat{u}_{N_b+1,N}] \\
\vdots & \ddots & \vdots \\
\phi_{N,1} + L_{\vec{x}}[\hat{u}_{N,N}] & \cdots & \phi_{N,N} + L_{\vec{x}}[\hat{u}_{N,N}]
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\
\vdots \\
\alpha_{N_b} \\
\alpha_{N_b+1} \\
\vdots \\
\alpha_N
\end{pmatrix} =
\begin{pmatrix}
g_1 \\
\vdots \\
g_{N_b} \\
f_{N_b+1} \\
\vdots \\
f_N
\end{pmatrix}
\]  

(11)

for \( N_b \) boundary points and \( N_i \) internal points, with \( N = N_b + N_i \) and \( \hat{u}_{ij} = \hat{u}(\vec{x}_i, \vec{\xi}_j) \). The solution of the boundary value problem is achieved after solving the resulting algebraic system for the coefficients \( \alpha \).
3.2 Solution of the Navier-Stokes equations by MAPS

Before solving the incompressible and isothermal Navier-Stokes equations by the MAPS, the procedure presented in [12] is done in order to obtain the particular solutions $\hat{u}_i^l$ and $\hat{p}^l$ by employing the Oseen’s decomposition formula applied to the Stokes problem. In this way, the approximated velocity and pressure fields, $\tilde{u}$ and $\tilde{p}$, can be expressed as a linear superposition of $N$ particular solutions located at $N$ trial points $\xi_k$, as:

$$u_i(x) = \sum_{k=1}^{N} \alpha_k^l \hat{u}_i^l(r_k)$$

$$p(x) = \sum_{k=1}^{N} \alpha_k^l \hat{p}^l(r_k)$$

where $r_k = |x - \xi_k|$.

By substituting the above expressions into a linearized version of the momentum equation (2), given by

$$\mu \frac{\partial^2 u_i}{\partial x_j \partial x_j} - \rho u_i^* \frac{\partial u_i}{\partial x_j} - \frac{\partial p}{\partial x_i} = 0,$$

with $\hat{u}^*$ as the solution at the previous iteration of the Picard algorithm, the following homogeneous linear superposing of functions, representing the approximated momentum equation, is obtained:

$$\sum_{k=1}^{N} \alpha_k^l \left[ \mu \frac{\partial^2 u_i^l(r_k)}{\partial x_j \partial x_j} - \frac{\partial p^l(r_k)}{\partial x_i} - \rho u_j^* \frac{\partial u_i^l(r_k)}{\partial x_j} \right] = \sum_{k=1}^{N} \alpha_k^l \left[ \phi(r_k) \delta_{il} + \Theta_l^i(r_k) \right] = 0,$$

and the last term on the left is

$$\Theta_l^i(r_k) = -\rho u_m^*(\bar{x}) \frac{\partial u_i^l(r_k)}{\partial x_m}.$$

The set of equations required to complete the collocation process is obtained by substituting the approximations (12) and/or (13) into the respective boundary condition. By collocating the resulting expression at the $N_b$ boundary nodes on $\Gamma_u$, for each of the components $k = 1, 2$, the first two set of lines on the matrix system (17) are found, while collocation of (15) at the $N_i$ internal nodes, for each of the components $k = 1, 2$, defines the last two set of equations of the matrix system

$$\begin{pmatrix} B^1 [\hat{u}_1^1, \hat{u}_2^1, \hat{p}^1] & B^1 [\hat{u}_1^2, \hat{u}_2^2, \hat{p}^2] \\ B^2 [\hat{u}_1^1, \hat{u}_2^1, \hat{p}_1^1] & B^2 [\hat{u}_1^2, \hat{u}_2^2, \hat{p}_1^2] \\ \phi + \Theta_1^1 & \Theta_1^2 \\ \Theta_2^1 & \phi + \Theta_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1^1 \\ \alpha_1^2 \\ \alpha_2^1 \\ \alpha_2^2 \end{pmatrix} = \begin{pmatrix} [g(\bar{x})_1] \\ [g(\bar{x})_2] \\ [0] \\ [0] \end{pmatrix},$$

(17)
where a general form of the boundary condition $B^k[u_1, u_2, p] = g(\bar{x})_k$, with $B^k$ as the boundary operator and $g_k$ the corresponding value of the boundary condition, was employed. Despite the fact that the continuity equation is not explicitly imposed in the resulting matrix system of equations, the formulation is mass conservative since the used superposition of particular solutions exactly satisfy the continuity equation.

4 SOLUTION ALGORITHMS

In the first part of this section the Schwarz Alternating Method is explained for its application to solve isothermal Navier-Stokes equation. A similar algorithm which is not shown here for brevity is used when solving the energy equation. The decoupled algorithm for solving the PDE system (mass, momentum and energy equations) is then detailed in the second part.

4.1 Schwarz Alternating Method

The Schwarz Alternating Method is a suitable strategy to improve the computational efficiency of the MAPS since local problems can be solved as though it were a global problem which only depends on the neighbouring subdomains by the boundary conditions. Let’s suppose a rectangular domain split in two subdomains ($\Omega^1$ and $\Omega^2$) as it is shown in Figure 2a. The original Navier-Stokes problem can be rewritten as follows

$$
\rho u_j^k \frac{\partial u_i^k}{\partial x_j} = -\frac{\partial p^k}{\partial x_i} + \mu \frac{\partial^2 u_i^k}{\partial x_j \partial x_j} \quad \forall \bar{x} \in \Omega^k, \tag{18}
$$

$$
u_i^k = u_{ib} \quad \forall \bar{x} \in \Gamma, \tag{19}
$$

where the superscript $k$ refers to the corresponding subdomain. The Schwarz solution is achieved by solving, in first place, the above equations for $k = 1$ and the boundary condition given by $u_i^1 = u_i^2 \quad \forall \bar{x} \in \Gamma^{12}$, then for $k = 2$ and the boundary condition given by $u_i^2 = u_{ib} \quad \forall \bar{x} \in \Gamma^{21}$, with $\Gamma^{12}$ and $\Gamma^{21}$ the overlapping subdomains.
\( u_2^i = u_1^i \quad \forall \vec{x} \in \Gamma^{21} \), and so on until the \( L_2 \)-norm of the difference between successive solutions go less than the specified tolerance. Since we are only dealing with Dirichlet boundary conditions, an additional equation must be solved after solving the Navier-Stokes equations in order to guarantee continuity of pressure throughout the global domain. Regarding pressure particular solution is obtained by integration, the approximated pressure is expressed as:

\[
p^m(\vec{x}) = \sum_{k=1}^{N^m} \alpha_{lm} \hat{p}^l \left( \| \vec{x} - \vec{\xi}_k \| \right) + c^m
\]

with the superscript \( m = 1, 2 \) indicating the corresponding subdomain and \( c^m \) as the integration constant. The following equation is obtained after evaluating the pressures at some point \( \vec{x}_{12} \) in the overlapping zone \( \Omega^{12} \):

\[
c_2 - c_1 = \sum_{k=1}^{N^1} \alpha_{1k}^{11} \hat{p}^l \left( \| \vec{x}_{12} - \vec{\xi}_k \| \right) - \sum_{k=1}^{N^2} \alpha_{2k}^{12} \hat{p}^l \left( \| \vec{x}_{12} - \vec{\xi}_k \| \right).
\]

The above equation can be solved after fixing the value of one of the integration constants. In case of splitting the domain into more than two subdomains, as shown in Figure 2b for \( s = 9 \) with \( s \) as the number of subdomains, the solution is obtained by solving the problem given by equations (18) to (19). In this case the superscript \( k \) changes in the sequential order shown in Figure 2b and the solution \( k+1 \) is obtained by using the boundary conditions with the variable value available in memory either after obtaining a solution for the \( j = 1, \ldots, i - 1 \) subdomains or as the initial value in the first Schwarz iteration. The matching equation for pressure (21) becomes an overdetermined equation system with \( s - 1 \) variables (integration constants) and one equation for each overlapped zone in the domain. For better understanding of the Schwarz scheme, the algorithm employed for solving a generic PDE with dependent variable \( \phi \) is presented:

- After setting an initial guess \( \phi^0(\vec{x}) \), compute the interpolation matrix on the left hand side of equation (17) when solving the Navier Stokes equations or (11) in case of scalar problems, for each of the \( s \) subdomains.

- For each subdomain \( (l = 1, \ldots, s) \) calculate the column vector on the RHS of equation (17) or (11) with the variable values available in memory and solve the equation system to obtained \( \alpha \) coefficients.

- For each subdomain \( (l = 1, \ldots, s) \) reconstruct \( \phi \) and store its value in memory

- Evaluate the \( L_2 \)-norm of the difference between the values of the variable \( \phi \) at the present and previous Schwarz iterations \( (||\phi^n - \phi^{n-1}||) \). If it is less than or equal to the set tolerance \( tol_{\phi} \), for all of the variables, stop the Schwarz algorithm, if not go to the second step with \( \phi^{n-1} = \phi^n \).
4.2 Sequential solution procedure

The following algorithm is implemented for solving in a decoupled way the equations 1,2 and 3:

• By using the MAPS as explained in section 2.2, solve the following form of the
momentum equation in order to obtain the fields \(\bar{u}^k\) and \(p^k\) based on the guess (first iteration) or the previous iteration variable values \(\bar{u}^{k-1}\) and \(T^{k-1}\)

\[
\mu \frac{\partial^2 u^k}{\partial x_j \partial x_j} - \rho u^{k-1}_j \frac{\partial u^k}{\partial x_j} - \frac{\partial p^k}{\partial x_i} = g_i \rho \beta (T^{k-1} - T_c). \tag{22}
\]

• By employing the MAPS for scalar problems (section 2.1) and with the velocity obtained in the previous step, solve the following form of the energy equation in order to obtain the updated temperature value \(T^k\)

\[
\alpha \frac{\partial^2 T^k}{\partial x_j \partial x_j} - u_j^k \frac{\partial T^k}{\partial x_j} = 0. \tag{23}
\]

• If \(\|\phi^k - \phi^{k-1}\| < tol\) for all \(\phi = \bar{u}, p, T\), stop, if not make \(\phi^{k+1} = \phi^k\) and go to the first step.

5 NUMERICAL RESULTS

Numerical results were obtained for \(Re = 400, 1000\) and \(Gr = 1 \times 10^2, 1 \times 10^4\) with a 41 \times 41-point nodal distribution refined towards the boundaries and \(c = 0.01\) for velocity and temperature approximation. Besides, the domain is split into \(s = 4 \times 4\) overlapped subdomains in order to applied the Schwarz alternating algorithm. As shown in Table 1, the mentioned dimensionless numbers were achieved by combining some values of the fluid
Table 1: Properties and Dimensionless number values for the solved cases

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$ (kg/m$^3$)</td>
<td>$4.00 \times 10^2$</td>
<td>$4.00 \times 10^2$</td>
<td>$1.00 \times 10^4$</td>
<td>$1.00 \times 10^3$</td>
</tr>
<tr>
<td>$\beta$ (1/oC)</td>
<td>$6.25 \times 10^{-5}$</td>
<td>$6.25 \times 10^{-3}$</td>
<td>$1.00 \times 10^{-5}$</td>
<td>$1.00 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\alpha$ (m$^2$/s)</td>
<td>$3.52 \times 10^{-3}$</td>
<td>$3.52 \times 10^{-3}$</td>
<td>$1.41 \times 10^{-3}$</td>
<td>$1.41 \times 10^{-3}$</td>
</tr>
<tr>
<td>$Re$</td>
<td>$4.00 \times 10^2$</td>
<td>$4.00 \times 10^2$</td>
<td>$1.00 \times 10^3$</td>
<td>$1.00 \times 10^3$</td>
</tr>
<tr>
<td>$Gr$</td>
<td>$1.00 \times 10^2$</td>
<td>$1.00 \times 10^2$</td>
<td>$1.00 \times 10^2$</td>
<td>$1.00 \times 10^2$</td>
</tr>
<tr>
<td>$Ri$</td>
<td>$6.25 \times 10^{-4}$</td>
<td>$6.25 \times 10^{-2}$</td>
<td>$1.00 \times 10^{-4}$</td>
<td>$1.00 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Figure 4: Numerical results for $Gr = 100$ and $Re = 400$ in comparison to the numerical solution reported by Iwatsu et al. [4], a. $T$ on $x_1 = 0.5$, b. $Nu$ on $x_2 = 0.0$ (○) and $x_2 = 1.0$ (*)

properties since boundary conditions different to zero were fixed as the unity, i.e. $U = 1$ and $T_H = 1$. The fixed properties were $\mu = 1 kg/m/s$, $Pr = 0.71$ and $\vec{g} = (0, -10)m/s^2$.

In Figure 3, the obtained velocity profiles are compared to the ones attained by Iwatsu et al. [4]. Numerical results are in good agreement to the reference solution regarding that the nodal distribution used here ($41 \times 41$) is coarser than the one used by Iwatsu et. al ($128 \times 128$). For $Re = 400$ and 1000 and $Gr = 100$, the buoyancy effect is not predominant since Richardson numbers are, respectively, $Ri = 6.25 \times 10^{-4}$ and $1.00 \times 10^{-4}$, which are much less than the unity. Therefore, the velocity field is close to the isothermal pattern. Despite of the coarseness of the nodal distribution employed, accurate results were found for temperature (Figure 4a) and the local $Nu$ on the bottom and top boundaries (Figure 4b). Since local $Nu$ is calculated based on the temperature derivatives, it is not as accurate as temperature profile, increasing the difference towards the corners in the case of the top boundary profile and near to the centre for the bottom boundary. Nevertheless, the highest difference between the obtained and the reference average $Nu$ values, presented in Table 2 for the top boundary, are 2.54% and 9.42% when comparing to the results reported in [4] and [5], respectively. By using the aforementioned solution algorithm, the highest $Ri$ values achieved before the Picard iterations diverge, were $Ri = 6.25 \times 10^{-2}$ and $1.00 \times 10^{-2}$ with $Re = 400$ and 1000 ($Gr = 1 \times 10^4$), respectively. In consequence,
Table 2: Average Nusselt number on top boundary

<table>
<thead>
<tr>
<th>$Re$</th>
<th>Reference</th>
<th>Grid</th>
<th>$Gr$</th>
<th>$1 \times 10^2$</th>
<th>$1 \times 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Iwatsu et al. [4]</td>
<td>$128 \times 128$</td>
<td>3.84</td>
<td>3.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>$41 \times 41$</td>
<td>3.75</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>Iwatsu et al. [4]</td>
<td>$128 \times 128$</td>
<td>6.33</td>
<td>6.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cheng [5]</td>
<td>$128 \times 128$</td>
<td>6.73</td>
<td>6.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>$41 \times 41$</td>
<td>6.20</td>
<td>6.13</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Approximated streamlines (a.) and temperature contours (b.) for $Gr = 1 \times 10^4$ and $Re = 1000$

the approximated streamlines and the temperatures contours, shown for $Re = 1000$ in Figure 5, present a forced convection-dominated behaviour since the buoyancy effect is not notorious. The flow structure and temperature contours are in good agreement with the results reported by Moallemi and Jang [1] in their Figure 2.

6 CONCLUSIONS

- The Method of Approximate Particular Solutions in conjunction to the Schwarz alternating algorithm were used to solve the two-dimensional mixed heat convection lid-driven cavity flow problem at $Re = 400$ and $Re = 1000$ for low $Ri$. The obtained results are in good agreement to the ones reported in literature, despite the nodal distributions used are much coarser than the ones employed by the reference authors.

- Future authors’ work will be focused on obtaining results for higher $Ri$ both by using denser nodal distribution and by implementing a more stable algorithm for the decoupled solution of the momentum and energy conservation equations.
REFERENCES


VIRTUAL REALITY 3D SIMULATION OF FLUID-PARTICLE INTERACTION

MOHAMED CHEKIRED*, VARVARA ROUBTOVA*

* Institut de recherche d’Hydro-Québec (IREQ)
1740, boul. Lionel-Boulet, Varennes,
Québec, J3X 1S1, Canada
Roubtsova.varvara@ireq.ca, web page http://groupe technologie.hydro.qc.ca

Keywords: Fluid Flow, Marker and Cell Method, Fluid-Particle Interaction.

Abstract. A phenomenological, pore-scale, virtual reality hydrodynamics laboratory, called SimSols, is being developed to simulate in 3D the processes that saturated and unsaturated soils undergo under static and dynamic loads. The main objective of the virtual laboratory is to explore the various features of porous media, in order to enhance our understanding of the processes at work at micro level and their influence at macro level. SimSols will mimic the phenomena observed in the soil, which will enhance earthwork design and explain atypical behaviors by providing a better understanding of these phenomena through visualization and without the usual challenges involved in studying them. Early virtual laboratory results have been validated by a series of experiments carried out at the École de Technologie Supérieure (ÉTS) in Montreal (QC), Canada. This paper presents 3D cases involving the interaction of glass beads with viscous fluid.

1 INTRODUCTION

It is important to understand the elementary soil-water interaction processes that take place in earth dams and the thresholds of these processes, as they control a range of phenomena that occur in this kind of structure. Especially critical are the processes involving seepage that lead to particle migration. Seepage can be unsteady during construction and steady during the initial stages following construction. The interaction of earth dams with groundwater flow is very complex. The velocity of this water and its flow rate are controlled by the permeability of the soil, and the seepage forces acting on the soil skeleton are accompanied by the force of buoyancy. The latest trends call for numerical simulation to be carried out at the finest possible resolution, which, in the case of seepage, means a change of model. Euler’s equations of flow for an incompressible fluid, in the form of Darcy’s law, take into account the relationship between the velocity of the liquid and the hydraulic gradient, and are used to measure seepage at macro level. The model yields an average velocity, and enables simulation of a real viscous-incompressible flow into pore channels, and its interaction with soil particles, at the finest resolution. This interaction is one of the primary factors involved in predicting the start of erosion.

One of the possible applications of SimSols is the simulation of the internal erosion process at micro level. A Euler-Lagrange formulation is used to simulate flow and particle migration.
Particle motion is simulated by the well-known Discrete Element Method (DEM) [1] using a Hertz-Mindlin [2, 3] contact model. The fluid phase is modeled as a continuum by solving the Navier-Stokes equations using the Marker and Cell (MAC) method [4]. No slip boundary conditions are used to calculate the influence of particles on the fluid. The influence of fluid on the particles is taken into consideration by integrating the pressure field and viscous stress around the particle. To solve the massive calculus involved, SimSols can both be run on GPU cards than on computing clusters. The programming codes for SimSols are in a continuous state of development. Every physical phenomenon that is added to the model is validated in the laboratory. In this article, we compare the results obtained from numerical simulation with those of laboratory tests for 3D flow into channels containing obstacles.

2 FLUID FLOW

The water flow through a pore channel is described by the Navier-Stokes equations for viscous-incompressible liquid, and is expressed by

\[
\rho \left[ \frac{\partial \vec{v}}{\partial t} + (\vec{v} \nabla) \vec{v} \right] = -\nabla p + \eta \nabla^2 \vec{v} + \vec{f}
\]

(1)

\[
\nabla \vec{v} = 0
\]

(2)

where \( \rho \) is the density of the liquid (kg/m\(^3\)), \( t \) is the time (s), \( \vec{v} \) is the flow velocity (m/s), \( p \) is the water pressure (Pa), \( \eta \) is the dynamic viscosity (Pa.s), and \( \vec{f} \) is the force of gravity (N). The presence of nonlinearities and small parameter for the second derivate are the well-known difficulties encountered in solving these equations numerically. The flow in the pore channels can be both steady and unsteady. It can also be both laminar and turbulent. The laminar-to-turbulent transition is particularly important in the erosion process, when finer particles start to leave the ground. This process increases the fluid velocity and the width of the channels. The reverse process is also possible, and occurs when the fines within the pore structure clog and narrow the pore channels, slowing the fluid. The challenge is that a solution for a transitional flow creates an additional difficulty in solving the Navier-Stokes equations.

3. NUMERICAL DISCRETIZATION

Today, there are many well-known numerical methods for solving the Navier-Stokes equations. Most of them use the stream function and vorticity as the variables. With this approach, the explicit solution of a continuum equation is avoided, but it is difficult to generalize for the 3D case [5]. While there is no such difficulty with the use of physical variables, the continuum equation in this case contains only velocity (eq. 2). There is no direct dependency between velocity and pressure, as expressed by density and the state equation for compressible flow. One of the more successful methods for solving the Navier-Stokes equations with physical variables is the Marker and Cell Method (MAC), developed at the Los Alamos Laboratory in 1965 [4]. This method uses a staggered grid. Since its early success, several modified schemes have been introduced. SimSols uses the scheme based on the splitting of physical parameters [6] [7]. This scheme comprises three steps.
Step 1
\[
\frac{\hat{v} - \hat{v}_n}{\Delta t} = - (\hat{v}_n \cdot \nabla)\hat{v}_n - \eta \nabla \times \hat{\omega}
\] (3)

Step 2
\[
\Delta p = \frac{\hat{D}}{t} \text{ but } D^* = 0
\] (4)

Step 3
\[
\frac{\hat{v}^* - \hat{v}}{\Delta t} = - \nabla p
\] (5)

where \( n \) is a current number of time steps; \( \Delta t \) is a time step; \( \hat{\omega} = \nabla \times \hat{v} \) and \( D = \nabla \cdot \hat{v} \).

The following physical interpretation of this scheme is taken from [8]. The momentum vector transmits only by convection and diffusion in step 1. The temporary velocity field (\( \hat{v} \)) is compressible, but has a real physical meaning. If we apply the operator \( \text{rot} \) to equation (1) and take into consideration that \( \text{rot} \nabla p = 0 \), we have \( \text{rot} \hat{\omega} = \text{rot} \hat{v}^* = \omega^* \), which means that this field has a real velocity-vorticity. The pressure field is found from the temporary velocity field in the second step. The momentum vector is transmitted only by the pressure gradient (no convection or diffusion) in the third step.

The boundary conditions for a solid wall are non-slip and non-penetration. The channel wall was positioned so that the mesh points would correspond to the perpendicular velocity components that lie on the wall. To calculate the parallel velocity components, a special approximation near the walls for temporal velocity [8] was used. The boundary condition for the Poisson pressure equation is the Neumann condition. For the inflow condition, the perpendicular velocity component is determined, and the other components are zero. The free outflow condition was used for the output boundary.

The same non-slip conditions can be imposed for the particle boundary, but they are not useful from the point of view of numerical efficiency. The approximation, as presented in [9], was used instead. In the simplest case, all the liquid grid velocities located ‘inside’ particles are forced to equal the particle velocity, or zero in the case of fixed obstacles, before the temporary velocity is evaluated, while all the other points retain the old values. The ‘interior’ points are then updated after the pressure has been determined in the same manner as all other points, in order to ensure incompressibility.

4 MODEL VALIDATION

The Streamline approach was used to describe and validate the 3D numerical results. However, it is quite difficult to apply this approach to analyzing the streamlines in the area where they become mixed, as the images obtained are often unusable. So, in order to validate the numerical results, two specific streamlines were chosen for visualization, and the more
representative section, the one that includes these streamlines, was modeled in our series of experiments.

The experiments were carried out at the École de Technologie Supérieure (ÉTS) in Montreal (QC), Canada. Flow tests were conducted at low pressure with several obstacles placed in a horizontal clear plexiglass channel, to enable visualization of the streamline patterns. The obstacles were either vertical glass tubes 16 mm in diameter and 10 mm in height, or precision glass beads 10 mm in diameter. For each obstacle, flow tests were repeated for a minimum of 4 flow rates. Flow lines were obtained using a dye tracer. All the tests were filmed in planar view, and pictures were taken in elevation view, systematically verifying that the dye tracer remained at the mid-height point in the channel. Great care was taken to ensure that no air bubbles were trapped in the channel. The measurements recorded included water flow rate, upstream pressure, downstream pressure, water temperature, and air temperature. Figure 1 shows the various configurations of streamline patterns obtained experimentally, which were compared to those obtained from numerical simulations for the tests carried out with obstacles in the form of cylinders, and flows with different flow rates. Since the diameter of a cylinder is greater than the height of a channel, the third dimension will greatly influence flow. The layers along the height of the flow have a parabolic velocity profile, but they do not mix at low Reynolds numbers. They begin to mix with an increase in Reynolds number. For the simulation, only the middle section is presented.
The flow around spherical particles follows the same trend. Figure 2 shows the streamlines located at the mid-height point in the channel. Figure 3, in which the obstacles are translucent, shows an example of streamlines of ten superposed layers at a Reynolds number of 6. In this case, the flow lines are different, because the diameter of the obstacle is different for each layer, but the layers don’t mix. For higher Reynolds numbers, the 3D effects are visible (Figure 4).
Figure 2. The streamlines for flow with spherical obstacles.
*The Reynolds number was calculated from the diameter of the spheres.

Figure 3. The streamlines for laminar flow at Reynolds numbers of 6 for ten layers of the channel with two spherical obstacles.
Figure 4. (a) Experimental view, and (b) simulation streamlines, for a Reynolds number of 120.

The third example concerns flow through a pyramid made of glass beads. The tests were carried out in a channel 42.9 mm in height and 50 mm in width. Figure 5 shows the streamlines for a different Reynolds number, which was calculated from the diameter of the glass beads. The streamlines are presented for various layers. Since the pictures are very similar, only one planar view, corresponding to the highest layer, is presented.
Figure 5. The streamlines for different Reynolds numbers for different layers of glass beads in a pyramidal-shaped obstacle (h is the distance from the bottom of the channel, and H is the height of the channel).

Figure 6 shows the frontal view, in order to compare the results of the experiment to those obtained from numerical simulation. The difficulty was to obtain a very accurate horizontal location for the two thin needles used to release the dye in the experimental setup. Following the comparison of experimental and numerical results, a deviation of about ±1 mm between the horizontal locations of the needles was detected. The two streamlines for various sections are presented. The image would be unusable if all the streamlines for the two layers are presented. This problem does not arise in the last case (Re = 34), and we can present all the streamlines in frontal view.
<table>
<thead>
<tr>
<th>Reynolds Number</th>
<th>Experimental data</th>
<th>Simulation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td>17</td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td>25</td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>34</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 6. Comparison between experimental and numerical results.

As shown for all the examples presented (Figures 1 to 6), the numerical results are in good agreement with those obtained in laboratory testing.

12 CONCLUSIONS

SimSols is a virtual laboratory that can be used for applications in which particles interact with viscous-incompressible fluid. In this paper, we have presented the results for the interaction of fluid with motionless particles. The experimental and numerical techniques for visualizing 3D flow were improved during this work. The results obtained show that the simplest boundary conditions work very well, and so it will be possible to calculate the drag force by integrating the viscous stress tensors around the particles. This is our next challenge. Also, the results for moving particles in the 3D case without fluid were presented in [10]. Simsols can provide detailed information on the flows in granular media, such as streamlines, pressures, and velocities, at any location that we cannot capture with conventional tests. In this paper, the results were aimed at gaining an understanding of the phenomena related to geotechnical fields, but may be applied to other areas where fluid and particles are involved.

REFERENCES


A LS-DYNA/FLUKA COUPLING FOR THE NUMERICAL SIMULATION OF HIGH ENERGY PARTICLE BEAM INTERACTION WITH MATTER

MARTINA SCAPIN*, LORENZO PERONI*, VITTORIO BOCCONE†,♦ AND FRANCESCO CERUTTI†

* Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Turin, Italy
e-mail: martina.scapin@polito.it, web page: http://www.polito.it

† CERN, EN-STI-EET
CERN, CH-1211 Geneva 23, Switzerland
Email: francesco.cerutti@cern.ch - Web page: http://www.cern.ch

♦ on leave for Université de Genève,
Département de Physique Nucléaire et Corpusculaire
24 Quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland
Email: boccone@cern.ch - Web page: http://dpnc.unige.ch

Key words: high energy deposition, cylindrical shock-wave, density variation, FLUKA, LS-DYNA.

Abstract. The interaction between intense high-energy particle beams and materials provokes a sudden non-uniform temperature increase. This induces a dynamic response of the structure entailing the generation of shock-waves, which start to travel in the component producing a strong reduction in density in the impacted zone. FLUKA is a numerical tool for the calculation of the energy delivered in matter based on the Monte-Carlo method. LS-DYNA is a non-linear FE explicit code, which can use as input for the thermo-mechanical analysis the results from FLUKA code.

In this work a soft-coupling method between the two codes is developed and applied for the numerical simulation of an accidental impact of a multi-bunch 7 TeV proton beam of LHC on a metallic collimator insert. The FLUKA model was obtained using a voxel structure approach, while the FEM simulation was a 3D Lagrangian analysis. The FLUKA energy distribution, calculated on the initial geometry, is updated in accordance with the density modification during the simulation. As a matter of fact, the next bunches will impact against a lower mass target and can penetrate more in depth in the material.

These effects were evaluated with an iterative soft-coupling of the two codes, performed in Matlab. The number of the primary protons that need to be simulated determines the precision of the FLUKA results. The first FLUKA simulation, performed with an unmodified material, is followed by the first FEM mechanical analysis. At this point, at each step, the algorithm: takes as input the density map resulting from the FEM calculations, re-defines the regions
with different density in the target material, using the voxel structure and then runs a new FLUKA calculation which will be used as input for the next step FEM analysis (and so on). The first results confirm that the density reduction in the maximum deposition area provokes a reduction in the particle beam/matter interaction with a reduction of deposited energy and a sort of tunnelling effect. In more details, the peak of the energy deposition moves into the component along the direction of the beam. The comparison with the uncoupled case shows that also the pressure is affected: the maximum value decreases since the shockwave penetrates more into the material. The results also show that to be able to appreciate the difference between coupled and uncoupled analysis, it is necessary to obtain a quite significant density reduction: this occurs when the shockwaves has the time to travel radially away from the hit zone producing a significant rarefaction.

1 INTRODUCTION

Large Hadron Collider (LHC) is the most powerful particle accelerator in the world and can accelerate two proton or ion beams up to an energy of 7 TeV/c and 2.76 TeV/u respectively [1]. The LHC was built by the European Organization for Nuclear Research (CERN) in a circular 27 km long tunnel about 100 meters underground the border between France and Switzerland. The entire proton beam of the LHC is not continuous but is divided into 2808 bunches, each having $1.15 \times 10^{11}$ protons.

The total energy stored in each beam at maximum energy is about 350 MJ, two orders of magnitude higher than the other large accelerator machine like Tevatron or HERA. This large amount of energy, sufficient to melt 500 kg of copper, is potentially destructive for any accelerator components having direct interaction with the beam (e.g. the collimation system) in case of uncontrolled beam loss. For this reason, it is needed to provide a realistic assessment of possible structural damage of the components in case a fraction of the full beam is lost on them. An accurate prediction of the reliability and robustness is quite difficult, since beam-induced damages for high energy and intensity occur in a regime in which the possibility to perform experimental tests is limited. For this reason, it is of fundamental importance to develop reliable methods and accurate models that could be efficiently applied to estimate the damage occurring during a beam impact.

When a High Energy (HE) particle beam interacts with a solid target the particles deposit their energy on the material. This provokes a dynamic response of the structure entailing thermal stress waves and thermally induced vibrations or even the failure of the component. The evolution of the phenomenon is quite similar to what might happen during an explosion. The impacted part of the component reaches extremely high values of pressure and temperature and undergoes changes of state. The sudden increase in pressure originates outgoing shockwaves that, travelling through the component, lead to a substantial density reduction in the impacted part. The fact that the amount of energy absorbed by the matter is strongly density dependent implies that if a significant reduction in density occurs, this aspect has to be considered in order to correctly evaluate the consequences caused on the material hit by several high energy proton beam bunches. The main consequence is the generation of a tunneling, which implies that the proton beam penetrates more in depth in the hit target. Obviously, the effects become more and more appreciable and significant increasing the number of bunches considered impacting against the target. The main objective of this work is
to describe a methodology to be applied for taking into account for this phenomenon.

The evaluation of thermal loads on the material is performed using the energy deposition maps obtained by FLUKA (a particle physics MonteCarlo simulation package, [2, 3]) as input for a thermo-mechanical analysis, performed using the commercial code LS-DYNA [4], which is a general purpose transient dynamic finite element program including an implicit and explicit solver with thermo-mechanical and highly non-linear capabilities.

Several studies were performed with the main goal of understanding the phenomenon evolution in case of energy deposition in the matter consequent to the impact of a high energy particle beam and a solid component. Recently Tahir et al. [5] described a procedure for taking into account the influence of the density variation of the material on the overall energy deposition of the impinging. Their method involves the coupling between the FLUKA and the hydrodynamic BIG2 code. The energy maps calculated by FLUKA code is used in the BIG2 code to study the corresponding thermodynamic and the hydrodynamic response of the target that leads to a reduction in the density. The modified density distribution is used in FLUKA to calculate new energy loss distribution and the two codes are thus run iteratively. A FLUKA simulation is performed every time the density along the target axis changes of about 15-20%. Unfortunately, no more details on the technical aspects of the procedure are presented.

2 NUMERICAL MODELS

The amount of energy deposited in a material by a high energy proton beam, as well as the penetration or stop lengths, is strongly affected by the density of the material itself.

The density influences the probability of the interaction between particles and matter, playing a key role in the calculation of the proton energy loss in case of impact against a target. This implies that two materials subjected to the same impact condition, but with different Z numbers, experience different energy absorption. In particular, higher the atomic number (Z number) of the material, then higher its energy absorption. Besides, it means also that, for the same material, if there is a density modification, it should be taken into account in order to recalculate the energy deposition on the material in these new conditions.

To take in account those effects, a soft coupling between FLUKA [2, 3] and FE code LS-DYNA [4] is developed, in collaboration between Politecnico di Torino and FLUKA Team at CERN. The routine is implemented in Matlab and automatically, runs FLUKA and LS-DYNA on a Linux platform (Fedora14).

The method is applied to simulate the beam impact against a tungsten geometry, which represents a parallelepiped of 21×35×1000 mm. The mesh is such that 21×35×200 elements are used, so the elements dimensions are 1×1×5 mm. The same discretization is used both for FE and FLUKA models. For the FLUKA calculations a voxel-based method [6] is used to build the model, while for the thermo-mechanical simulation in LS-DYNA Lagrangian 3D solid elements with one integration point are used. The scheme of the LS-DYNA model is reported in Fig. 1.
The FLUKA calculation is performed for a proton bunch at 7 TeV. The beam dimensions are taken from the realistic Tertiary Collimator (TCT) accidental scenario (left of the IR5) interaction region, where the full width at half maximum of the beam are 0.12 cm and 0.076 cm along the x and y axis, respectively. For each step, a total of about 20000-30000 primaries are simulated. The time required for the simulation is reduced by splitting the CPU-demanding tasks over about 30-40 CPU cores.

In order to correctly simulate the thermo-mechanical response of the material it is necessary to take into account both the hydrodynamic behavior, using a dedicated equation of state (EOS), and the deviatoric behavior, using a dedicated strength material model. For these simulations the chosen equation of state is a polynomial form, in which the coefficients are obtained fitting a multi-phase tabular equation of state taken from the SESAME library [7], and the material model is the Steinberg-Guinan (S-G) model [8].

3 PROCEDURE DESCRIPTION

The methodology developed for the soft coupling between the two codes implies that, first of all, a sufficient number of primaries, needed to achieve a good precision on the energy deposition and beam size and intensity are defined. Then, the first FLUKA calculation is run, considering that the entire target is solid. As next step, the first FE mechanical analysis is run. At this point the iterative procedure starts and will terminate when the desired number of bunches impact against the target. Each iteration contains the total duration of a bunch pulse: the deposition phase (0.5 ns) and the successive free phase (25 ns). In this work the impact of a maximum number of 60 bunches is simulated. At each iteration, the coupling algorithm:

• takes as input the density map resulting from the FE calculation and defines discrete density levels: each level corresponds to an independent FLUKA material;
• defines, using a voxel structure, the regions with different density in the target block and associates to each voxel the corresponding material with the correct density;
• stores the energy deposition for each voxel, and runs the new FLUKA simulation;
• takes as input the energy map resulting from FLUKA calculation and defines discrete energy levels;
• generates the new FE model, associating to each FE element the corresponding energy and interpolating the SESAME EOS for getting the polynomial coefficients for each element (see e.g. [9, 10]);
• restarts the mechanical LS-DYNA simulation for the next bunch;
• analyzes the results in order to get the density map for the construction of the new FLUKA model.

The update of the FLUKA map can be performed at each bunch, but preliminary studies suggested that updating every 5 bunches is sufficient to obtain stable results.

4 RESULTS

In Figs. 2-4, some results of the numerical analysis are reported in the $x$-$z$ section of the target. In more details, the results are in terms of energy calculated by FLUKA, pressure and density maps, corresponding to the bunches number 1, 10, 20, 30, 40, 50 and 60. The results are shown for the end time of each simulation (after the free phase in which there is not the deposition, 25 ns).

Looking the results of Fig. 2, it is possible to notice that the energy distribution on the target changes, both in values and shape, during the deposition phase. This is a direct consequence of the density variation, reported in Fig. 3. As a matter of fact, the material, in which a great amount of energy is deposited, is subjected to a significant density reduction during the free expansion phase, in which the shock-wave propagates. This implies that the material becomes more transparent to the particles of the next bunch and the shower generated by the interactions of the bunch itself with the material. This provokes the so called tunnelling effect. The consequences of this are that the proton beam penetrates more in depth

![Figure 2](image_url)
in the material in the axial \( (z) \) direction (beam direction) and the energy is more diluted over the target.

**Figure 3:** Density distribution on the tungsten target in the \( z \)-\( x \) section corresponding to the maximum energy deposition in the \( y \) direction: results for the bunches number 1, 20, 30, 40, 50 and 60 after the free phase (the \( x \) dimension is amplified)

The results in terms of density emphasize the tunnelling: the density modification involves higher longitudinal coordinates increasing the number of bunches. After 60 bunches the total length of the target experiences a reduction in density in the zone around the beam axis.

The results in terms of pressure (Fig. 4) show that the maximum of pressure remains more or less in the same longitudinal position with respect of the first bunch, but the pressure wave starts to travel in the \( x \) direction. Since the pressure wave generated is cylindrical (due to the shape of the energy deposition), the same happens also in the \( y \) direction. The fact that the energy deposited by the following bunches is lower and more widespread, with respect to that calculated for the first bunch impacting the solid material, implies that the pressure increment, consequent to the next bunches, is reduced in the zone, in which the first bunch deposited a great amount of energy. On the other hand it should increase in the part of the target, in which there is an increment in density.
5 CONCLUSIONS

In this work a new methodology is applied for the numerical simulations of the multi-bunch impact of the LHC beam on a solid tungsten target. The idea was to realize a soft coupling between the FLUKA code, used for the evaluation of the thermal load on the target, and the FE code LS-DYNA, used for evaluating the thermo-mechanical response. The routine for the coupling was developed in Matlab and iteratively runs the two codes on the same Linux platform.

The main objective was to build a reliable numerical tool for predicting the consequences on the target of the tunnelling effect, produced by the propagation of the cylindrical shock-wave, generated by the impact. The shock-wave propagation leads to a substantial density reduction in the impacted part of the target, which has to be taken into account due to the fact that the amount of energy absorbed by the matter is strongly density dependent. The update of the FLUKA calculation is performed every 5 bunches, to which corresponds a significant density reduction.

The results obtained showed that the procedure is able to describe evolution of the tunnelling effects. As a matter of fact, the energy distribution over the target changes, both in values and shape, during the deposition phase: the material, in which a great amount of energy is deposited, is subjected to a significant density reduction and becomes more transparent to
the next proton bunch, reducing the amount of absorbed energy. The consequences are that the proton beam penetrates deeper in the material in beam axis direction and the energy is more diluted over the target.

REFERENCES


A NEW FUNCTIONAL FOR IMPROVING CELL AREA DISTRIBUTION

Barrera Sánchez P.* , Domínguez-Mota F. J† and González Flores G.*

*Departamento de Matemáticas, Facultad de Ciencias
Universidad Nacional Autónoma de México
Circuito Exterior de Ciudad Universitaria 04510, D.F., México
e-mail: pablobarrera@ciencias.unam.mx - guilmerg@yahoo.com
web page: http://www.matematicas.unam.mx/unamalla

† Facultad de Ciencias Físico-Matemáticas, “Mat. Luis Manuel Rivera Gutiérrez”
Universidad Michoacana de San Nicolás de Hidalgo
Ciudad Universitaria Morelia, Michoacán, México
e-mail: dmota@umich.mx

Key words: Numerical grid generation, discrete generation method, variational grid generation, convex area functional.

Abstract. In this work we present a new area functional that help us improve the cell area distribution in structured grids over plane irregular regions, which avoids very small and very large cell areas as much as possible. We present some results and an implementation of this functional in a preliminary version of the latest UNAMalla system.

1 Introduction

One of the main problems in numerical grid generation is to improve grid quality. In many practical cases, this is equivalent to modify the grid geometry in such a way that the cell area values are as less spread around the mean as possible.

Before producing quality grids, a central issue that needed to be solved was the variational generation of convex grids. An important contribution on this direction is due to S. Ivanenko, who proposed an ad hoc modification of Winslow’s functional for generating convex grids [10]; it was improved later by Barrera et al. in [5] by adding a general initialization procedure. A review on smoothness and convex functionals which have a strong control over the cell areas is presented in [3].

A discussion about the properties of the area functionals which are useful for generating convex grids can be found in [1] and [4]; however, all this functionals focus on controlling only the minimum cell area. A functional which controls both the minimum and maximum area cell values to produce quality grids was proposed in [2]. In this paper, we introduce
a new area functional which also controls the minimum and maximum area cell values but focusing only on those grid cells whose area values lie outside a control interval and, in consequence, improving the overall performance of the optimization process.

2 Problem formulation

In this section we describe briefly the main concepts required to present the new functional.

2.1 Basic theory for continuous mappings

Given a simply connected polygonal region \( \Omega \), it is possible to define a bijective mapping from the boundary of the unit square \( B = [0, 1] \times [0, 1] \) onto the boundary of \( \Omega \). If this mapping is bijectively extended from \( B \) onto \( \Omega \), then a grid on \( B \) defines naturally a grid on \( \Omega \).

For modelling purposes, we are mainly interested in simply connected domains \( \Omega \) defined by irregular boundaries; an example is shown in figure 1. For such domains, a fundamental aspect is to avoid grid folding in the optimization process; in other words, we require homeomorphic mappings as described in the following theorem:

**Theorem.** If \( x \) is a 2D mapping such that

- \( x : B \rightarrow \Omega \)
- \( x|_{\partial B} = \partial \Omega \)
- \( J(\xi, \eta) > 0, \quad \forall (\xi, \eta) \in B \)
then $x(\xi, \eta)$ is a homeomorphism from $B$ onto $\Omega$. For a proof of this theorem you can see Bobylev [7].

**Figure 2**: Mapping from a simple region onto a domain of interest.

Given a uniform mesh on $B$, the following theorem poses the conditions under which $x$ defines and unfolded grid for $\Omega$ as well as the guidelines for the selection of adequate mappings for meshing $\Omega$.

**Theorem.** Let the unit square $B$ be subdivided into $nc$ simple regions $B_i$ such that

1. $B = \bigcup_{i=1}^{nc} B_i$, \quad Int(B_i) \cap Int(B_j) = \phi.$
2. $x : B \mapsto \Omega$ is continuous.
3. $x_i$ is smooth on $B_i$,
4. $x_i = x|_{B_i}$,
5. $x : \partial B \mapsto \partial \Omega$ is a homemorphism,
6. $J_i(\xi, \eta) > 0$, \quad $\forall(\xi, \eta) \in B_i$, \quad $\forall i = 1, \ldots, nc$,

Then $x$ is a homeomorphism from $B$ onto $\Omega$.

**2.2 Discrete formulation**

Let $B$ be the unit square and $U(m, n)$ the uniform mesh of size $m \times n$ on $B$ given by

$$U(m, n) = \left\{ \left( \frac{i}{m}, \frac{j}{n} \right) \mid 0 \leq i \leq m, 0 \leq j \leq n \right\}$$

where

$$\partial U(m, n) = \partial B \cap U(m, n).$$

A discrete grid $G$ of size $m \times n$ on $\Omega$ is a mapping

$$G : U(m, n) \mapsto \mathbb{R}^2$$
such that

\[ G(\partial U) \subset \partial \Omega \]

and

\[ \partial G = G(\partial U) \subset \partial \Omega. \]

By considering a positive orientation on the boundary of \( \Omega \), we get an induced orientation on the boundaries of the cells \( c_{ij} = G(B_{ij}) \), and also on the four triangles defined by the cell vertices.

\[ \begin{align*}
& \text{Figure 3: The four triangles defined by the vertices of a quadrilateral cell.}
\end{align*} \]

Besides, will also say that \( G \) is convex if each one of the triangles has positive area and non degenerate except, possibly in the corners cells, see [3]. In the other hand, we are interested on controlling the convexity, smoothness and orthogonality of the grid cells by minimizing a suitable functional defined on the set of all the discrete grids on \( \Omega \).

Minimization is the basis of the variational grid generation method, which is one of the few methods that can be succesfully applied to produce a structured convex grid when the boundary of the domain \( \Omega \) is an irregular curve; the standard functionals have the form

\[ F(G) = \sum_{q=1}^{N} f(\Delta_q), \tag{1} \]

where \( f(\Delta_q) \) depends on the vertives if the triangle \( \Delta_q \), and \( N \) is four times the total number of grid cells since the four triangles defined by the four vertices of every grid cell are considered by Barrera et. al [1, 3, 4, 5, 6], and Ivanenko et. al [8, 9, 10].

When \( f \) is a function only of the areas of the triangles in the cells, \( F \) is referred to as an area functional. Its standard expression is given by

\[ F(G) = \sum_{q=1}^{N} f(\alpha_q), \tag{2} \]

where \( G \) represents the grid where the functional is evaluated and \( \alpha_q \) is the oriented area of the \( q \)-th grid triangle. As mentioned, there are four triangles in every grid cell defined by its vertices, as it can be seen in figure 3.
2.3 Variational setting

The variational grid generation problem can be posed as the minimization problem

\[ G^* = \arg \min_{G \in M(\Omega)} \sum_{q=1}^{N} f(\triangle_q) \]

defined on the set of admissible grids \( M(\Omega) \) for \( \Omega \).

To solve this problem numerically, some quantities must be defined first. For the generic grid cell triangle \( A, B, C \in \mathbb{R}^2 \) we define the length measure as

\[ \lambda(\triangle(A, B, C)) = \|A - B\|^2 + \|C - B\|^2, \]

the area measure

\[ \alpha(\triangle(A, B, C)) = (B - A)^t J_2 (B - C) = 2 \cdot \text{area}(\triangle(A, B, C)), \]

where

\[ J_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \]

and the orthogonality is measured with

\[ o(\triangle(A, B, C)) = (B - C)^t (B - C). \]

These quantities on the triangles define the classic area functional

\[ F_A(G) = \sum_{q=1}^{N} \alpha(\triangle_q)^2, \]

the classic length functional

\[ F_L(G) = \sum_{q=1}^{N} \lambda(\triangle_q), \]

and the classic orthogonality functional

\[ F_O(G) = \sum_{q=1}^{N} o(\triangle_q)^2. \]

There are also some useful linear combinations. For instance, the area-orthogonality functional [11]:

\[ F_{AO}(G) = \sum_{q=1}^{N} \left[ \frac{\alpha(\triangle_q)^2}{2} + \frac{o(\triangle_q)^2}{2} \right], \]
and the area-length functional

$$F_{AL}(G) = \sum_{q=1}^{N} \left[ \sigma \alpha (\Delta_q)^2 + (1 - \sigma) \lambda (\Delta_q) \right]$$

Some other functionals are continuous extensions of Winslow’s functional [10], like the smoothness functional

$$F_H(G) = \sum_{q=1}^{N} \frac{\lambda (\Delta_q)}{\alpha (\Delta_q)}$$

and the quasi-harmonic functional

$$F_{H\omega}(G) = \sum_{q=1}^{N} \frac{\lambda (\Delta_q) - 2 \alpha (\Delta_q)}{\omega + \alpha (\Delta_q)}$$

which was designed with a flexible barrier to approximate the harmonic functional. There is an existence theorem for optimal convex grids. In [1], it is proven that if \( f : \mathbb{R} \to \mathbb{R} \) is a convex ans decreasing positive function, then there exists \( \omega > 0 \) large enough, such that the minimizers of

$$S_\omega(G) = \sum_{q=1}^{N} f (w \cdot \alpha_q), \quad (3)$$

are convex grids.

For the implementation of an optimization process, we can even use a \( C^1 \) function \( f_\omega(\alpha) \) such that \( f \equiv 0 \) for \( \alpha \geq \alpha_l \) and \( f' (\alpha) < 0 \), for \( \alpha < \alpha_l \).

A function \( f \) which turned out to be very useful is given by

$$f(\alpha) = \begin{cases} 1/\alpha, & \alpha \geq 1 \\ (\alpha - 1)(\alpha - 2) + 1, & \alpha < 1; \end{cases} \quad (4)$$

a thorough discussion on the functional \( S_\omega \) defined by (4) can be found in [4]; it features a “mobile” barrier which is the main tool to generate convex grids (See fig. 4).

### 2.4 Combination of functionals

In order to combine different geometrical properties in the optimal grids, it is convenient to minimize linear convex combinations of \( S_\omega \) with a classic functional \( F_c \), where the latter is either the length, orthogonality of area-orthogonality functional:

$$F(G) = \sigma S_\omega(G) + (1 - \sigma) F_c(G).$$
Even though it is possible to generate convex grids by minimizing $S_\omega$, numerical experimentation has shown that in very irregular regions, the minimum value of $\alpha$ in the convex grids is very close to zero, an example of this phenomenon is sketched in figure 5.

Since this values of $\alpha$ are closely related to the accuracy of the numerical solution of partial differential equations using these convex grids, small values should be avoided if possible. This can be done using a convex function $f$ which focuses only on those cells whose minimum value of $\alpha$ is less than a threshold $\alpha_l$. In this way, the global values of smoothness and orthogonality are kept, whereas the area values are corrected (See fig. 6). This leads to propose the functional

$$f_l(\alpha) = A \left[(\alpha_l - \alpha)_+\right]^2$$

where $A > 0$ is a relatively large coefficient.

3 Two-sided area functional

Next, the idea presented in the previous section can be extended to avoid relatively large $\alpha$ values in some cells, i.e., we can impose an upper bound $\alpha_r$ for these values (See...
The new functional $F_b$ defined by $f_b$ can be used in two ways:

1) As a convex grid generator first giving a larger weight to a classic functional to reflect its properties and controlling the $\alpha$ values at the end.

2) As a cell corrector for those cell whose $\alpha$ values lie outside of $[\alpha_l, \alpha_r]$.

In other words, $F_b$ can be used either as a grid preprocessor or prostprocessor.

4 Combination of $F_b$ with classic functionals

$F_b$ can also be combined with the classic functionals $F_c$ of area, length and orthogonality

$$F(\omega G) = \sigma F_b(\omega G) + (1 - \sigma) F_c(G)$$
using an adequate normalization to reflect the properties of both functionals.

The optimization process produces a sequence of grids which converges to a convex grid. However, a large set of runs has provided enough empirical evidence to assert that the convergence rate is decreased in the last iterations due to the fact that $F_l(\omega G)$ becomes notably smaller than $F_c(G)$ since the former is positive only on small cells. Thus, to improve the convergence rate, an extra parameter $\sigma_{eq}$ is added to reduce the value of the classic functional. This yields the functional

$$F(\omega G) = \sigma F_b(\omega G) + (1 - \sigma)\sigma_{eq} F_c(G);$$

initially, we set $\sigma_{eq} = 1000.0$ and in the optimization process $\sigma_{eq}$ is updated accordingly to

$$\sigma_{eq} = \lambda \cdot \sigma_{eq}$$

where $0 < \lambda < 1$. Therefore, when $\sigma_{eq}$ is close to one, the optimal grids of $F(\omega G)$ are close to the optimal grids of the classic functionals; as $\sigma_{eq}$ decreases, $F_b$ becomes the important component in the combination.

It must be noted than $F(\omega G)$ can also be used either as a pre or postprocesor, in both bases with very satisfactory results.

The functional $F(\omega G)$ is implemented in UNAMALLA [12], and it has proven to be a useful tool for generating and improving convex grids.

4.1 Examples

An important technical issue in UNAMALLA is the fact that the optimal grids generated satisfy the condition that their minimum value of $\alpha$ is larger than a preset critical
value $\varepsilon$; in [4] this kind of optimal grids are referred to as $\varepsilon$-convex grid. The algorithm used for their generation is described in detail in [4].

Three examples of the grids generated by minimizing and the corresponding area distributions $F_b$ are shown in figures 9, 10 and 11.

5 Conclusions

The proposed functional, used as a pre or post grid processor, is a powerful for the generation of convex grids having area control: if a nearly-convex grid with strong area, smoothness of area-orthogonality properties, $F_b$ can modify the grid to force the $\alpha$ values to lie within the interval $[\alpha_l, \alpha_r]$ without losing the global properties we are looking for. The $\alpha$ histogram provides very useful information to decide how to select $\alpha_l$ and $\alpha_r$.

An important aspect is the adequate selection of the four boundary segments which represent the geometrical sides of $\Omega$, since this selection affects directly the properties of the grids and their cells.

Acknowledgements

We want to thank Posgrado en Ciencias Matemáticas of UNAM and CIC-UMSNH Grant 9.16 for the financial support for this work.
REFERENCES


Figure 11: Optimal bilateral grid for the Upper Arkansas subbasins.


AN ITERATIVE COUPLING APPROACH TO THE SIMULATION OF ELASTIC WAVE PROPAGATION

JULIETA ANTÓNIO *, ANTÓNIO TADEU* AND PATRÍCIA FERREIRA†

* Department of Civil Engineering, Faculty of Sciences and Technology, University of Coimbra, Pólo II, Rua Luís Reis Santos, 3030-788 Coimbra, Portugal
e-mail: julieta@dec.uc.pt, tadeu@dec.uc.pt- web page: http://www.dec.uc.pt

† Institute for Research and Technological Development in Construction Sciences – ITeCons, 3030-289 Coimbra, Portugal
email: patriciaferreira@itecons.uc.pt – web page: http://www.itecons.uc.pt

Key words: Elastic Wave Propagation, TBEM/MFS Iterative Coupling, Cracked Medium.

Abstract. This paper presents an iterative coupling formulation between the normal derivative integral equation (TBEM) and the method of fundamental solutions (MFS) for the transient analysis of problems involving elastic wave propagation in the presence of multi-inclusions. The proposed formulation overcomes the limitations of each method, requires less computer memory and may use less CPU time than a full direct coupling scheme requires. First, the formulation is presented and then its efficiency is assessed by computing the number of iterations and measuring the CPU time required to achieve the correct solution. Finally, the formulation is used to simulate the propagation of waves generated by a blast load in the vicinity of an empty cavity driven in a cracked medium.

1 INTRODUCTION

Modeling wave propagation in heterogeneous media with multiple inclusions is a challenge in many branches of engineering and science.

Some analytical approaches are known for inclusions of regular geometry, such as circular, cylindrical cylinders and spheres, [1, 2]. However, when inclusions have irregular and complex geometries the problem can only be solved using numerical methods such as the boundary element method (BEM), the finite difference method (FDM) or the finite element method (FEM).

But for a very large number of inclusions it is unfeasible to use the FEM and the FDM. Computation in regions containing complex scattering configurations are often simplified by using several smaller, separate, artificial boundaries, each enclosing a different obstacle [3].

The BEM is one of the most suitable techniques for modeling homogeneous unbounded systems containing irregular interfaces and inclusions since the far field conditions are automatically satisfied and only the boundaries of the interfaces and inclusions need to be discretized [4]. However, the BEM has some drawbacks in that it requires the correct evaluation of singular or hypersingular integrals and generates complex, fully populated linear
systems. Moreover, the classical formulation of the BEM degenerates to lead to ill-conditioned systems in the modeling of thin bodies or open boundary problems (zero thickness structures). In such cases the dual boundary element method or the normal derivative integral equation [5] (here called TBEM) must be used, which causes additional hitches in solving the resulting equations.

The meshfree methods that require neither domain nor boundary discretization have been subjects of recent research in many areas of computational science and approximation theory [6]. The MFS seems to be particularly effective for studying wave propagation since it overcomes some of the mathematical complexity of the BEM and provides acceptable solutions at substantially lower computational cost [7].

Coupling different methods to benefit from the advantages of each is one strategy researchers have adopted to improve the results with a view to speeding up analysis and ensuring efficiency, stability, accuracy and flexibility [8-10]. Even if the flexibility of the model is improved, the conventional (direct) coupling methods require the equations for all subdomains to be assembled into a single, global, equation system. Researchers have proposed a number of iterative methods to avoid having to assemble and solve a global, coupled equation system. Smaller and better conditioned systems of equations can be obtained by analyzing the subdomains separately [11-13].

In the present work the transient analysis of elastic wave propagation problems in the presence of multi-inclusions is undertaken by coupling the TBEM and the MFS. The problem is solved iteratively. At each step, each inclusion is solved individually, which leads to small system of equations and thus to reduced matrix storage requirements.

In the next sections the iterative coupling formulation is presented and its performance is assessed by comparing the CPU time with the times needed for a full coupling technique. The applicability of the proposed formulation is illustrated by solving physical systems that involve varying numbers of inclusions (cracks).

2 ITERATIVE COUPLING FORMULATION

Consider two empty irregular two-dimensional cylindrical inclusions, a crack and a cavity, embedded in a homogeneous elastic medium (Medium 1) with density \( \rho \) (Figure 1) and allowing longitudinal (P-wave) and shear waves (S-wave) to travel at velocities \( \alpha \) and \( \beta \), respectively (Figure 1).

The system is subjected to a dilatational line source placed at point O, \((x_s, y_s)\). The problem is solved using an iterative process. At each iteration step, each inclusion is solved individually, assuming there are no other inclusions present. The incident field is the scattered field generated by all the other previously solved inclusions. At the first iteration, the direct incident field generated by the source and exciting the field also needs to taken into account.

Iteration 0 - Step 1: The incident field only illuminates the crack, and the second inclusion is assumed to be absent. The traction boundary element method (TBEM) is used to model the crack and this formulation can be expressed by the following equation:

\[
aw^{(i)}(x_\phi, \omega) = - \int_{x_{i,1}} u^{(i)}(x, \omega) H(x, n_{s2}, x_\phi, \omega) ds + u_{inc}^{(i)}(x_\phi, n_{s2}, x, \omega)
\]

(1)

In this equation, \( i, j = 1, 2 \) correspond to the normal and tangential directions relative to the
inclusion surface. In these equations, \( \mathbf{n}_{a2} \) is the unit outward normal to the boundary \( S_i \) at the collocation points \( x_a, (x_a, y_a) \), while \( \mathbf{n}_{ai} \) is the unit outward normal along boundary \( S_i \), at \( x \), defined by the vector \( \mathbf{n}_{ai} = (\cos \theta_{ai}, \sin \theta_{ai}) \). The Green’s functions \( \overline{H}_j(x, \mathbf{n}_{ai}, \mathbf{n}_{a2}, x_a, \omega) \) are defined by applying the traction operator to the fundamental solutions for tractions \( H_j(x, \mathbf{n}_{ai}, x_a, \omega) \), in direction \( j \) on the boundary \( S_i \) at \( x \), caused by a unit point force in direction \( i \) applied at the collocation point, \( x_a \). \( u_j^{(0)}(x, \omega) \) corresponds to displacements in direction \( j \) at \( x \). The superscript used in \( u_j^{(n \omega)} \) indicates the number of the iteration. The incident field, \( \overline{u}_{inc}(x, \mathbf{n}_{a2}, x_a, \omega) \), can be evaluated in a similar way to the evaluation of \( \overline{H}_j \), in terms of stresses. As noted by Guiggiani [14] the coefficient \( a \) is zero for piecewise straight boundary elements.

![Figure 1: The geometry of the problem](image)

The boundary integral equation (1) can be solved by discretizing the boundary into straight boundary elements, with one nodal point in the middle of each element. The use of \( N \) boundary elements leads to a system of \([2N \times 2N]\) equations \( (B_{u}^{(0)} = u_{inc}^{(0)}) \),

\[
[-\overline{H}_j^{(i)}][u_j^{(0)}] = [-u_{j,inc}^{(0)}]
\]

(2)

where \( k, l = 1, N \), \( \overline{H}_j^{(i)} = \int_{C_i} \overline{H}_j(x_i, \mathbf{n}_{ai}, \mathbf{n}_{a2}, x_k, \omega) dC_j \), \( C_i \) is the length of each boundary element and \( u_{j,inc}^{(0)} = \overline{u}_{inc}(x_i, \mathbf{n}_{a2}, x_i, \omega) \).

The solution of this system of equations gives the nodal displacements \( u_j^{(0)} \) along the boundary \( S_i \), which allows the scattered displacement field to be defined at any receiver \( x_{rec} \),

\[
u_{r,01}(x_{rec}, \omega) = -\int_{S} u_{j,01}^{(0)}(x, \omega) H_j(x, \mathbf{n}_{ai}, \mathbf{n}_{a2}, x_{rec}, \omega) \, ds
\]

(3)

In this equation, the subscripts \( 01 \) in \( u_{r,01}^{(0)}(x_{rec}, \omega) \) define the iteration order (0) and identify the inclusion structure that produces it (1).

Iteration 0 - Step 2: The cavity is subjected to the direct incident field and illuminated by the scattered field generated at the crack after being submitted to the incident field (step 1). The cavity is modeled using the MFS that calculates the response of the inclusion as a linear combination of fundamental solutions simulating the displacement field generated by \( NS \) virtual sources. In the exterior elastic medium the scattered displacement fields are given by:
\[ u^{(0)}_j(x, \omega) = \sum_{n=1}^{NS} \sum_{j=1}^{2} \left[ a^{(0)}_{uj,n,ext} G_j(x, x_{n,ext}, \omega) \right] \]  

(4)

where \( G_j(x, x_{n,ext}, \omega) \) are the fundamental solutions which represent the displacements at points \( x \) in the medium, in direction \( i \), caused by a unit point force in direction \( j \) applied at positions \( x_{n,ext} \). \( n_{ext} \) are the subscripts that denote the load order number and \( a^{(0)}_{uj,n,ext} \) are unknown amplitudes of the virtual sources.

The amplitudes of the unknown virtual loads \( a^{(0)}_{uj,n,ext} \) can only be evaluated if null tractions are imposed along the boundary \( S_2 \) along the \( NS \) collocation points \( x_{col} \). This must be done taking into account the scattered field generated at inclusion 1, the crack, which can be regarded as an incident field that strikes the second inclusion

\[
\bar{u}^{(0)}_{ij;2} \left(x_{col}, n_{x_2}, \omega \right) = - \int_{S_2} u^{(0)}_j \left(x, \omega \right) \overline{H} \left(x, n_{a_1}, n_{a_2}, x_{col}, \omega \right) \, ds
\]

So equation (4) needs to be modified accordingly,

\[
\bar{u}^{(0)}_{ij;2} \left(x_{col}, n_{x_2}, \omega \right) + u_{inc} \left(x_{col}, n_{x_2}, x_i, \omega \right) + \sum_{n=1}^{NS} \sum_{j=1}^{2} \left[ a^{(0)}_{uj,n,ext} \bar{G}_j \left(x_{col}, n_{a_1}, x_{n,ext}, \omega \right) \right] = 0
\]

(5)

The Green’s functions \( \bar{G}_j \left(x_{col}, n_{a_2}, x_{n,ext}, \omega \right) \) are defined by applying the traction operator to \( G_j \left(x, y, x_{col}, y_{col}, \omega \right) \). In these equations, \( n_{x_2} \) is the unit outward normal to the boundary \( S_2 \) at the collocation points \( x_{col} \). This leads to a system of \( [2NS \times 2NS] \) equations \( \left[ \bar{G}_{ij} \right]=\left[a^{(0)}_{uj,n,ext}\right] \), which allows the unknown amplitudes \( a^{(0)}_{uj,n,ext} \) to be defined.

\[
\left[ -\bar{G}_{ij} \right] \left[ a^{(0)}_{uj,n,ext} \right] = \left[ -u_{inc} \right]
\]

(6)

where \( n = 1, NS \), \( u_{inc}^{(0)} = u_{inc} \left(x_{col}, n_{a_2}, \omega \right) + u_{inc} \left(x_{col}, n_{a_2}, x_i, \omega \right) \).

The scattered field at \( x_{rc} \) can be computed as

\[
u_{1,02} \left(x_{rc}, \omega \right) = \sum_{n=1}^{NS} \sum_{j=1}^{2} \left[ a^{(0)}_{uj,n,ext} G_j \left(x_{rc}, x_{n,ext}, \omega \right) \right]
\]

(7)

At the end of this iteration the total displacement field at the receiver would be

\[
u_{1} \left(x_{rc}, \omega \right) = u_{inc} \left(x_{rc}, x_i, \omega \right) + \sum_{n=1}^{M} \nu_{1,0n} \left(x_{rc}, \omega \right)
\]

(8)

where \( M = 2 \) (the number of inclusions).

Iteration k - Step 1: The first inclusion is only illuminated by the field scattered by the second inclusion in the conditions defined in iteration k-1 in Step 2. The incident field is the scattered field generated in the previous iteration by the second inclusion

\[
u_{1}^{(k)} \left(x_{g}, \omega \right) = - \int_{h} u_{j}^{(k-1)} \left(x, \omega \right) \overline{H} \left(x, n_{a_1}, n_{a_2}, x_g, \omega \right) \, ds + \bar{u}_{ij;2} \left(x_0, n_{x_2}, x_{n,ext}, \omega \right)
\]

(9)

The scattered pressure field at the receiver \( x_{rc} \) can then be calculated as
\[ u_{i,41}(x_{rc}, \omega) = - \int_{S} u^{(i)}_{j,41}(x, \omega) H_{j}(x, n_{a1}, n_{a2}, x_{rc}, \omega) \, ds \]  \hspace{1cm} (10)

Iteration k - Step 2: The second inclusion is now only illuminated by the field scattered by the first inclusion at Step 1. The stress field generated by the first inclusion in Step 1 is the only incident field that strikes the cavity which leads to

\[ \bar{u}^{(i)}_{1,12}(x_{rc}, n_{a2}, \omega) + \sum_{n=1}^{NS} \sum_{j=1}^{M} a_{n,j} \bar{G}_{j}(x_{rc}, n_{a2}, x_{rc}, \omega) = 0 \]  \hspace{1cm} (11)

At the end of iteration \( k \) the total displacement field at the receiver would be

\[ u_{i}(x_{rc}, \omega) = u_{i,inc}(x_{rc}, x_{rc}, \omega) + \sum_{i=r=0}^{M} \sum_{n=1}^{NS} u_{i,r,n}(x_{rc}, \omega) \]  \hspace{1cm} (12)

The iterative process continues until the contribution of the scattered field to the displacement at a certain receiver reaches a predefined threshold.

The proposed iterative coupling requires only the solution of the individual inclusions’ linear system of equations. Given the example used to illustrate the algorithm procedure, the two individual systems of \([2N \times 2N]\) and \([2NS \times 2NS]\) equations would only need to be solved once. The full coupling would require solving a system of \([2(N + NS) \times 2(N + NS)]\) equations.

3. PERFORMANCE OF THE PROPOSED FORMULATION

The performance of the proposed iterative coupling algorithm (MFS/TBEM) was verified by using it to solve the elastic field produced by a steady state blast line load emitting different excitation frequencies and placed in the presence of three circular empty cracks embedded in the vicinity of an empty cavity in an unbounded elastic medium.

The null-thickness, 90° arc-shaped cracks, are centered at (5.0 m, 20.0 m), have radii of 6.00 m, each has a length of 3 \( \pi \) m and they are equally spaced. The empty cavity is centered at (20.0 m, 9.0 m) and has a radius of 4.0 m.

Each crack is discretized as an open line with 200 TBEM boundary elements, while the cavity boundary is modeled using 160 virtual sources placed at 0.8 m from the inclusion surface, using the MFS.

The host elastic medium, with a density of 2200 kg/m³, allows P-wave and S-wave velocities of 1651.4 m/s and 1011.3 m/s, respectively. This system is illuminated by a wave field generated by a dilatational line load placed within the subdomain defined by the three cracks, at (0.0 m, 20.0 m). Displacements are computed over a grid of 18268 receivers arranged along the \( x \) and \( y \) directions at equal intervals and placed from \( x = -5.0 \) m to \( x = 25.0 \) m and from \( y = -10.0 \) m to \( y = 30.0 \) m.

Computations were performed for different excitation frequencies with a small imaginary part of the form \( \omega_{l} = \omega - i\eta \) (in which \( \eta = 0.7\lambda_{c}\omega = 0.7 \times 2\pi x 4 \)).

The CPU time was computed and compared with the times obtained for a full coupling formulation. Figure 2 shows the real and imaginary parts of the \( x \) and \( y \) – displacement field
obtained using the full coupling formulation for a frequency of $f = 200.0 \text{ Hz}$. Figure 3 shows the number of iterations and the CPU time required at each receiver for the coupling iterative formulation.

**Figure 2**: $x$ and $y$-displacements using the full coupling formulation for a frequency of $f = 200.0 \text{ Hz}$: a) real part; b) imaginary part. *average CPU time = 3.190 s*

**Figure 3**: Performance of the iterative coupling for a frequency of $f = 200.0 \text{ Hz}$: a) number of iterations; b) CPU time (s). *average CPU time = 2.591 s*

As can be seen in Figure 3, the number of iterations varies with the position of each receiver. Different simulations have been performed (not illustrated) where it is found that the CPU time and the number of iterations increase the greater the number of cracks. The number of iterations needed is higher when the excitation frequency is higher. As the frequency increment increases the number of iterations and CPU time decrease. This is because a smaller $\Delta \omega$ is associated with a larger time window, which accounts for a larger number of multi-reflections. In all cases the iterative coupling performs better than the full coupling.

### 4 TIME DOMAIN RESPONSES

To illustrate the applicability of the proposed algorithm, time domain displacements have been computed for the same geometry described in section 3.
This system is illuminated by a wave field generated by a dilatational line load placed in the subdomain defined by the three cracks, at \((0.0 \, \text{m}, 20.0 \, \text{m})\), modeled as a Ricker wavelet with a characteristic frequency of 500 Hz. The computations are performed in the frequency domain for frequencies ranging from 4.0 Hz to 2048.0 Hz, with a frequency increment of 4.0 Hz, which determines a total time window of 0.25 s. The resulting displacement is obtained over a grid of receivers arranged as described before.

A series of snapshots taken from computer animations is presented in Figure 4 to illustrate the resulting wave field at different time instants in terms of \(x\) - and \(y\) - displacement components \((u_x\) and \(u_y)\).

![Wave field snapshots](image)

**Figure 4**: Time domain displacements \(u_x\) (left column) and \(u_y\) (right column) for a characteristic frequency of 500 Hz a) \(t = 5.49\, \text{ms}\); b) \(t = 13.12\, \text{ms}\).

These displacement fields correspond to the incident field generated by the 2D source plus the scattered field generated by the thin cracks and empty inclusion.

The color scale adopted ranges from blue (lower displacement values) to red (higher
displacement values). The waves excited by the dilatational source first hit the surface of the crack that is furthest to the left. At $t = 5.49\text{ms}$ the wavefront has reached the other two cracks. Additional diffractions can be seen at the edges of the cracks, as well as additional reflections at their surface (see Figure 4a). The wave front has reached the rightmost gap between the cracks.

The wave energy trapped within the subdomain defined by the concave part of the cracks generates a complex wave field due to the multiple reflections, whereas the energy diffracted at the edges of the cracks spreads out through the gaps between the cracks, propagates away and reaches the surface of the empty cavity, whence it is reflected back (see Figure 4b at $t = 13.12\text{ms}$).

5 CONCLUSIONS

An iterative formulation coupling the TBEM and the MFS has been developed to solve elastic wave propagation within a domain that incorporates cavities and cracks. The cavities are modeled using the MFS while the TBEM was used to model the cracks. The iterative procedure enables computation of the scattered wave field generated by a large number of inclusions because it uses a series of systems of equations that are smaller than those required when a full coupling is applied.

The effectiveness of the proposed iterative coupling formulation was verified by computing the CPU time and comparing it with the time taken for a full coupling formulation. The CPU time and the number of iterations grow the higher the frequency and the greater the number of inclusions. The iterative coupling requires less CPU time than the full coupling and is more convenient when more cracks are modeled. The applicability of the proposed iterative formulation has been illustrated by computing the responses generated by the elastic wave propagation in the vicinity of a cavity and cracks.

REFERENCES

[8] Rüberg, T. and Schanz, M. Coupling finite and boundary element methods for static and


AUTOMATIC DIFFERENTIATION BASED FORMULATION OF COUPLED PROBLEMS

B. HUDOBIVNIK∗ AND J. KORELC †

∗ †Faculty of Civil and Geodetic Engineering (UL FGG)
University of Ljubljana
Jamova 2, 1000 Ljubljana, Slovenia
Email: ∗ blaz.hudobivnik@fgg.uni-lj.si, † joze.korelc@fgg.uni-lj.si, web page: http://www3.fgg.uni-lj.si/

Key words: Automatic Differentiation, Symbolic environment, Coupled Problems, Multi-physics Problems, Formulation Approach

Abstract. Our work will show that complex transient coupled problems can be formulated and solved effectively with AceGen and AceFEM using an automatic differentiation based formulation (ADB-formulation). From scalar pseudo-potential function consistent tangent matrix for strongly coupled problems can be derived, leading to quadratically convergent Newton-Raphson type procedure. Another problem considered is the implementation of finite element. Typically, all equations are written inside a single finite element and a single pseudo-potential is defined. Such implementation is efficient but rigid, therefore, a different implementation was considered. Within the second approach we wrote a separate finite element for each field, but in a way that quadratic convergent Newton-Raphson procedure is preserved. The paper presents examples where unified and field-by-field implementations are compared according to computational efficiency. The results show that with increasing ratio between the complexity of constitutive equations and discretization, generated code size and evaluation time of implementations become comparable.

1 INTRODUCTION

With the use of advanced software tools and technologies it is possible to reach automatisation of finite element method (FEM)[3]. Different approaches and tools have been developed which enable more effective problem solving. For the formulation of coupled problems one of the most important software tools is automatic differentiation [2]. The automatic differentiation procedure (AD) has two possible modes. The first is forward AD and the second backward AD. Numerical cost of the first is proportional to number of independent variables, and numerical cost of the second to the number of functions
differentiated [3]. The result of AD procedure is called “computational derivative” and is written as \( \hat{\delta}f(a) \). To establish a relation between different types of differentiations and result of automatic differentiation, definition of differentiating exceptions is introduced. For example, formulation
\[
\frac{\delta f(a, b(a))}{\delta a} \bigg|_{\frac{\partial b}{\partial a} = 0}
\]
determines that the “real” dependence \( b(a) \) defined by algorithm is ignored and \( b \) is taken as constant.

2 ADB FORMULATION OF GENERAL WEAK FORM AND PSEUDO-POTENTIAL FORM

For solving physical problems with FEM we need a weak form of differential equation to describe physical problem. Let us define a general weak form in a shape of
\[
\int\Omega \delta a(p) \cdot b(p) d\Omega + ... = 0
\]
(1)

Where \( a \) and \( b \) are tensors of arbitrary order, and \( \delta a \) is directional derivative of variation of tensor \( a \). Because \( \delta a \) is a fictive quantity, we cannot directly apply the automatic differentiation procedure. Thus it needs to be written in discretized form. Let \( p \) be a vector of unknown quantities of the problem and \( \delta a = \frac{\partial a(p)}{\partial p} \delta p \). Discretized form of general weak form is then written as
\[
\int\Omega \delta a(p) \cdot b(p) d\Omega + ... = \sum_{i=1}^{n_t} \left( \int\Omega \frac{\delta a(p)}{\delta p_i} \cdot b(p) d\Omega \right) \delta p_i + ... = 0
\]
(2)

From (2) follows \( n_t \) algebraic equations \( R \), which can be solved with standard methods, e.g. Newton-Raphson method. If the partial derivative (2) is replaced by computational derivative, the ADB formulation of weak form is obtained as follows
\[
R = \int\Omega \frac{\delta a(p)}{\delta p} \cdot b(p) d\Omega + ... = \int\Omega \frac{\delta W(p)}{\delta p} \cdot b(p) d\Omega + ... = 0
\]
(3)

Term \( \frac{\delta a(p)}{\delta p} \) is not appropriate for backward differentiation, because cost of differentiation grows linearly with the number of components of tensor \( a \). The uncontrolled growth of derived expressions can also occur [3]. Formulation (3) needs to be written in appropriate shape, so that as little as possible scalar functions are differentiated. To accomplish that, scalar product of tensors \( a \) and \( b \) is introduced, so-called pseudo-potential in the shape of \( W(p) = a(p) \cdot b(p) \). Result is so-called ADB formulation of discretized weak form:
\[
R = \int\Omega \left. \frac{\delta (a(p) \cdot b(p))}{\delta p} \right|_{\frac{\partial b}{\partial p} = 0} d\Omega + ... = \int\Omega \left. \frac{\delta W(p)}{\delta p} \right|_{\frac{\partial b}{\partial p} = 0} d\Omega + ... = 0
\]
(4)
In equation (4) the introduced differentiation exception \( \frac{D}{\partial p} b = 0 \) assures that automatic differentiation returns equations of problem, which correspond to weak form (3). In the formulation of finite element method, the contribution of individual element \( R_e \) to the global residual \( R \) is obtained by numerical integration over element domain, \( R_e = \sum_{g=1}^{n_e} w_g R_g \), where \( w_g \) is Gauss weight and \( R_g \) is residual calculated in each integration point and is written as

\[
R_g = J_g \frac{\delta W(p)}{\delta p} \bigg|_{\frac{\partial b}{\partial p} = 0}
\]

where \( J_g \) is Jacobian determinant.

### 2.1 General implementation of coupled problems in environments

**AceGen/AceFEM**

Let us take coupled problem, which is defined by \( n_c \) unknown fields \( \phi = \{ \phi_1, ..., \phi_{n_c} \} \). According to FEM, \( \phi_i \) is interpolated on element domain with the use of interpolation functions \( N_j; \phi_i^h = \sum_{j=1}^{n_j} N_j p_{i,j} \), where \( p_{i,j} \) is nodal the unknown of \( i \)-th field and \( j \)-th node in \( e \)-th finite element and \( n_j \) is the number of element nodes. The number of all unknowns of finite element is then \( n_p = \sum_{i=1}^{n_c} n_{p_i} \), where \( n_{p_i} \) is the number of unknowns used for discretisation of \( i \)-th field \( \phi_i \). That means the vector \( R_g \), which represents the contribution of \( g \)-th integration point to residual vector of \( e \)-th element \( R_e \), has \( n_p \) components and tangent matrix \( K_g \) has \( n_p^2 \) components. Consequently, the size of symbolically generated code of element grows with the square of the number of unknowns of the problem, which can become unmanageable with the large number of fields of the element. Complexity of the problem can be reduced in multiple ways.

Instead of explicit expressions for all components of \( R_g \) and \( K_g \), we can generate program code only for characteristic \( i \)-th component of residual \( (R_g)_{i} \) and characteristic \( i,j \)-th component of tangent matrix \( (K_g)_{i,j} \) (as described in [4]).

Additionally, we can generate equations that correspond to different fields or sets of fields of the problem, separately in separated finite elements, and later join them together when global residual and global tangent matrix are being assembled. With this in mind, a subset \( G_K \subseteq \{1, ..., n_c \} \) is defined. It determines fields \( \phi^{(K)} = \{ \phi_i : i \in G_K \} \), and their corresponding equations will be formulated inside K-th element source code. The unknowns of K-th source code are \( p^{(K)}_e = \bigcup_{i \in G_K} p^{(K)}_i \), where \( p^{(K)}_i \) is the vector of unknowns used to discretize field \( \phi_i \), \( p^K_e = \bigcup_{K=1}^{n_G} p^{(K)}_e \) is the vector of all unknowns on finite elements level and \( n_G \) is the number of subsets of \( \phi \). Residual and stiffness matrix belonging to subset \( G_K \) are defined as: \( R^{(K)}_g = \{ R^{(K)}_g : i \in G_K \} \) and \( K^{(K)}_g = [K^{(K)}_{g,j} : i \in G_K \land j \in \{1, ..., n_c \}] \). Accordingly \( R^{(K)}_g \) has dimension \( n_p^{(K)} = \sum_{i \in G_K} n_{p_i} \) and \( K^{(K)}_g \) has dimension \( n_p^{(K)} \times n_p^{(K)} \). Thus, \( K^{(K)}_g \) is not a square matrix.

According to ADB formulation of weak form shown in previous chapter (2), pseudo-potential of K-th element is defined as
\[ W_g^{(K)} = \sum_{l \in G_K} W_g^{(l)}(p_e) = \sum_{l \in G_K} \sum_m a_m^{(l)} \cdot b_m^{(l)}, \]  

where \( m \) goes through a number of terms of weak form that correspond to the \( l \)-th field. Residual and tangent matrix in \( g \)-th integration point of \( K \)-th element source code are given as

\[
R_g^{(K)} = J_g \frac{\delta W_g^{(K)}(p_e)}{\delta p_e^{(K)}} \Bigg|_{\frac{\partial b_m^{(l)}}{\partial p_e^{(K)}} = 0, \forall l \in G_K \land \forall m}
\]

and

\[
K_g^{(K)} = \frac{\delta R_g^{(K)}}{\delta p_e}
\]

**Algorithm 1** ADB formulation of \( K \)-th element source code

```plaintext
for \( g := 1 \) to \( n_g \) step 1 do
  \( J_g := \det \left( \sum \frac{\partial \delta \Sigma}{\partial \delta \Xi} \right) \)  // Calculation of Jacobian determinant
  \( b_m^{(l)}(p_e) := \ldots \)  // Definition of auxiliary functions \( b_m^{(l)} \), which will be constant when pseudo-potential is differentiated
  \( W_g^{(K)} := \sum_{l \in G_K} \sum_m a_m^{(l)} \cdot b_m^{(l)} \)  // pseudo-potential is defined as sum of scalar tensors \( a_m^{(l)} \) and \( b_m^{(l)} \)
  \( R_g^{(K)} = J_g \frac{\delta W_g^{(K)}(p_e)}{\delta p_e^{(K)}} \Bigg|_{\frac{\partial b_m^{(l)}}{\partial p_e^{(K)}} = 0, \forall l \in G_K \land \forall m} \)  // Differentiation with respect to unknowns of \( k \)-th set with appropriate exceptions
  \( K_g^{(K)} = \frac{\delta R_g^{(K)}}{\delta p_e} \)  // Differentiation with respect to all unknowns \( p_e \) of the problem
end for
```

In general we can separate formulations into two groups: formulations where all equations are written inside one element (unified approach) and formulation where equations for each field or group of fields are formulated in separated elements (field-by-field approach). In the first example it applies \( n_G = 1 \). \( R_g \) and \( K_g \) are evaluated in two generic formulas \( R_{gi} \) and \( K_{gi,j} \), which are both written inside one element. In the second case physically separated element is generated for each set \( \phi^{(K)} \). Each element has \( n_p^{(K)} \) unknowns, called primary variables of element \( p_e^{i \in G_K} \). In addition, element has access to other variables \( p_e^{\in \{1,...,n_c\} \setminus G_K} \) through secondary additional nodes. Because there are \( n_G \) residuals and \( n_G \) tangent matrices, we need to generate \( 2 \times n_G \) generic formulas in total to calculate residual and tangent matrix in all elements of sets. These generic formulas are usually simpler than the formulas in the first case, but at the same time some quantities, e.g. interpolation functions, need to be calculated several times.

3 Thermo-Hydro-Mechanical coupled problem

3.1 Weak form of THM coupled problem

At THM formulation porous medium is used, so such material model must be used that account for solid, water and air. Labels \( w \) and \( s \) relate to liquid and solid material
phases. Unlabelled material quantities refer to effective or averaged quantities of one-phase medium $\phi = \sum n^i \phi^i$, where $n^i$ is volume fraction and $\phi^i$ material quantity of phases $w$ and $s$. Weak form of balance equation for the case of NeoHooke type large strain material [8, 7, 5] can be given with equation:

$$\int_{\Omega} \delta E : S d\Omega - \int_{\Omega} \delta E : S^{pw} d\Omega - \int_{\Omega} \rho \delta u \cdot g d\Omega - \int_{\Gamma_u} \delta u \cdot \mathbf{t} d\Gamma_u = 0$$  \hspace{1cm} (8)

Where $S$ is total second Piola-Kirchoff stress tensor, $S^{pw}$ volumetric second Piola-Kirchoff stress tensor caused by water pore pressures $p_w$, calculated as $S^{pw} = -\alpha_b J \mathbf{C}^{-1} (p_w - p_0)$, where $J$ is the jacobian determinant, $\mathbf{C}$ the right Cauchy-Green deformation tensor and $\alpha_b$ Biot's coefficient. $E$ is Green-Lagrange strain tensor referring to initial configuration. Operator $":\"$ denotes matrix product $A : B = \text{tr}(A^T B) = \text{tr}(AB^T)$, also known as Frobenius inner product or matrix contraction. $I$ is identity matrix of dimension $3 \times 3$. Because plasticity is considered, hyper-elastic constitutive equations must be supplemented by additional non-linear coupled equations, forming relation between time dependent history variables $\mathbf{h}$ and constitutive equations. The constitutive model considering termo-plasticity is taken from [7] and supplemented with equations of effective stresses of [5].

Weak form of differential equation of non-stationary heat flow is given by [6] and contribution to temperatures from mechanical work by [7], and is written:

$$\int_{\Omega} \left( \frac{\partial T}{\partial t} \right) \delta T d\Omega + \int_{\Omega} k_T \nabla \delta T \cdot \nabla T d\Omega - \int_{\Omega} (Q + Q_w) \delta T d\Omega + \int_{\Gamma} \delta T \left( -D_{mech} + \mathcal{H} \right) d\Gamma - \int_{\Gamma} \delta T \left( q^T + \alpha_c (T - T_\infty) \right) d\Gamma = 0$$  \hspace{1cm} (9)

Where $T$ is temperature, $k_T$ effective heat conductivity, $Q$ internal heat source, $q^T$ surface heat source and $(\rho c)$ averaged heat capacity, where $c$ is specific heat capacity and $\rho$ density. Term $\alpha_c (T - T_\infty)$ accounts for heat convection and radiation on surface. Therefore, temperature on the surface tends towards temperature $T_\infty$, where $\alpha_c$ is pseudo convective coefficient. Conversion of mechanical elastic and plastic work into heat is covered by mechanical dissipation $D_{mech}$ and elastic-plastic structural heating $\mathcal{H}$, given by [7]. Term $Q_w = -\rho w c_w k_w \left( -\nabla p_w + \rho w g \right) \cdot \nabla T$ is added to inner heat source $q$, to capture the influence of liquid pressures.

Weak form of water flow on porous medium is taken from [6]. It is obtained from mass balance equation for compressible liquid water. The used governing equation of one-phase liquid flow in saturated porous medium is written as:
\[
\int_{\Omega} \delta p_w \left( \frac{\alpha_b - n}{K^s} + \frac{n}{K^w} \right) \frac{\partial p_w}{\partial t} d\Omega + \int_{\Omega} \frac{k_w}{\mu_w} \nabla \delta p_w \cdot \nabla p_w d\Omega + \int_{\Omega} \delta p_w \alpha_b tr \left( \nabla \frac{\partial u}{\partial t} \right) d\Omega
\]

\[
- \int_{\Omega} \delta p_w \alpha_t \frac{\partial T}{\partial t} d\Omega - \int_{\Omega} \frac{k_w}{\mu_w} \rho_w \nabla \delta p_w \cdot g d\Omega + \int_{\Gamma} \delta p_w \frac{q_w}{\rho_w} d\Gamma_w = 0
\]

(10)

Where \( p_w \) is water pressure in medium, \( n \) porosity of medium, \( K^w \) bulk modulus of water, \( k_w \) is intrinsic permeability and \( \mu_w \) dynamic viscosity of liquid. Term \( \int_{\Omega} \delta p_w \alpha_b tr \left( \nabla \frac{\partial u}{\partial t} \right) d\Omega \) in equation (10) contributes to change in pressures due to speed of deformation and term \( -\int_{\Omega} \delta p_w \alpha_t \frac{\partial T}{\partial t} d\Omega \) contributes to pressure changes due to temperature expansion of medium and liquid, where \( \alpha_t \) is average thermal expansion coefficient of medium.

3.2 Implementation of THM finite element

Weak form of coupled thermo-hydro-mechanical problem is defined with equations (8), (9) and (10). Variables of the problem are chosen as \( \phi = \{ u, v, p_w, T \} \), where \( u, v \) and \( w \) are displacements, \( T \) temperature of medium and \( p_w \) pressure of liquid in pores. For this paper we will consider three different implementations of THM finite element shown on example of 4-node isoparametric quadrilateral element (Figure 1).

\[
A.1) \phi^{(1)} = \{ u, v, p_w, T \} \\
B) \phi^{(1)} = \{ u, v \} \\
\phi^{(2)} = \{ p_w \} \\
\phi^{(3)} = \{ T \}
\]

Figure 1: Three different implementations of finite elements considered (A.1, A.2 and B)

A.1) First standard finite element implementation is chosen with four nodes and all equations defined in a single element. Vector of unknown scalar fields is therefore \( \phi = \phi^{(1)} = \{ u, v, p_w, T \} \). Pseudo-potential is written as a sum of pseudo potentials
of physical fields \( W = W^u + W^{pw} + W^T \) (defined later in chapter (3.3)) (“Unified approach”).

**A.2** Implementation A.2 is from the aspect of the derivation of \( R_g \) and \( K_g \) equivalent to implementation A.1. Difference between A.1 and A.2 is in organisation of element nodes. Separate node is defined for each physical field, totalling to 12 nodes.

**B** Three subsets of \( \phi \) are defined: \( \phi^{(1)} = \{ u, v \} \), \( \phi^{(2)} = \{ p_w \} \) and \( \phi^{(3)} = \{ T \} \). They are discretised by sets of unknowns of variables defined in three separate elements, each with its own vector of nodal unknown degrees of freedom \( p_i^e \). Pseudo potential is defined for each element, i.e. \( W^{(1)} = W^u \), \( W^{(2)} = W^{pw} \) and \( W^{(3)} = W^T \).

For implementations A.1 and A.2, \( K \) in \( R \) are derived in one element from single pseudo potential:

\[
R_g = \begin{cases} 
R^u_g & \\
R^{pw}_g & \\
R^T_g & 
\end{cases} 
\]

and

\[
K_g = \begin{bmatrix} 
K_{gu} & K_{gpw} & K_{gT} \\
K_{pugw} & K_{pwpw} & K_{pWT} \\
K_{Tug} & K_{Tpwpw} & K_{TTg} 
\end{bmatrix}
\]

In case B, there are three separate residuals \( R^u_g \), \( R^{pw}_g \) and \( R^T_g \) and tangent-stiffness matrices: \( K^u_g = \begin{bmatrix} K_{gu} & K_{gpw} & K_{gT} \end{bmatrix} \), \( K^{pw}_g = \begin{bmatrix} K_{pugw} & K_{pwpw} & K_{pWT} \end{bmatrix} \) and \( K^T_g = \begin{bmatrix} K_{Tug} & K_{Tpwpw} & K_{TTg} \end{bmatrix} \).

### 3.3 Pseudo-potential of THM problem

For each of the weak forms defined in chapter 3.1, a proper pseudo potential can be written, which is derived as sum of products of auxiliary functions \( b_{im} \) and functions \( a'_{im} \) (see chapter 2). Additionally, the ADB formulation of spatial gradient is defined as

\[
\nabla_x(\varnothing) = \left[ \frac{\delta \varnothing}{\delta \Xi} \right]_{\Xi = J^{-1}F^{-1}} = J^{-1} \hat{\delta}_{x}^{\Xi}
\]

where \( \Xi = \{ \xi, \eta, \zeta \} \) are reference coordinates which are expressed in shape functions \( N_i = N_i(\xi, \eta, \zeta) \) (e.g. \( N_i^Q = 1/4\{(1-\eta)(1-\xi), \ldots, (\eta+1)(1-\xi)\} \)). \( J_g = \frac{\delta X}{\delta \Xi} \) is Jacobian matrix of transformation from \( \Xi \) to \( X \) and \( x \) is vector of spatial coordinates. Material gradient can be derived similarly as \( \nabla_X(\varnothing) = \left[ \frac{\delta \varnothing}{\delta X} \right]_{\Xi = J^{-1}F^{-1}} = J^{-1} \hat{\delta}_{X}^{\Xi} \).

#### 3.3.1 Pseudo-potential of mechanical problem

Let us take weak form of balance equation (8). The corresponding pseudo-potential has the form of \( W^u_g = W_u + b_1^u : E - b_2^u \cdot u \), where \( b_1^u = S^{pw} \) and \( b_2^u = \rho g \) are auxiliary variables, which must be kept constant during automatic differentiation of \( R \). This will be achieved with appropriate ADB formulation shown below. \( W_u \) is elastic strain energy function, describing appropriate material model, in our case elastic and hyperelastic.
3.3.2 Pseudo potential of temperature conduction problem

Similarly as for mechanical problem, we define pseudo potential of non-stationary heat conduction from (9) in the shape of:

\[ W_T = b_{T1} \cdot \nabla_x (T) + b_{T2} T, \]

where \( b_{T1} \) and \( b_{T2} \) are auxiliary variables.

\[ b_{T1} = k_T \nabla_x (T) \quad (11) \]

\[ b_{T2} = \frac{(\rho c_p) \Delta t (T - T_p)}{(\nabla_x (p_w) + \rho_w g) \cdot \nabla_x (T) - D_{mech} + H - Q} \quad (12) \]

3.3.3 Pseudo potential of liquid flow

As for previous cases, the auxiliary variables are defined as:

\[ b_{pw1} = k_s \mu \nabla_x (p_w) - k_s \mu g \rho_w \quad (13) \]

\[ b_{pw2} = \alpha \beta (\frac{n}{K_s} + \frac{n}{K_w}) \frac{1}{\Delta t} (p_w - p_{wp}) - \frac{\alpha}{\Delta t} (T - T_p) \quad (14) \]

Pseudo potentials then follow from (10) as:

\[ W_{pw} = b_{pw1} \cdot \nabla_x (p_w) + b_{pw2} p_w. \]

3.4 ADB formulation of residuals of THM problem

In case of A.1 and A.2 implementations, there is one pseudo potential per integration point \( W_g = W_g^u + W_g^T + W_g^{pw} \). Residual of subsets \( G_K, R_g^{(K)} \) can then be written in the ADB formulation as:

\[ R_g = J_g \frac{\partial W_g}{\partial p_e} \bigg|_\delta \]

\[ \frac{\partial b_{g1}}{\partial p_{ce}} = 0, \frac{\partial b_{g2}}{\partial p_{ce}} = 0, \frac{\partial b_{T1}}{\partial p_{ce}} = 0, \frac{\partial b_{T2}}{\partial p_{ce}} = 0, \frac{\partial b_{pw1}}{\partial p_{ce}} = 0, \frac{\partial b_{pw2}}{\partial p_{ce}} = 0, \frac{\partial D_{mech}}{\partial p_{ce}} = 0, \frac{\partial H}{\partial p_{ce}} = 0, \frac{\partial Q}{\partial p_{ce}} = 0 \quad (15) \]

For case B, there are three separate pseudo potentials, one for each element and accordingly three residuals, written in the ADB formulation as:

\[ R_u^g = J_g \frac{\partial W_u^g}{\partial p_u} \bigg|_\delta \]

\[ \frac{\partial b_{u1}}{\partial p_{ce}} = 0, \frac{\partial b_{u2}}{\partial p_{ce}} = 0, \frac{\partial b_{u3}}{\partial p_{ce}} = 0, \frac{\partial b_{T1}}{\partial p_{ce}} = 0, \frac{\partial b_{T2}}{\partial p_{ce}} = 0, \frac{\partial D_{mech}}{\partial p_{ce}} = 0, \frac{\partial H}{\partial p_{ce}} = 0, \frac{\partial Q}{\partial p_{ce}} = 0 \]

\[ R_T^g = J_g \frac{\partial W_T^g}{\partial p_T} \bigg|_\delta \]

\[ \frac{\partial b_{T1}}{\partial p_{ce}} = 0, \frac{\partial b_{T2}}{\partial p_{ce}} = 0, \frac{\partial b_{T3}}{\partial p_{ce}} = 0, \frac{\partial D_{mech}}{\partial p_{ce}} = 0, \frac{\partial H}{\partial p_{ce}} = 0, \frac{\partial Q}{\partial p_{ce}} = 0 \]

\[ R_{pw}^g = J_g \frac{\partial W_{pw}^g}{\partial p_{pw}} \bigg|_\delta \]

\[ \frac{\partial b_{pw1}}{\partial p_{ce}} = 0, \frac{\partial b_{pw2}}{\partial p_{ce}} = 0, \frac{\partial b_{pw3}}{\partial p_{ce}} = 0, \frac{\partial D_{mech}}{\partial p_{ce}} = 0, \frac{\partial H}{\partial p_{ce}} = 0, \frac{\partial Q}{\partial p_{ce}} = 0 \quad (16) \]
4 EXAMPLES

4.1 Example introduction

Unified (A.1) approach and field-by-field (B) approach will be compared according to numerical efficiency on 56 simulations. Four constitutive material models will be used: linear elastic (LE), small strain elasto-plastic (LP), hyper-elastic (HY) and finite strain elasto-plastic (JC). Four combinations of physical fields "D", "DT", "DW" and "DTW" will be compared, where "D" is mechanical, "T" temperature and "W" hydrous field. Each model and example will be formulated for 2D domain using quadrilateral (Q1) and for 3D domain using hexahedral (H1) element with first order interpolation. All simulations are made using the same topology, initial and boundary conditions and material data, appropriately administered for 2D and 3D models and used fields.

The test example is a block with dimensions of $2 \times 3 \times 2$. 2D model has dimension of $2 \times 2$ with the thickness of 3 units, discretized with mesh of $80 \times 40 = 3200$ elements per subset $G_K$. The mesh of 3D model is discretized with the mesh of $16 \times 8 \times 16 = 2048$ elements per subset $G_K$.

4.2 Comparison of Computational efficiency

We compared the code size of elements defined in individual example. For A.1 approach only one element was defined per simulation, but for B approach two to three separate elements have to be defined. The total code size is therefore a sum of code sizes of element source codes of each subset $G_K$. Secondly, we are interested in evaluation time. We cannot compare different models directly. However, we can compare the time needed to assemble element tangent matrices $K$ and residuals $R$ of all subsets $G_K$. As above, this means that we must summarize the contribution of all elements defined for each subset $G_K$.

<table>
<thead>
<tr>
<th>Approach</th>
<th>A.1</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element</td>
<td>D</td>
<td>DT</td>
</tr>
<tr>
<td>Q1-LE</td>
<td>6457</td>
<td>10105</td>
</tr>
<tr>
<td>Q1-LP</td>
<td>14989</td>
<td>20223</td>
</tr>
<tr>
<td>Q1-HY</td>
<td>8755</td>
<td>13715</td>
</tr>
<tr>
<td>Q1-JC</td>
<td>54704</td>
<td>70666</td>
</tr>
<tr>
<td>H1-LE</td>
<td>20084</td>
<td>27136</td>
</tr>
<tr>
<td>H1-LP</td>
<td>34318</td>
<td>43459</td>
</tr>
<tr>
<td>H1-HY</td>
<td>25876</td>
<td>36212</td>
</tr>
<tr>
<td>H1-JC</td>
<td>112327</td>
<td>131085</td>
</tr>
</tbody>
</table>

First we can compare the code size of elements. The code size values are listed in Table 1. Growth of code can be observed with increasing number of fields, topology dimension and constitutive model complexity. All field-by-field approach simulations
have larger total code sizes than their unified counterparts. This is expected, because some equations that are identical in all elements must be calculated for each individual element separately. These equations can represent considerable percent of code, if element constitutive equations are simple, which causes larger differences between code of A.1 and B.

When code size of elements is compared for unified and field-by-field approach, it can be observed that for linear elastic and hyperelastic models the code size of approach B is larger for factors from 1.11 (Q1-LE: DW) to 1.44 (Q1-LE: DTW), but when plastic models are considered, increase in code size is for smaller values between 1.04 (H1-LP: DW) and 1.28 (Q1-LP: DTW).

Here the influence of the number of fields on code size is compared. When the code size of mechanical element is comparable to the code size of temperature and pressure elements (e.g. for non plastic models), adding all physical fields increases the code by values from 2.03 to 2.74 for approach A.1 and for approach B from 2.67 to 4, but when mechanical element is more complex (e.g. plastic models), the addition of physical fields increases the code by smaller values than before, in approach A.1 from 1.4 to 1.9 and B from 1.6 to 2.4. The complexity of material model has the largest effect on code size increase, because for finite strain elasto plastic models, the code size is by factors 3.4 (when all fields are present) to 8.5 (when only one field is present) larger than for linear elastic models.

A.1 and B approaches return identical results in the same number of steps and iterations. The number of steps was 133 for all examples, but the average number of iterations per step increased from 1 (linear elastic with one field) to 3.3 (finite strain elasto plastic with all fields), because the problem becomes non-linear. Once assembled, global matrices have the same dimensions and values in both A.1 and B approaches, only different organisation of degrees of freedom. Therefore, the main difference comes from assembly time of matrices. Normalized element assembly times are given in Table 2. Generally, element assembly times of tangent matrices and residuals for one iteration are smaller for
approach A.1. The difference in time comes from larger total code size of B as described above. The largest difference between A.1 and B is for models Q1-LE with two physical fields. With increasing complexity and number of physical fields the difference becomes lower and approaches 1. In some cases (e.g. "H1-HY: DTW"), the factor goes below 1, which is a consequence of optimisation procedure of source code, which works better on smaller source codes.

The results lead to the conclusion that formulation with field-by-field approach is appropriate to use for solving strongly coupled problems. A major disadvantage shows only in larger code size and consequently longer assembly, times when the number of physical fields is small and constitutive equations are simple, but on the other hand, when complex model is used with more fields, the assembly times become comparable.

5 CONCLUSIONS

We have shown that with an appropriate automatic differentiation based formulation of the problem and physically separated (field-by-field) implementations of equations, it is possible to efficiently describe coupled problems of arbitrary complexity in a way that quadratic convergent Newton-Raphson procedure is preserved. Proposed field-by-field implementation was also numerically compared to the standard unified. The code size is on average by 17% larger and element assembly time by 30% longer than that of unified, but with increasing complexity of material model (e.g. finite strain), and number of physical fields, the difference drops to 16% for size, while the assembly times become practically identical. Since the goal is to be able to solve strongly coupled problems with arbitrary number of physical fields, the difference will decrease further with increasing complexity of problem.

REFERENCES


Computational optimisation strategies of tailored fibre placement in polymer matrix composites based on local shear stress minimisation

T. RETTENWANDER*, M. FISCHLSCHWEIGER† AND G. STEINBICHLER*

* Institute of Polymer Injection Moulding and Process Automation
  Johannes Kepler University Linz
  Altenberger Straße 69, 4040 Linz, Austria
  e-mail: ipim@jku.at, www.jku.at/ipim

† Center for Lightweight Composite Technologies
  ENGEL Austria GmbH
  Steyrer Straße 20, 4300 St. Valentin, Austria
  Web page: www.engelglobal.com

Key words: Tailored fibre placement, computational optimisation, shear stress minimization, finite element method.

Abstract. The aim of this paper is to present optimisation strategies of Tailored Fibre Placement (TFP) in polymer matrix composites using FEM analysis with the principal stress criterion. In this work we will shed some light on the differences of typical engineering approaches which are state of the art for TFP and semi-consistent schemes where stress states and fibre orientation are coupled.

The criterion is presented in a coupled procedure, where several iterations lead to a good agreement of fibre directions and principal stress vectors. Furthermore some deficiencies of the applied optimisation scheme are addressed and proposals for overcoming these problems are given.

INTRODUCTION

Trees and their ability of adaptive growth have generated some major attention in engineering applications. Especially their ability to deal with their orthotropic material properties in an optimum way have led to an optimisation criterion how to adapt fiber alignment in composite materials.

Considering a publication of Mattheck [1] trees have the ability to align their fibres along the direction of the uniaxial force flow. Therefore trees exhibit an extraordinary structure were fibres are subjected only to beneficial loads, namely to tensile and compressive stress. Shear stresses are minimized. This also implies that this adaptation of the structure minimises the material usage to withstand external or internal loads.

In many composite structures the material with its anisotropic properties is not fully
exploited. Often the layup and stacking sequence with only straight fibres is a compromise of varying material angles to carry multiaxial stresses in the structure. Usually the stacking ends in a quasi-isotropic structure to ensure a multiaxial load carrying. With such a layup the outstanding mechanical properties due to the high anisotropy of for example one single layer with a unidirectional fibre orientation are not fully used. This leads to a limitation of weight and cost effectivity of the composite structure. Therefore optimisation of the material layup for different externally applied loadings is of high interest in composite structure engineering.

Optimisation techniques which are usually used for optimising composite structures have been addressed several times in literature [1, 2, 3, 4 and 5]. A recent review of several methods can be found in [6].

From our point of view a very promising approach to optimise fibre alignment in technical applications is to implement a criterion for minimising local shear stresses. This corresponds to the structural optimisation as this is done in nature by trees for example. Consequently the optimisation strategy requires the determination of the main stress trajectories with proper computational methods. Computer aided methods based on finite elements have been developed by Kriechbaum [7] and Reuschel [8] for this purpose. These methods calculate principal stresses in a linear plane stress analysis in a local point (Gauss point) according to an external loading, where the underlying material behaves purely isotropic. Then fibres are aligned in the direction of principal stresses and hence they should predominantly be subjected to tension and compression. Shear stresses should disappear in the composite and therefore failure due to excessive shearing should be omitted. Since this approach is a highly coupled problem between the structure’s stress state, the orthotropic material behavior and fiber orientation angle an iterative solution algorithm is chosen in these publications. Several plane stress analyses are carried out and fibre aligned in a reordered way following the principal stress trajectories of the end state.

However, for engineering approaches often the optimisation according to the above mentioned procedure is used with the neglection of an iterative stress state-fiber orientation coupling. In such a case good optimisation results can be obtained for simple structures, which are loaded such that even local stress states are characterized by a high pronouncement of a single stress trajectory. However, the more complex the geometry of the structure and the loadings are, the more problematic is the neglection of the stress state- fibre orientation coupling. [9]

This has been already demonstrated by Tosh et al. [10] by calculating a plate with defect (hole) where the plate is under tension. In the before mentioned publication also an example of a pin loaded whole is given. This increases the complexity and shows clearly that for this case a coupling of the stress state and the fibre orientation is required. The general example of a loaded plate with a hole has been investigated in [8] and [11] and shows potential for optimisation by principal stress trajectories. Crothers et al. [11] showed in experiments an improvement of the specific structure strength of 45 % by using the fibre steering technique.

The improvement in a laminate structure using above mentioned method can be seen by having a look at the failure body of the Tsai-Wu failure criterion [12], which is widely used for anisotropic composite materials. The body is displayed in Figure 1. By having a look at the cut section of the body stresses in fibre direction ($\sigma_1$) and perpendicular to the fibre direction ($\sigma_2$) form an ellipsoidal failure plane. The area of the failure plane is dependent on the state of shear stress level. By increasing the shear stress ($\tau_{12}$) level (indicated by the
The optimisation criterion with principal stress trajectories aims to avoid these shear stresses and therefore improve overall composite bearing capacity.

The aim of this paper is to shed some light on the optimisation strategy where the coupling of local stress states with the fibre orientation is taken into account. Therefore a complex structure with loading cases with a high potential for local multiaxial stress states is investigated. Special focus is put on studying the fiber realignment according to coupling effects over several coupling iterations. Also the number of iterations required for a converging system should be addressed. Furthermore the differences between optimisation results obtained by a typical engineering approach and those results obtained by the coupling strategy are compared and discussed.

In the current coupling approach the applied criterion uses the principal stress information of only one layer of a predefined composite laminate structure and hence updates an existing predefined laminate according to the orientation information of the reference layer. For complex structures the usage of both, major and minor principal stresses for orientation information seems of particular importance. This is done in the used optimisation procedure by the predefined layup.

**MODEL**

**Geometry Selection and Model Setup**

For the investigation of the principal stress trajectory approach a complex U shaped cantilever beam is investigated. Figure 2 illustrates this U Beam. The geometry represents an example of a more complex structure, where optimal fibre trajectories are less clear. The U Beam is fixed on one side and a displacement boundary condition is set on the other side as this is shown in Figure 2. For the later investigation three observer elements have been selected as reference and are labelled in Figure 2.
Figure 2: U Beam: Model geometry which is set up for the investigation of principal stress trajectories.

The mechanical properties used for the unidirectional (UD) fibre reinforcement (layer) are summarised in Table 1.

### Table 1: mechanical properties of a UD layer

<table>
<thead>
<tr>
<th>Property</th>
<th>Value (MPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young’s Modulus 0°</td>
<td>40000</td>
</tr>
<tr>
<td>Young’s Modulus 90°</td>
<td>8000</td>
</tr>
<tr>
<td>In-plane shear modulus</td>
<td>4000</td>
</tr>
<tr>
<td>Poisson’s Ratio</td>
<td>0.25</td>
</tr>
<tr>
<td>Ult. Tensile Strength 0°</td>
<td>1000</td>
</tr>
<tr>
<td>Ult. Tensile Strength 90°</td>
<td>30</td>
</tr>
<tr>
<td>Ult. Comp. Strength 0°</td>
<td>600</td>
</tr>
<tr>
<td>Ult. Comp. Strength 90°</td>
<td>110</td>
</tr>
<tr>
<td>Ult. In-plane shear strength</td>
<td>40</td>
</tr>
</tbody>
</table>

The stacking sequence has been chosen in the following way: both, the direction of the major principal stresses and minor principal stresses have been used to orientate a 0/90 laminate structure, where the laminate structure is kept as a symmetrical assembly (0/90)_s. The optimized aligned fibres have been stacked as well in a symmetric order with a constant ply thickness. No thickness variation has been applied to directly compare the optimized structure with the non-optimized (0/90)_s structure.

The first layer (cover layer) is the reference layer for all plies over the laminate following the major principal stress directions. The second and the third layer are perpendicular to the first layer with exact the same thickness and therefore following the minor principal stress directions. The fourth layer again follows the fibre direction from the first layer.

### Results and Discussion

The geometry is analysed regarding its potential to optimisation with the principal stress criterion. In total 15 iterations has been carried out whereupon the iteration algorithm has
been executed fully automatically, using FEM-analysis, post processing and rewriting the FEM input deck.

Figure 3 shows the evolution of fibre alignment of the first iterations starting from 0 degree layer to the orientated layer according to major principal stress trajectories in the following coupling iterations.

For the structure the evolution of the maximal local composite shear stress over the coupling iterations is shown in Figure 4. As it is expected the maximum shear stress is reduced by this approach, however reaching the minimum level of 0 shear stress fails.

In the next figure the overall composite failure calculated by the Tsai-Wu criterion [12] is illustrated for 15 iterations. Using the Tsai-Wu criterion allows coupling the stresses in the failure hypothesis, which is contrary to maximal stress criterion. An increasing global Tsai-Wu failure is detected for the local most critical element in the structure.
Referring to the above presented results there is some mismatch to the expected behaviour by the optimisation criterion. For clarifying this behaviour a detailed discussion of 3 elements across the structure is presented. The position of the elements in the structure is shown in Figure 2. Element 1 shows intermediate critical effects regarding the global failure behavior according to Tsai-Wu. Element 2 is a representation for a typical structure continuum without any critical effects. Element 3 is situated in a multiaxially loaded zone and shows the highest shear stresses in the structure.

For the three elements the Tsai-Wu Failure Criterion develops as shown in Figure 6. For elements 1 and 2 the trend of failure behaves to be nearly the same. There is a reduction in failure of the fibres aligned to the major principal stress (layer 1 and 4). However, the composite failure in layer 2 and 3 increases. These layers are aligned perpendicular to layer 1. For element 3 the failure criterion does not reach a constant level and failure behaviour deteriorates for all layers.

For interpretation of the achieved ply based failure behaviour a more detailed observation of the before mentioned elements is made. In element 1 the composite stresses develop as
shown in Figure 7. P1, Major is the major principal stress in the specific layer. Normal X is the stress in fibre direction, Normal Y is the stress perpendicular to the fibre direction and Comp XY is the shear stress. All stresses are given as averaged element stresses. It can be stated that the shear stress reduces to approximately zero in the element, which is in accordance with the goal of the applied optimisation approach. For layer 1 and 4 the stress in fibre direction totally adjusts to the value of the major principal stress, which is fairly high in this element. The fibres can easily handle this high stresses and have no negative effect on the failure behaviour of element 1 (see Figure 6). However the Normal X stress in layer 1 and 4 are directly coupled to the Normal Y stress in layer 2 and 3. Therefore it leads to high stresses transverse to fiber direction, which are higher than the maximum bearing capacity of the matrix-fibre interface and lead to failure of these layers.

To verify the optimisation criterion, the principal stress angle is calculated for the element, which should be reduced to zero along the iterations. For Element 1 the angle is reduced to a value close to zero in all layers, which is shown in Figure 8.

Same results are also achieved in element 2. In elements 1 and 2 the iteration progress already reaches a constant level after 2 iterations.

Figure 7: Composite stresses in element 1, layerwise inspection
In some regions the coupling of principal stress fibre alignment and shear stress reduction to zero is not valid. In the structure a specific level of shear stresses is sustained as shown in Figure 9. Considering element 3 possible reasons are investigated. The position of element 3 is shown in Figure 2.

For element 3 no specific optimum can be found along 5 iterations. Stresses in general show a fluctuation over the iterations (Figure 10). For element 1 the normal stress in fibre direction (Normal X) adjusts to the principal value. For the other layers across the thickness a deviation to the principal stress value is detected. This proposes that a deviation in the
principal angle is present, which is indicated in Figure 11. The optimum value zero of the principal angle is not reached in any layer of this element. Actually the element is highly multiaxially loaded and no convergence can be reached. The principal stress state also shows a very inhomogenous development across the thickness, where the major directions are varying. A small deviation in the principal angle (see layer 4) already results in shear forces in the composite. The optimisation by the principal stress information of only one layer (top layer) seems therefore detrimental to specific elements.

Figure 10: Composite stresses in element 3, layerwise inspection

Figure 11: Principal angle for element 3
Conclusions and Outlook

The application of principal stress optimisation criterion has been applied to a complex structure, wherein principal stress information is determined for one layer and the information is applied to a predefined stacking sequence of (0/90)\textsubscript{S} laminate with constant ply thickness. The aim to reduce composite shear stresses can be achieved in most parts of the structure with this procedure. Observer elements show these conclusions. In complex loaded elements stress fluctuations over the coupling iterations could be observed and convergence could not be achieved with the semi-consistent approach.

Also an increasing global failure level could be shown during the optimisation procedure because of the non-idealised orientation of the fibres of the plies referenced to the first layer.

An engineering approach by defining a (0/90)\textsubscript{S} laminate is detrimental for the full reduction of shear stresses in the laminate. Principal stresses vary across the thickness especially in multiaxially loaded areas. Therefore a layerwise changing vector information has to be applied to fully adjust fibre orientation to ply based acting stresses. Further it seems that a high potential for increasing the lightweight degree is a coupled optimisation of tailored fibre placement and thickness adjusting of the layer.

In this work the optimisation is based on a semi-consistent approach. Currently the authors work on a full consistent scheme where the stress state and fibre orientation are coupled in the local integration. Therefore the above stated problems occurring by using the semi-consistent approach should be solved with the full consistent optimisation strategy for TFP.

REFERENCES

Karlsruhe (1999).


DEVELOPMENT OF A MAXIMUM ENTROPY APPROACH FOR THE THERMOMECHANICAL MODELLING OF THE ROTARY FRICTION WELDING PROCESS

M. FOCA∗, G. RACINEUX† AND L. STAINIER†

∗Institut de Recherche en Genie Civil et Mecanique (GeM, UMR 6183 CNRS)
Ecole Centrale de Nantes
44300 Nantes, France
e-mail: mathieu.foca@ec-nantes.fr

†Institut de Recherche en Genie Civil et Mecanique (GeM, UMR 6183 CNRS)
Ecole Centrale de Nantes
44300 Nantes, France
e-mail: guillaume.racineux@ec-nantes.fr & laurent.stainier@ec-nantes.fr

Key words: Thermo-dynamical coupling, Maximum Entropy, Meshfree, Rotative Frictional Welding

Abstract. A multi-physics modelling of rotary friction welding process based on a Maximum Entropy approach is proposed. This approach will be able to solve coupled thermo-mechanical problems. Because strains are very high locally around the welded area, the remeshing time in a classical finite element method is very important. The use of this meshless method should reduce simulations time and the numerical diffusion phenomena.

1 INTRODUCTION

The Rotary Friction Welding (RFW) is used to assemble two parts one of which at least has a symmetry of revolution. It uses the thermical effects created by the friction between the rotating part and the fixed one. The heating power is resulting of resisting torque and rotation speed. The melting temperature is not reached because of the material creep during the welding and forging. One of the advantages of the RFW compared to other welding process such as resistance welding or laser beam welding is that the welded area has an exceptional quality. Nowadays, two kinds of rotary friction welding are mainly used: direct drive and inertial welding.

2 MODELLING

In order to obtain a good numerical modelling of the process, all thermal and mechanical aspects have to be taken care of. Therefore, the thermo-mechanical coupling
happening during the welding cannot be ignored. Indeed, there are very high strains and temperatures can reach high values near the welded area. So having a calculation code able to take care of all these parameters is very important. The whole study will be performed in axisymmetric 2D because of the symmetry of revolution of the moving part. To evaluate the efficiency of our method, the results we obtain will be compared to the results of an equivalent simulation performed on Abaqus, which will be considered as the industrial reference. For such a simulation, Abaqus remeshes the structured as soon as the upset is reaching a certain value: for example, it does not simulate more than 0.5mm with a same mesh. Then we obtain a mesh evolution showed on figure 2. The two main drawbacks of the use of finite element method is that remeshing takes a lot of time and results in propagation of numerical errors. This last aspect is really important in our context because an objective is to perform metallurgy analysis after the welding and it requires a very precise temperature field during the simulation.

The maximum entropy method has been preferred to other meshless method such as Smoothed Particled Hydrodynamics (SPH) or Element Free Galerkin because we have to deal with contacts and therefore to exactly define the edges.

The long-term objective is to be able to solve general thermo-mechanical problems. Bibliographic references [1] [2] shows that the method gives good results on mechanical problem with an adiabatic behaviour.

3 EXPERIMENTAL ASPECT

Experiments are lead in partnership with ACB on a RDS60 machine in order to compare the results we obtain by simulation. In a first time, P295GH steel parts are melt. It can be thick or thin tubes or solid cylinder. Once the influence of each parameter is set, titanium alloy TA6V will be weld.
The RFW process is made of four stages:

- **Phase 1**: Force increasing.
- **Phase 2**: Friction phase at constant force and rotation speed.
- **Phase 3**: Force increasing until the imposed forge force.
- **Phase 4**: Cooling under constant force.

The following parameters are set on the machine. Indicated values correspond to our first experimental test plan.

- **Force applied to the moving part**: between 35kN and 155kN.
- **Rotation speed** $V$: between 600 tr/min and 900 tr/min.
- **Material consuming** $U$: material quantity ejected during the welding, between 4mm and 8mm.

The objective is to compare the following aspects:

- **Heating time**.
• Material consuming speed.
• Temperature field.
• Influence of the forge force.
• Size of the Thermical Affected Area (TAA).

Moreover, the use of a thermal camera gives us the temperature field on the part surface around the welded area.

4 MAXIMUM ENTROPY METHOD

4.1 Origins

The MaxEnt method uses the entropy defined by Shannon [4] in the information theory. In this theory, entropy is defined as a measure of uncertainty. To find a solution maximizing the entropy means that we want to find the most uncertain solution. Here is an example. Let’s take a set of event \( A = \{A_1, ..., A_n\} \) linked to the probabilities \( \{p_1, ..., p_n\} \). Let’s consider the two following cases:

\[
B_1 = \begin{pmatrix}
A_1 & A_2 \\
0.5 & 0.5
\end{pmatrix}
\quad \text{and} \quad
B_2 = \begin{pmatrix}
A_1 & A_2 \\
0.9 & 0.1
\end{pmatrix}
\]

The most uncertain result is obviously obtained in the case \( B_1 \). Therefore the entropy is higher in case \( B_1 \) than \( B_2 \). Based on three criteria, Shannon [4] established the entropy \( H \) expression as following:

\[
H(B) = H(p_1, ..., p_n) = -\sum_{a=1}^{n} p_a \log p_a
\]

We obtain:

\[
H(B_1) = 0.301 \quad \text{et} \quad H(B_2) = 0.141
\]

4.2 Linear Maximum Entropy problem

Consider a set of distinct nodes \( X = \{x_a, a = 1, ..., N\} \subset \mathbb{R}^d \). The smallest convex space containing \( X \) is defined as following [5]:

\[
\text{conv} X = \{x \in \mathbb{R}^d | x = X\lambda, \lambda \in \mathbb{R}_+^N, 1.\lambda = 1\}
\]

where \( \mathbb{R}_+^N \) is the positive orthant, \( 1 \in \mathbb{R}_+^N \) is the vector full of 1 and \( X \) is a \( d \times N \) matrix in which columns represent the coordinates of nodes in \( X \). As \( X \) is finite, \( \text{conv} X \) is convex and compact.
Let \( u : \text{conv}X \to \mathbb{R} \) whom values \( \{u_a = u(x_a), a = 1, ... , N\} \) are known on \( X \). We want to build approximations of \( u \) (which may represent a displacement field for example) having the following form:

\[
u^h(x) = \sum_{a=1}^{n} p_a(x) u_a\tag{3}
\]

where functions \( p_a : \text{conv}X \to \mathbb{R} \) are the *shape functions*. Those functions must respect the following conditions:

\[
\sum_{a=1}^{n} p_a(x) = 1 \tag{4}
\]

\[
\sum_{a=1}^{n} p_a(x)x_a = x \tag{5}
\]

Arroyo and Ortiz [1] proved that it is possible to make a link between those relationship and Shannon’s entropy. In order to respect Jaynes [6] maximum entropy principle, the following problem (ME) must be solved:

\[
\text{(ME)} \quad \text{Maximise} \quad H(p) = -\sum_{a=1}^{n} p_a \log p_a
\]

such as \( p_a \geq 0, a = 1, ... , N \)

\[
\sum_{a=1}^{n} p_a(x) = 1
\]

\[
\sum_{a=1}^{n} p_a(x)x_a = x
\]

Moreover, in order to control the correlation degree between the shape functions values in \( x \) and values at close nodes, a locality notion has to be add. Then the following (RAJ) problem [7] must be solved:

\[
\text{(RAJ)} \quad \text{For} \; x \; \text{fixed, minimise} \quad U(x, p) \equiv \sum_{a=1}^{n} p_a |x - x_a|^2
\]

such as \( p_a \geq 0, a = 1, ... , N \)

\[
\sum_{a=1}^{n} p_a(x) = 1
\]

\[
\sum_{a=1}^{n} p_a(x)x_a = x
\]

To obtain the best compromise between the two aspects, the following problem must be solved:

\[
\text{(LME)}_\beta \quad \text{For} \; x \; \text{fixed, minimise} \quad f_\beta(x, p) \equiv \beta U(x, p) - H(p)
\]

such as \( p_a \geq 0, a = 1, ... , N \)

\[
\sum_{a=1}^{n} p_a(x) = 1
\]

\[
\sum_{a=1}^{n} p_a(x)x_a = x
\]
Arroyo et Ortiz [1] also proved that the solution to the (LME)\(\beta\) problem is unique and has the following expression:

\[
p_{\beta a} = \frac{1}{Z(x, \lambda^*(x))} \exp\left[-\beta |x - x_a|^2 + \lambda^* (x - x_a)\right]
\]  

(6)

where:

\[
Z(x, \lambda) = \sum_{a=1}^{n} \exp\left[-\beta |x - x_a|^2 + \lambda (x - x_a)\right]
\]  

(7)

\[
\lambda^*(x) = \arg\min_{\lambda \in \mathbb{R}^d} \log Z(x, \lambda)
\]  

(8)

\(\lambda\) correspond to Lagrange multipliers imposing (4) and (5) conditions. 

\(\beta\) must be set during the simulation depending on the interpolation degree we want to. \(\beta = 0\) means that the shape function support is \(\text{conv}X\) whereas \(\beta = +\infty\) means that the entropy aspect is neglected. For all \(\beta \in [0, +\infty]\) and for all \(x \in \text{conv}X\), the solution exists and is unique. The parameter \(\gamma\), linked to \(\beta\) as following, is generally used:

\[
\gamma = h^2 \beta
\]  

(9)

where \(h\) the characteristic length between two nodes.

5 FIRST RESULTS

In references, MaxEnt method has only be used to solve mechanical problem [1] [2]. Before solving coupled thermo-mechanical problems, we check that we have good results on simple thermal cases. Matlab code shared by Arroyo and Ortiz is used as base.

5.1 Shape functions visualisation

5.1.1 Shape functions in 1D

In this 1D example, we have five nodes at following coordinates:

\[x_a = 0.01a, \quad \forall a \in [0; 1; 2; 3; 4]\]

Here, \(h = 0.01\). The figure 3 represents the five shape functions for different value of \(\gamma\). Values in table 5.1.1 prove that conditions (4) and (5) are satisfied.

5.1.2 Shape functions in 2D

In the same way, shape functions in 2D are presented on figure 4.
Figure 3: Shape functions in 1D for different values of $\gamma$, $\gamma = \{0, 0.5, 1.8, 5\}$.

### Table 1: Shape functions values on a 1D example, $\gamma = 1.8$.

<table>
<thead>
<tr>
<th>Function</th>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 3</th>
<th>Node 4</th>
<th>Node 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function 1</td>
<td>1.000</td>
<td>1.247e-1</td>
<td>5.605e-4</td>
<td>6.828e-8</td>
<td>1.562e-27</td>
</tr>
<tr>
<td>Function 2</td>
<td>4.401e-5</td>
<td>7.511e-1</td>
<td>1.241e-1</td>
<td>5.558e-4</td>
<td>1.740e-18</td>
</tr>
<tr>
<td>Function 3</td>
<td>5.249e-11</td>
<td>1.236e-1</td>
<td>7.507e-1</td>
<td>1.236e-1</td>
<td>5.294e-11</td>
</tr>
<tr>
<td>Function 4</td>
<td>1.740e-18</td>
<td>5.558e-4</td>
<td>1.241e-1</td>
<td>7.511e-1</td>
<td>4.401e-5</td>
</tr>
<tr>
<td>Function 5</td>
<td>1.562e-27</td>
<td>6.828e-8</td>
<td>5.605e-4</td>
<td>1.247e-1</td>
<td>1.000</td>
</tr>
</tbody>
</table>
5.2 Resolution of a thermical problem

The following system must be solved in order to solve a thermical problem:

\[
[C]\{\dot{T}\} + [K]\{T\} = \{q_e\}
\]  

(10)

with \(T\) the vector of nodal temperatures, \(C\) the matrix of heat capacity, \(K\) the matrix of conductibility and \(q_e\) the vector of imposed flux.

5.2.1 Calculation of conductibility matrix \(K\)

The calculation of the matrix \(K\) is given by:

\[
\forall (a, b), K_{a,b} = \sum_{a=1}^{n} k \nabla p_{\beta a} \cdot \nabla p_{\beta b} dV
\]  

(11)

with \(k\) is the heat conductivity of the material, \(\nabla p_{\beta i}\) is the gradient of the shape function at node \(i\), obtained by derivation of the shape function \(p_{\beta i}\).

5.2.2 Calculation of the heat capacity matrix \(C\)

The calculation of the matrix \(C\) is given by:

\[
\forall (a, b), C_{ab} = \sum_{a=1}^{n} \rho C_v p_{\beta a} p_{\beta b} dV
\]  

(12)
with $\rho$ is the density of the material, $c_v$ is the specific heat of the material, $p_{\beta_i}$ is the value of the shape function at node $i$.

In order to compare with Abaqus results, the same $C$ matrix must be used. Abaqus is using a $C'$ matrix condensed on the diagonal as following.

$$\forall (i, j) \in [1, n]^2, C'_{ij} = \delta_{ij} \sum_{k=1}^{n} C_{ik}$$  \hspace{1cm} (13)

### 5.2.3 Numerical integration

We solve the system by a temporal integration using finite differences. In particular, we use the generalized method of the trapezius:

$$\begin{cases}
[C']\{\dot{T}_{n+1}\} + [K]\{T_{n+1}\} = \{q_v(t_{n+1})\} \\
\{T_{n+1}\} = \{T_n\} - (1 - \alpha)\Delta t\{\dot{T}_n\} + \alpha\Delta t\{\dot{T}_{n+1}\}
\end{cases}$$  \hspace{1cm} (14)

with:

- $\alpha = 0$ : explicit scheme
- $\alpha = 1/2$ : Crank-Nicholson scheme
- $\alpha = 1$ : Euler-implicit scheme

Then we use the Crank-Nicholson scheme.

### 5.3 Test 1D, comparison MaxEnt/FEM

In order to verify the validity of our model, we compare results obtained by the two methods on simple cases. Here, we only present one of the tests.

Consider a bar of an homogeneous material of length $L = 0.03m$. Finite elements mesh and MaxEnt nodes are defined on figure 5. In this example, we note $T_0 = 273K$. At first, the temperature of the bar is set to $T = T_0$ and then $\forall t \geq 0$, the temperature on one side is set to $T = 400K$: Initial conditions:

$$T(x,0) = 0, \forall x \in ]0, L]$$
$$\dot{T}(x,0) = 0, \forall x \in [0, L]$$

Boundary conditions:

$$T(0,t) = T_0, \forall t \geq 0$$

The figure 6 compare FEM and MaxEnt results. Results are very similar, which proves that MaxEnt method is valid.
Figure 5: FEM mesh (—) and MaxEnt nodes and integration points (●) for the bar example.

Figure 6: Comparison FEM (—) / MaxEnt (– · –) on a simple thermical problem.

6 CONCLUSION

In this work, an application of the maximum entropy principle for the resolution of coupled thermo-mechanical problems has been proposed. The consistency on mechanical problems [1][2] and on simple thermical problems encourages us to continue. After we implement the code to compute this method, we should be able to simulate the rotary friction welding process and we should have results accurate enough then to perform metallurgical studies.

Thanks

This work is a part of the Friction Rotative à Entrainement Direct (FRED) project, funded by OSEO, in collaboration with ACB, ENSAM (LAMPA) and Polytech’ Nantes (IMN).
REFERENCES


EVALUATION OF AUTOMOTIVE WEATHERSTRIP BY COUPLED ANALYSIS OF FLUID-STRUCTURE-NOISE INTERACTION

HEON YOUNG KIM*, HAK JIN KIM AND TAE HYUNG KIM

*Department of Mechanical and Biomedical Engineering
Kangwon National University
192-1 Hyoja 2-Dong, Chuncheon, Gangwon-Do, Republic of Korea
e-mail: khy@kangwon.ac.kr, web page: http://cae.kangwon.ac.kr

Key words: Door weatherstrip, Acoustic isolation, FSI analysis, Permanent deformation, SEA

Abstract. Automotive weatherstrip plays a major role in isolating the passenger compartment from water, dust and noise, etc. Among them, the wind noise through weatherstrip is the most severe factor making the passenger uncomfortable. Weatherstrip should be in contact between the door and the body frame, and sufficient contact area is needed to minimize the wind noise through weatherstrip. But there are several factors that make it difficult to ensure sufficient contact area. First, weatherstrip rubber deteriorates as time goes by and residual stress in the rubber becomes relaxed which results in the decrease of the contact area. Second, the gap between the door and the body increases due to pressure difference at high speed. In order to predict and reduce wind noise through weatherstrip, nonlinear behaviour of rubber at high speed and the effect of rubber deformation to wind noise should both be analyzed. In the paper, rubber deformation with time is obtained by hyperelastic and viscoelastic analyses, while the gap between the door and the body frame of the vehicle going at a high speed was predicted by the coupled analysis, Fluid-Structure Interaction (FSI). And also Statistical Energy Analysis (SEA) calculates the amount of wind noise numerically caused by rubber deformation under high speed condition.

1 INTRODUCTION

Car door weatherstrips have important functions like blocking noise transmission and foreign matter intrusion, as well as preventing vibration transmission. Recently, as the emotional perspective in vehicle performance evaluation has been emphasized, efforts to enhance sound insulation performance of door weatherstrip have increased (1-2). The double sealing method is currently used to improve sound insulation in vehicles that are semi-medium size or larger. The door seal, which is installed on the door frame, has sufficient pressure distribution with a deflection of approximately over 6 mm, so the impact it receives from door opening is small. However, the body seal, which is installed on the body frame, has a deflection of approximately 3 mm, so if the amount of door opening occurs more than that, the seal opens, reducing sound insulation performance.

Door opening refers to the phenomenon in which the door frame widens from inside out when pressure drops due to the vehicle’s external flow at high speed, and is greatly influenced by the angle of the A filler. That is, the larger the A filler’s angle, the bigger the change in...
external flow, which causes the door frame to widen. This paper introduces an analytical prediction method for the sound insulation performance of the door weatherstrip by considering door opening effects. Fluid-Structure Interaction (FSI) analysis was used to predict the deformation of a door weatherstrip, which occurs due to the pressure difference between the inside and outside of a vehicle at high speed. Statistical Energy Analysis (SEA) was applied to evaluate the sound insulation performance when the weatherstrip is opened. Also, FSI analysis was conducted by considering the aging and permanent deformation of EPDM rubber, which is the main material of the weatherstrip, and the influence that these characteristics have on sound insulation performance were evaluated.

Figure 1: Illustration of the door and target point.

2 WEATHERSTRIP DEFORMATION PREDICTION USING FSI ANALYSIS

FSI analysis (4) was applied to predict the weatherstrip deformation which occurs when driving at high speed. The analysis was conducted using ADINA®, a commercial program. The cross section of a door, including the weatherstrip, was selected as the analysis model, see Fig. 1. As the part where the A filler and loop area meets, the cross section A- A’ is where door opening most frequently occurs.

2.1 Finite Element Modeling

After conducting window installation and closing analysis, a two-dimensional cross section was modeled to predict the air flow when weatherstrip deformation occurs by dividing the vehicle into the car frame and seal and the fluid area. Car components like the body frame and seal used 1 mm 3-node elements and 4-node elements, whereas the fluid (air) area was modeled using 1 mm 3-node elements. The materials model of the seal applied the stress-strain curve obtained from the uniaxial tension test, and its behavior was expressed by considering the hyperelasticity model. Ogden model (4-6) was used as the energy equation, and the stress-strain curve is as shown in Fig. 2. Being air, a density of 1.983e-08 kg/mm3, and viscosity of 1.205e-12 N . s/mm2 was applied to the fluid area.
To considering door opening, the analysis was divided into the window installing procedure and door closing procedure, as shown in Fig. 3. Fixating the body frame and body seal, the installation analysis of the window was conducted in step 1, while door closing was simulated by moving window, glass run, and door seal in step 2. Finally, in step 3, to consider door opening, an analysis was conducted by moving models, which were moved when closing the door, on the Y axis from 0.5 mm to 3 mm at a 0.5 mm interval. Because the axis of flow direction and the axis of the analysis model cross section were different, the pressure value, instead of the fluid speed, was applied to analyze external flow following high speed driving. Using the data in the references, pressure was applied to the inside of the car as much as the amount reduced from the outside, and the amount of pressure change that occurs at the door loop area when driving at a high vehicle speed (180 km/h) is approximately -1000 Pa, as shown in Fig. 3 (3).

![Figure 2: Stress-strain curve of seals.](image)

![Figure 3: Boundary condition of the structure and fluid model.](image)
2.2 Analysis Result

When moving the door outwards at a 0.5 mm interval, the door seal and body seal both maintained contact up to 2.5 mm, but at the door opening condition 3 mm, the body seal was separated from the frame due to fluid flow, as shown in Fig. 4. Also, the fluid inside escaped through the open space between the body frame and body seal. Fig. 5 shows the result of measuring the reaction force of the body frame on the body seal during the analysis time. At the door opening condition 3 mm, contact was maintained, but when pressure was applied in the interior space, the reaction force was eliminated.

![Figure 4: Analysis result of door open condition. (3 mm-open)](image)

![Figure 5: Reaction force of normal body seal.](image)

3 WEATHERSTRIP WITH AGING OR PERMANENT DEFORMATION

To predict the permanent deformation, a weatherstrip was aged for long time in high temperature chamber. Compared to the stiffness obtained from the material tests of an initial product sample, the aged sample had approximately 40% lower stiffness (6). Also, to obtain the deformed door seal shape, the viscoelasticity analysis was conducted in time domain.

In case of the aged and permanently deformed model, the body seal was opened at door opening condition 2 mm. The reaction force was decreased by 50% compared to the initial state.
as shown in Fig. 7. Considering that the amount of permanent deformation on the body seal, it was confirmed that change in form, rather than the change in material stiffness, is a major factor of body seal opening.

Figure 6: Measurement of permanent deformation.

Figure 7: Reaction force of aging body seal.

4 EVALUATION OF SOUND INSULATION USING SEA

Generally, it is impossible to evaluate the response characteristics on the entire system using FEM and BEM because many modes exist in the high frequency range. Also, the accuracy of the results, which is based on element size and analysis time, decreases the analysis efficiency when using FEM and BEM due to using many elements (1,7). On the other hand, SEA is appropriate for noise analysis at high frequency range, which has large number of vibration modes. Also, it is a very practical method for analyzing the noise of the entire system, rather than individual parts which form the system. The basic principle of SEA is that it divides the analysis system into subsystems, and computes the energy transmission and loss between each system, as shown in Fig. 8 (1,7). Many vibration modes exist in the case of door seal sound insulation evaluation analysis, and because the goal is a sound insulation performance
evaluation at the high frequency range of over 1000 Hz, SEA was used (9).

As shown in Fig. 9, the models for evaluation of sound insulation were obtained from the final shapes of the FSI analysis for the initial state and permanently deformed body seals under the given opening condition. The sound insulation performances in these two cases were then compared to each other. The modeling of cross sections with various thicknesses was done by using several components. After generating a three dimensional model by extruding a two dimensional shape to a certain length in the machine direction, the spaces between the frame and door weatherstrip were defined as subsystems in the SEA model (7). Also, a junction was defined for each subsystem (9). External sound pressure was decided by using the sound pressure measured in the door loop area inside the car during the wind tunnel test. Because the measured interior sound pressure was 73.3 dB, the measured result and analyzed result of the sound pressure transmitted inside were identical when the external sound pressure was 112.7 dB. Therefore, when the body seal was separated from the frame, sound pressure transmitted to the interior was evaluated by setting the external sound pressure condition at 112.7 dB.

The SEA result is shown in Fig. 10. In the case of the shape after aging, where the body seal is apart from the frame, the sound pressure transmitted to the interior was 85.1 dB. As shown in Table 1, the noise transmissibility increased approximately 10% compared to the initial state, in which the body seal was in contact with the frame. Therefore, it was confirmed that the most important factor in sound insulation performance evaluation was whether or not the body seal and frame were in contact.

Figure 8: Boundary condition of the structure model.

Figure 9: SEA model.
Figure 10: Comparison of sound pressure for two scenarios.

Table 1: Comparison of sound pressure for two scenarios

<table>
<thead>
<tr>
<th></th>
<th>Seal close (Initial state)</th>
<th>Seal open (Aged state)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average sound pressure (dB)</td>
<td>73.3</td>
<td>85.1</td>
</tr>
<tr>
<td>Noise transmissibility (%)</td>
<td>65.0</td>
<td>75.5</td>
</tr>
</tbody>
</table>

5 CONCLUSION

This paper proposes an analytical method for evaluating sound insulation performance of a door weatherstrip by considering door opening. Door weatherstrip deformation caused by external pressure drop at high speed was identified through FSI analysis, and characteristics were compared to understand the effect of aging and permanent deformation of the material. Also, by using SEA to comparatively evaluate sound insulation performance when there is/isn’t body seal contact, it was found that whether or not there is contact with the frame following permanent deformation of the door weatherstrip is an important factor for sound insulation performance.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(no. 2010-0024134)
REFERENCES


LOW INTRUSIVE COUPLING OF IMPLICIT AND EXPLICIT INTEGRATION SCHEMES FOR STRUCTURAL DYNAMICS: APPLICATION TO LOW ENERGY IMPACTS ON COMPOSITE STRUCTURES

T. CHANTRAIT*, J. RANNOU* and A. GRAVOUIL†

*Onera, CC/DMSM/MNU
29 avenue de la division Leclerc
92322 Châtillon, France
e-mail: teddy.chantrait@onera.fr - Web page: http://www.onera.fr

†Laboratoire de Mécaniques des Contacts et des Structures (LaMCoS, CNRS UMR 5259)
Institut National des Sciences Appliquées de Lyon
20, avenue Albert Einstein
F69621, Villeurbanne cedex, France
e-mail: anthony.gravouil@insa-lyon.fr - Web page: http://lamcos.insa-lyon.fr

Key words: Codes coupling, Explicit, Implicit, Low intrusive, Composite, Low energy impact

Abstract. Simulation of low energy impacts on composite structures is a key feature in aeronautics. Unfortunately they are very expensive: on the one side, the structures of interest have large dimensions and need fine volumic meshes (at least locally) in order to capture damages. On the other side small time steps are required to ensure the explicit algorithms stability which are commonly used in these kind of simulations [4]. Implicit algorithms are in fact rarely used in this situation because of the roughness of the solutions that leads to prohibitive expensive time steps or even to non convergence of Newton-like iterative processes. It is also observed that rough phenomenons are localized in space and time (near the impacted zone). It may therefore be advantageous to adopt a multiscale space/time approach by splitting the structure into several substructures owning there own space/time discretization and their own integration schemes. The purpose of this decomposition is to take advantage of the specificities of both algorithms families: explicit scheme focuses on rough areas while smoother (actually linear) parts of the solutions are computed with larger time steps with an implicit scheme. We propose here an implementation of the Gravouil-Combescure method (GC) [1] by the mean of low intrusive coupling between the implicit finite element analysis (FEA) code Z-set and the explicit FEA code Europlexus. Simulations of low energy impacts on composite stiffened panels are presented. It is shown on this application that time step ratios up to 5000
can be reached. However, computations related to the explicit domain still remain a bottleneck in terms of cpu time.

1 INTRODUCTION

Low energy impacts can be very harmful in particular for composite structures used in the aerospace industry. In fact, they can cause significant damage (matrix cracking, fiber failure, delamination...) inside the composite or on the side opposite to the impact. However, the residual print it leaves on the impacted side can be almost undetectable to the naked eye. The damage caused can therefore lead to early failure of the structure while they can be unnoticed during a visual inspection, this is related to the concept of BVID (Barely Visible Impact Damage). We thus understand the importance for manufacturers to control such situations. Numerical simulations of this phenomenon can be a great help especially to orient and to rationalize tests campaigns by the use of virtual testing. Various researches are led in scientific and industrial communities to simulate these impacts but these one remain currently difficult to implement on an industrial scale because they are very difficult to manage. Indeed, sources of non-regularities introduced in the models (contact, softening damage laws, cohesive zone models...) make convergence difficult to achieve for implicit algorithms. Simulations performed in this context are therefore mostly conducted though explicit time integration [3, 4]. This make it possible to better take into account these sources of non-regularities but they require the use of time steps depending especially on the smallest mesh element to ensure stability. Moreover, very fine meshes are usually required (at least locally) to capture the non-linear phenomena occurring during impact. This thus leads to a very large number of increments which can be prohibitive. Note however that these non-linear phenomena occur on a very localized area around the impact point. Adopting a space/time multiscale strategy thus appears to be advantageous to solve this kind of multiscale problems. This can be performed through domain decomposition where each subdomain owns its own time discretization. The purpose of this decomposition is to focus on numerical computation where non-linear phenomena appear [5]. Explicit resolution in the area close to the impact is required because of the roughness of the solution. However, on the complementary area where the solution is smoother, implicit integration is appropriate. Larger time steps can then be used thus saving cpu time. The work done here is based on the Gravouil-Combescure (GC) method [1] and a low-intrusive coupling [6] between implicit finite element analysis codes Z-set\(^1\) and explicit FEA code Europlexus\(^2\) has been realized. The strategy is schematically shown in Figure 1. It shows one section of an impacted plate and a typical mesh. This mesh is divided into two domains: an impacted domain (center) which is processed by Europlexus with fine time-stepping and a complementary domain which is processed by

\(^1\)Z-set is developed by Mines ParisTech, Onera and NW Numerics & Modeling
\(^2\)Europlexus is developed by CEA and the Joint Research Centre in Ispra, Italy
Z-set with larger time steps.

2 OVERVIEW OF THE USED MULTISCALE SPACE/TIME METHOD

Different multiscale space/time approaches can be found in literature [1, 7, 8, 9]. These can be viewed as extensions of dual domain decomposition methods without overlapping, conventionally used for parallel computing [10]. Indeed, the domain decomposition methods consist in splitting spatially a structure into several subdomains and search solution on each subdomain as independently as possible (an example of structure split into two domains is shown in dotted box in Figure 2). A key point in domain decomposition methods is to determine the special boundary conditions that must be applied on the subdomains interface. This boundary condition should indeed ensure both kinematic continuity and interface equilibrium. An additional interface problem is then solved at each time step to determine this boundary condition and its size is determined by the spatial discretization of the interface. At this stage we may note that from the discrete point of view, imposing the continuity of the displacement, velocity or acceleration is not equivalent. The interface problem can then be solved directly after building interface operator (based on domains Schur complements) or be solved through iterative methods (e.g. conjugate gradient method) which do not require explicit construction of the interface operator.

The additional feature of multiscale space/time methods is that each subdomain have its own time discretization. A diagram is shown in Figure 2 where a fine time stepping is associated with the subdomain $\Omega_1$ and a coarse time stepping associated with the
subdomain $\Omega_2$. Let $\Delta t$ be the fine scale time step and let $\Delta T$ be coarse scale one. $m$ is the ratio between these two time steps: $m = \frac{\Delta T}{\Delta t}$. The interface problem has to be rewritten in order to satisfy the continuity and the balance conditions in time. The GC method proposes to ensure these conditions at every fine scale time step. This allows to naturally handle non-linearities in the explicit domain which is not the case for the other methods. Interfacial velocity field of the coarse time scale is not known at each fine scale step of the fine time scale, so it is linearly interpolated [2]. The interface problem to solve is only slightly modified in comparison with a dual domain decomposition methods with a single time scale. If the problem to solve on the implicit domain is linear, the interface operator remains constant during the computation. Therefore, because of the large number of steps to achieve due to the CFL condition, the interface operator worth being explicitly build.

**Figure 2**: Overview of multiscale space/time methods.

3 IMPLEMENTATION OF THE GRAVOUIL-COMBESCURE MULTISCALE SPACE/TIME METHOD

The particularity of the approach proposed here is to adopt a code coupling point of view rather than algebraic one. We thus achieve a structure/structure code coupling where each subdomain is associated with different computer codes which have their own characteristics. The advantage of this type of coupling is to extend the GC method to large problems while taking advantages of features existing in each FEA codes (contact algorithms, cohesive zone elements, material models...). The potential of the GC method is enhanced through application to a large low-energy impact problem (see section 4). The implementation of the method has led to the identification of the minimal entry points that codes must provide. With this minimal interface, the method can be easily implemented without being intrusive. Only the development of this interface could be
few intrusive. These entry points include:

- A method to apply a nodal force boundary condition at the beginning of each increment on area of mesh.

- A method to get nodal kinematic values on area of the mesh at the end of each step (at least the velocities).

- A method to validate or invalidate an increment.

In addition, each code must provide a trace operator.

This interface was developed in the two mentioned FEA codes (Z-set and Europlexus) through a Python layer as can be schematically seen in Figure 3. Python scripts containing MPI instructions needed to exchange data between the codes. These scripts contain the GC method and tools to solve interface problem are also written. The interface problems are solved in the script attached to the code handling the finest time discretization (i.e. the explicit code). This allows to minimize MPI communications. Different test cases were then carried out on 2D and 3D structures. Some proposed in [7] were reproduced in order to validate the implementation of the method.

4 APPLICATIONS OF THE MULTISCALE COUPLING METHOD ON STIFFENED COMPOSITE PANEL IMPACTED

Simulations with different time steps ratios were performed on a stiffened composite panel relatively representative of an aircraft’s subassembly. Figure 4 illustrates the geometry, the mesh and the domain decomposition adopted. The central impacted area of the panel (in blue) as well as the impactor are processed with Europlexus and the complementary part which has relatively larger time step is processed with Z-set. The stacking sequence is [90/45/0/-45]s for the stiffeners and [90/45/0/0/-45]s for the skin. The size of elements (hexahedrons and wedges) ranges from 5mm to 0.25mm in the impact area. Note that each ply is modeled with one element in the thickness direction.

The size of the problem is about one million degrees of freedom with 90% of nodes located in the implicit domain. The size of the interface is about 3000 degrees of freedom.
Figure 4: Example of domain decomposition on stiffened composite panel split in two areas (one implicit subdomain and nine explicit subdomains).

Figure 5: Elastic-damageable behavior illustrated through traction/compression in fiber and matrix direction.
A 10J impact at 10m/s is applied. Time step of the impacted area is set up to $2 \times 10^{-8}$s which is in accordance with the CFL condition. We then perform computations with different time step ratios with $m$ ranging from 500 to 5000.

A elastic-domageable material model based on Onera Progressive Failure Model (OPFM) \cite{11} is used on the explicit domain. Its parameters was identified on T700/M21 whose behavior is shown in Figure 5 for deformation controlled loading in traction/compression. One of the results obtained with this model for a time step ratio equal to 500 is shown in Figure 6. We observe qualitatively cone-shaped matrix damage typically observed on this kind of impact.

We present in figure 7 the time evolution of the composite panel deflection for two time step ratios ($m=500$ and $m=5000$). Errors are also represented, they are defined by the absolute value of the difference between the reference solution considered (here with a time step ratio $m=500$) and the current solution normalized by the maximum displacement. The solutions appear to be very similar over the whole time range. However we note some differences on the error at 0.0036s and 0.0052s. These differences may be explained by the large time step ratio used for $m=5000$: only 60 increments are performed which is very coarse. Moreover, impactors rebounds are observed there, they are thus only poorly captured. We can see in figure 7 that the error still remains relatively low.
5 CONCLUSION AND PERSPECTIVES

Structure/structure code coupling has been implemented using the Gravouil-Combescure algorithm. Both implicit (\textit{Z-set}) and explicit (\textit{Europlexus}) codes were involved. The minimal interface which should be provided by the FEA codes to implement this method were identified. We have also highlighted the potential of this method on large calculations. The cpu time distributions presented in Figure 8 were measured on stiffened panel simulations for different time step ratios (50, 500 and 5000) with a linear elastic model and a domain decomposition slightly different from the one presented.

We observe that the choice to solve the interface problem on the finest time step is not too penalizing in term of cpu time with this interface size (about 3000 degrees of freedom). We also note that the overall computation cpu time decreases when the time step ratio increases. This shows that in this case the computation work is focused on the explicit area. Indeed, for time step ratios \( m=500 \) and \( m=5000 \), times spent in the implicit process are very small (less than three hours for \( m=500 \) and one hour for \( m=5000 \)). However, this time saving tends to reach a limit which corresponds to the time spent in the explicit process. Work is currently underway to improve load balancing in order to reduce the computation time spent in the explicit process. Extensions that allow to use several explicit codes has been made. Calculations performed on the composite panel was made with these developments. The explicit part have been split in several subdomains as shown in Figure 4 (the explicit contains nine ring-shape subdomains centered around the point of impact). However, they all have the same time step (\( m=1 \)) which is equivalent in this case to perform a classical parallelization of the explicit code \textit{Europlexus}. This
parallelization is however carried out by means of the algorithm developed (outside of the code). We also continue to enrich the model of the impacted area with feature such as cohesive zone elements or full OPFM model in order to perform experimental and numerical comparisons.

**Acknowledgements**

The authors would like to thank Vincent Faucher from CEA and Steve Belon and Roland Ortiz from Onera/DADS for their help on Europlexus. This work have been performed in the framework of the Onera project “PRF transition statique/dynamique dans les matériaux composites”

**REFERENCES**


MODAL ANALYSIS OF THE FGM BEAM-LIKE STRUCTURES WITH EFFECT OF THE THERMAL AXIAL FORCE

JUSTÍN MURÍN*, MEHDI AMINBAGHAI†, JURAJ HRABOVSKÝ*, VLADIMÍR KUTIŠ*, JURAJ PAULECH*

*Department of Applied Mechanics and Mechatronics of IPAE, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, Ilkovičova 3,812 19 Bratislava, Slovakia
  e-mail: justin.murin@stuba.sk, juraj.hrabovsky@stuba.sk, vladimir.kutis@stuba.sk, juraj.paulech@stuba.sk
  web page: http://aladin.elf.stuba.sk/Katedry/KMECH/

† Institute for Mechanics of Materials and Structures, Vienna University of Technology, Karlsplatz 13, A-1040 Vienna, Austria
  e-mail: mehdi.aminbaghai@tuwien.ac.at

Key words: FGM beam finite element, Coupled multiphysics problems, Modal analysis, FGM actuator.

Abstract. The modal analysis of the FGM beam-like actuator is presented, where effects of the thermal axial force and the shear force are considered. The temperature load is assumed to be lower as the critical buckling temperature. The longitudinal variation of material properties has been assumed which can be caused by the varying constituent’s volume fraction and the temperature dependence of the constituent’s material properties. Our new FGM beam finite element has been used in the proposed analysis. An influence of the material properties variation and the thermal axial forces on the actuator eigenfrequency and eigenform has been studied and discussed.

1 INTRODUCTION

Mechatronic systems represent complex integrated intelligent systems making use of a synergy between information technology, electronics, mechanics, communication and control. Mechatronic is one of the most dominant research and application areas in nowadays' engineering, consumer electronics and services. For an optimal utilization of their enormous potential it is necessary to examine, analyze, model, control and optimize their structure and parameters for a wide range of applications. Development of dominant mechatronic parts like sensors and actuators still continues and is strongly dependent on the design of new materials and applications of modern approaches from the information, communication and control technologies.

Mechatronic systems work in the multi-physical domains. Based on the types of interaction they are divided into thermal-mechanical, electro-thermal, electro-magneto-mechanical, piezoelectric and fluid-structural systems. The modelling and control of these complex...
systems requires a continuous research and development of new and effective numerical techniques. The most widely used numerical methods in this area are the finite element method (FEM), finite volume method (FVM) and meshless methods. Besides the multi-physical domain character, progress in material engineering also plays a crucial role in the mechatronic systems design. It is, in particular, the ability to precisely define local material properties - functionally graded material (FGM) or change properties according to a controlling parameter (usually temperature) - shape memory alloys (SMA). However, the most important part in the mechatronic system design is their modelling and simulation.

The variation of FGM's material properties can be achieved via a controlled uneven mixing of two or more components e.g. using powder metallurgy, plasma spray applications, etc. or by a change of components' material properties through temperature. Such a material has much better efficiency than its components. Problems that occur at layer-interfaces of classical multi-layer composites are circumventing [1-3]. For these reasons it is of great interest to implement such new materials in the design of mechatronic parts, especially in the case of small dimensions where it is impossible to change mechanical and other physical attributes through a change in cross-section or complicated geometry (elastic joints, stiffness and dynamics properties etc.). A more intelligent and sophisticated function is reached with the implementation of composites made of FGMs. It is inevitable to create new advanced models and finite elements for their precise and effective multiphysical analyses [4], [7]. A number of international and domestic conference contributions emphasize this necessity [8], [9], [10].

Many papers dealing with modal analysis of single FGM beams can be found in literature, e.g. [11], [12], [13]. Mostly transversal variation of material properties has been considered. In papers [14], [15], spatial variation of material properties has been assumed. In [16], new 2D beam finite element has been established, which can be used in modal analysis of the beams made of FGM with transversal and longitudinal variation of macroscopic material properties. Effects of the internal axial force, shear force and elastic foundation have been taken into account.

In the contribution, which is an extension of the work [16], the new beam finite element will be used in modal analysis of the actuator which is built of FGM beams with longitudinal variation of material properties. The effects of the material properties variation and thermal axial force on the eigenfrequency and eigenform will be studied.

2 DESCRIPTION OF THE 2D FGM BEAM FINITE ELEMENT

This chapter is focused on description of the 2D FGM beam finite element, which is based on differential FGM beam equations of transversal and axial vibration. All quantities in following equations are the polynomial functions of \(x\). Homogenization process of the varying material properties and the calculation of other effective finite element parameters is fully described in [14].
2.1 Transversal free vibration

The main equations of the 2nd order beam theory containing the inertia effects (according to the Figure 1) are:

\[ R' = -q + kw - \mu \omega^2 w \]  
\[ M' = Q + m + \bar{\mu} \alpha \phi \]  
\[ \phi' = -\frac{M}{EI} - \kappa' \Rightarrow M = -EI \phi' - EI \kappa' \]  
\[ w' = \varphi + \frac{Q}{G A} \Rightarrow Q = G \tilde{A} w' - G \tilde{A} \phi \]

Eqs. (1) and (2) present the equilibrium equations for bending in the deformed configuration. Eqs. (3) and (4) are the constitutive relations of the 2nd order beam theory.

Here \( q \) is the distributed transversal load; \( m \) is the distributed bending moment; \( \kappa' \) denotes any applied beam curvature and \( k \) is the modulus of elastic Winkler foundation. Further, \( \mu = \rho A = \rho_L(x)A \) denotes the mass distribution; \( \bar{\mu} = \rho I = \rho_L(x)I \) is the mass inertial moment distribution, where \( \rho_L(x) \) is the homogenized mass density distribution. \( \omega \) is the natural eigenfrequency; \( R \) is the transversal force; \( Q \) is the shear force; \( M \) is the bending moment. The angle of cross-section rotation is \( \varphi \); \( w \) is the beam’s transverse displacement; \( B = EI = E^{MH}(x)I \) is the homogenized bending stiffness and \( \tilde{G}A = G_L(x)k^\prime(x)A \) is the reduced homogenized shear stiffness. \( I \) is the moment of inertia, \( A \) is the cross-section area, \( E^{MH}(x) \), \( E^{NH}(x) \) and \( G_L(x) \) is the homogenized elasticity modulus for bending, axial and shear loading, respectively. The calculation of the shear correction function \( k^\prime(x) \) is presented in [15]. The first derivative with respect to \( x \) is denoted by superscript (‘).
The relation between the transversal $R$ and shear $Q$ force is:

$$Q = -(\kappa + N''_{II})w' - N''_{II}\psi + R \quad (5)$$

where $N''_{II} \equiv N$ is the resultant axial force of the 2nd order beam theory (it has to be known and is caused by thermal loads in our case), $\psi$ is the beam rotation imperfection, and $\kappa = \kappa(x)$ is the elastic foundation modulus for the beam rotation.

Setting the expression (5) into the equations (1) – (4) we get:

$$\mu \omega - + \omega q = 0 \quad (6)$$

$$\mu \omega ^{2} - G\kappa \omega - G\kappa \varphi - M = 0 \quad (7)$$

$$E\varphi + M = - E\kappa \quad (8)$$

$$\kappa + N'' + G\kappa w' - G\kappa \varphi - R = - N''_{II}\psi \quad (9)$$

We get four coupled differential equations which can be solved (after common boundary conditions consideration) for the transfer functions: $w = w(x)$, $\varphi = \varphi(x)$, $M = M(x)$ and $R = R(x)$.

In the modal transversal vibration analysis the right side of the equations (6) – (9) is equal to zero. In the finite element derivation the reduced shear stiffness was simplified: $G\kappa \equiv G\kappa_{L}(x)k^{m}A$, where instead the shear correction function $k^{m}(x)$ [15] its average value $k^{m} = \frac{1}{L_{0}} \int k^{m}(x)dx$ have been applied (calculation of the average shear correction factor is described in [15]);

After some mathematical operations only one homogeneous differential equation of the 4th order of the homogenized FGM beam deflection with non-constant coefficients has been obtained

$$\eta_{0}w''' + \eta_{1}w'' + \eta_{2}w' + \eta_{3}w + \eta_{4}w' + \eta_{5}w = 0 \quad (10)$$

The non-constant coefficients $\eta_{0}$ to $\eta_{3}$ and appropriated parameters of the differential equation (10) are described in [14] in detail.

If the variation of all beam parameters is polynomial, the solution of this differential equation has a form [17]

$$\begin{bmatrix} w(x) \\ w'(x) \\ w''(x) \\ w'''(x) \end{bmatrix} = \begin{bmatrix} b_{0} & b_{1} & b_{2} & b_{3} \\ b_{0}' & b_{1}' & b_{2}' & b_{3}' \\ b_{0}'' & b_{1}'' & b_{2}'' & b_{3}'' \\ b_{0}''' & b_{1}''' & b_{2}''' & b_{3}''' \end{bmatrix} \begin{bmatrix} w_{1} \\ w'_{1} \\ w''_{1} \\ w'''_{1} \end{bmatrix} \quad (11)$$

where functions $b_{j}$, $b_{j}'$, $b_{j}''$ and $b_{j}'''$, ($j \in \langle 0,3 \rangle$) are the solution functions of the differential equation (10) and are called transfer functions. The dependence of the $w' = w'(x)$, $w'' = w''(x)$ and $w''' = w'''(x)$ on the $\varphi = \varphi(x)$, $M = M(x)$ and $R = R(x)$ is described in [14] from which the transfer matrix expression has been obtained:
The kinematical and kinetic variables at node $i$ are denoted by index “$i$” in (12). By setting $x = L$ in (12) the dependence of the nodal variables at node $k$ on the nodal variables at node $i$ will be obtained (see Figure 2).

2.1 Axial free vibration

The equilibrium equation for the axial vibration (according to Figure 1) and the constitutive equation of the FGM beam are:

$$N' = n + (k_x - \mu \omega^2)u$$ (13)

$$u' = \frac{N}{EA} + \varepsilon'$$ (14)

Here, $n$ is the axial distributed load; $N$ and $N'$ denote the axial force and its first derivative respectively. The modulus of elastic foundation in the axial direction is $k_x = k_x(x)$; $u = u(x)$ and $u'$ refer to the axial displacement and its first derivative. $\varepsilon'$ is the axial applied strain. $EA = E_L^{NH}(x)A$ is the homogenized beam stiffness in axial direction, and $\omega$ is the natural frequency.

By combination of the equations (13) and (14) we get the differential equation

$$\eta_2 u'' + \eta_1 u' + \eta_0 u = n$$ (15)

with non-constant polynomial coefficients: $\eta_2 = EA$, $\eta_1 = E' A$, $\eta_0 = \mu \omega^2 - k_x$. In the modal axial vibration analysis the right side of the equation (15) is equal to zero. The solution of the differential equation (15) for $n = 0$ can be expressed by transfer functions $\vec{b}_j$ and has the form:

$$\begin{bmatrix} u(x) \\ u'(x) \end{bmatrix} = \begin{bmatrix} \vec{b}_0 & \vec{b}_1 \\ \vec{b}_0' & \vec{b}_1' \end{bmatrix} \begin{bmatrix} u_i \\ u_i' \end{bmatrix}$$ (16)

The $\vec{b}_j$-functions ($j \in \{0,1\}$) are the solution functions of the differential equation (15). If the $u'(x)$ and $u(x)$ are replaced with the expression (14), we get:

$$\begin{bmatrix} u(x) \\ N(x) \end{bmatrix} = \begin{bmatrix} \vec{b}_0 & \vec{b}_1 \\ EA\vec{b}_0' & \frac{EA}{E_i A_i} \vec{b}_1' \end{bmatrix} \begin{bmatrix} u_i \\ N_i \end{bmatrix}$$ (17)

By setting $x = L$ in (17) the dependence of the nodal variables at node $k$ on the nodal variables at node $i$ will be obtained (see Figure 2). The transfer functions become the transfer
constants which can be calculated by a simple numerical algorithm [17]. \(E_i\) is the initial value of the homogenized elasticity modulus \(E^{NH}_L(x)\) at node \(i\).

### 2.1 Finite element matrix derivation

Figure 2 shows two nodal finite element with 6 degrees of freedom. The finite element equation in local coordinate system \(x, y\) (18) has been obtained by combination of the equations (12) and (17) and it has a form:

\[
\begin{bmatrix}
N_i \\
R_i \\
M_i \\
N_k \\
R_k \\
M_k
\end{bmatrix}
\begin{bmatrix}
B_{1,i} & 0 & 0 & B_{4,i} & 0 & 0 \\
0 & B_{2,2} & B_{2,5} & 0 & B_{2,5} & B_{2,6} \\
0 & B_{3,2} & B_{3,3} & 0 & B_{3,5} & B_{3,6} \\
B_{4,1} & 0 & 0 & B_{4,4} & 0 & 0 \\
0 & B_{5,2} & B_{5,3} & 0 & B_{5,5} & B_{5,6} \\
0 & B_{6,2} & B_{6,3} & 0 & B_{6,5} & B_{6,6}
\end{bmatrix}
\begin{bmatrix}
\bar{u}_i \\
\bar{w}_i \\
\bar{\varphi}_i \\
\bar{u}_k \\
\bar{w}_k \\
\bar{\varphi}_k
\end{bmatrix}
= \begin{bmatrix}
\bar{F}'_i \\
\bar{U}'_i
\end{bmatrix}
\tag{18}
\]

The non-constant terms \(\bar{B}_{i,j}\) (functions of \(\omega, N^{NH}, k^{sm}, \bar{k}, k, \bar{k}\) and other beam parameters) of the symmetric local finite element matrix \(\bar{B}'_{loc}\) are not expressed in detail here from space spending point of view. Those are calculated numerically. \(\bar{F}'_{loc}\) and \(\bar{U}'_{loc}\) is the vector of the local element forces and vector of the local element displacements, respectively.

The global finite element matrix \(\bar{B}'_{glob}\) is obtained by usual transformation of the local matrix \(\bar{B}'_{loc}\), \(\bar{B}'_{glob} = \bar{T}'^T \bar{B}'_{loc} \bar{T}'\). \(\bar{T}'\) is the well known transformation matrix, \(\bar{T}'^T\) is its transposed form. The global finite element equation reads

\[
\bar{F}'_{glob} = \bar{B}'_{glob} \bar{U}'_{glob}
\tag{19}
\]

where \(\bar{F}'_{glob}\) and \(\bar{U}'_{glob}\) is the vector of global forces and vector of the global displacements, respectively. Finally, the algebraic system of equations of whole beam structure will be established by a usual way.
The beam structure natural eigenfrequency $\omega_{Ki}$ (for the calculated thermal forces $N^\mu$ in the finite element) has been iterative calculated by software MATEMATICA [18]. The natural eigenfrequency $\omega$ will be increased until all the boundary conditions have been fulfilled. In this state, the natural frequency $\omega$ responds to the $i$-th natural eigenfrequency $\omega_{Ki}$. As the natural eigenfrequency is known, the eigenfrequency and corresponding eigenmode can be calculated by a usual way.

3 NUMERICAL EXPERIMENTS

The actuator (Figure 3) is loaded with thermal load caused by Joule heat. Undeformed and deformed form of the actuator is shown in the Figure 3a. Thermoelastic deformation induces the vertical displacement $\delta$ or the action force in the point $m$. Maximal action force arises when displacements at this point are restrained (Figure 3b). The design of actuator requires not only electro-thermo-structural analysis but also modal analysis. For its performance, mechanical model of beam structure according to Figure 3b have been considered. Three different analyses depending on the type of material have been analysed in order to find its influence on eigenfrequencies of the system. In the first two analyses actuator is the made from one chosen component, in the third one it is made by mixing of two components.

The actuator has been considered as the beam structure (shown in Figure 3b). It consists of 7 parts - beams. Their square cross-section is constant $b = h = 10\,\mu\text{m}$. Lengths of the parts are: $L_i = 300\,\mu\text{m}$, $i = 1 - 7$. The angles $\alpha_i$ and $\alpha_2$ are: $\alpha_1 = 70^\circ$, $\alpha_2 = 20^\circ$.

![Figure 3: The geometry of the actuator](image)

3.1 Case I – actuator with constant material properties

In this case two analyses have been made. Firstly, the actuator has been made only from aluminum Al6061-TO with constant material properties: the elasticity modulus $E = 69.0\,\text{GPa}$, the mass density $\rho = 2700\,\text{kgm}^{-3}$, the Poisson’s ratio $\nu = 0.33$, the coefficient of thermal expansion $\alpha_T = 23.5 \times 10^{-6}\,\text{K}^{-1}$. In the next analysis, the actuator has been made from titanium carbide TiC that constant material properties are: the elasticity modulus $E = 480.0\,\text{GPa}$, the mass density $\rho = 4920\,\text{kgm}^{-3}$, the Poisson’s ratio $\nu = 0.20$ and the coefficient of thermal expansion $\alpha_T = 5.9 \times 10^{-6}\,\text{K}^{-1}$. 
The actuator (Figure 3) clamped at the nodes \( i, k, r, p \) and simply supported at the node \( m \) has been studied by modal analysis. The first three eigenfrequencies have been found (see Table 1) using our new FGM beam finite element. The 1st order beam theory \( (N^I = 0) \) has been taken into account. Only one our new finite element was used for each part. The same problem has been solved using 10 BEAM3 elements of the FEM program ANSYS [19].

<table>
<thead>
<tr>
<th>Eigenfrequency [Hz]</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Al6061-TO</td>
<td>New finite element</td>
<td>ANSYS</td>
</tr>
<tr>
<td>1st</td>
<td>290682</td>
<td>291325</td>
<td>567956</td>
</tr>
<tr>
<td>2nd</td>
<td>392640</td>
<td>394158</td>
<td>767167</td>
</tr>
<tr>
<td>3rd</td>
<td>395534</td>
<td>397081</td>
<td>772821</td>
</tr>
</tbody>
</table>

As shown in Table 1, the values obtained by both finite elements agree very well with each other.

### 3.3 Case II – FGM actuator

The FGM actuator with the same geometry as in previous cases has been considered (as shown in Figure 3). Material of the beams consists of two components: aluminum Al6061-TO as a matrix and titanium carbide TiC as a fibre. Material properties of the components are constant (not temperature dependent), same as in previous experiments. There are considering two different longitudinal variation of the fibres volume fraction and have been chosen as the polynomial function of the local beam axis \( x \):

\[
a) \quad v_f(x) = 1 - \frac{1}{150} x + \frac{1}{90000} x^2 \\
b) \quad v_f(x) = \frac{1}{100} x - \frac{1}{30000} x^2
\]

that are shown in Figure 4. The first variation of the fibres volume fraction (denoted by a) has been considered in parts 1, 4, 5 and 7 (with initial point \( i, p, k \) and \( r \)) and the second variation of the fibres volume fraction (denoted by b) in parts 2, 3 and 6 (with initial point \( j, m, n \) and). According to Figure 4a-b zero values of the fibres volume fraction at the points \( j, l \) and \( n \) have been assumed.

![Figure 4: Fibre volume fraction variation](image-url)
The effective material properties of the homogenized beams (as a function of their local $x$-axis) have been calculated by the direct integration method [14] and we have got for the first a) variation of the fibres volume fraction (distribution of the elasticity modules are shown in Figure 5a):

$$E_{L}^{NH}(x) = E_{L}^{HH}(x) = 4.8 \times 10^8 - 2.74 \times 10^{12} x + 4.5666667 \times 10^{15} x^2 \quad [\text{kPa}]$$

$$G_{L}^{HH}(x) = 1.84615 \times 10^8 - 1.05385 \times 10^{12} x + 1.75641 \times 10^{15} x^2 \quad [\text{kPa}]$$

$$\rho_{L}^{H}(x) = 4920 - 1.48 \times 10^7 x + 2.466667 \times 10^{10} x^3 \quad [\text{kgm}^{-3}]$$

$$\alpha_{TL}^{H}(x) = \frac{0.01128 - 0.000072486x + 1.2081 \times 10^{-7} x^2}{480 - 2.74x + 0.004566667x^2} \quad [\text{K}^{-1}]$$

and for the second b) variation of the fibres volume fraction (distribution of the elasticity modules are shown in Figure 5b)

$$E_{L}^{NH}(x) = E_{L}^{HH}(x) = 6.9 \times 10^7 + 4.11 \times 10^{12} x - 1.37 \times 10^{16} x^2 \quad [\text{kPa}]$$

$$G_{L}^{HH}(x) = 2.65385 \times 10^7 + 1.58077 \times 10^{12} x - 5.26923 \times 10^{15} x^2 \quad [\text{kPa}]$$

$$\rho_{L}^{H}(x) = 2700 + 2.22 \times 10^7 x + 7.4 \times 10^{10} x^3 \quad [\text{kgm}^{-3}]$$

$$\alpha_{TL}^{H}(x) = \frac{0.0004071 + 0.000108729x + 3.6243 \times 10^{-7} x^2}{69 + 4.11x - 0.0137x^2} \quad [\text{K}^{-1}]$$

**Figure 5:** Homogenized elasticity modules

Because of only longitudinal variation of the constituents volume fraction in this case the homogenized elasticity modulus (for axial and transversal loading) are equal each other. The coefficients of thermal expansion $\alpha_{TL}^{H}(x)$ were not obtained as a polynomial so expansion to a Taylor series has to be used to convert them into the polynomial form.

The average shear correction factor [15] for all beams is $k^m = 0.83$ (constant Poisson ratio has been assumed for simplicity). The coupled modal analysis of the FGM actuator clamped at the nodes $i, k, r, p$ and simply supported at the node $m$ has been studied. The reference temperature is $T_{ref} = 20 ^\circ C$. The thermal forces $N^{H}$ in the beams and the critical buckling
temperature $T_{cr}$ have been calculated by the BEAM3 finite elements of the FEM program ANSYS [19]. A constant temperature load has been assumed on all parts of the actuator. Thermal axial forces have been evaluated for different temperature $T = 40, 60$ and $80^\circ C$ and then have been used as input axial forces in the modal analysis. Thermal forces evaluated for different temperature in the actuator are presented in the Table 3. The load temperatures in Table 3 are lower as the critical buckling temperature of the FGM actuator, which is of $128^\circ C$. So the pre-buckling thermal loading has been assumed.

**Table 3:** Thermal forces for different temperature

<table>
<thead>
<tr>
<th>Thermal forces [μN]</th>
<th>40 °C</th>
<th>60 °C</th>
<th>80 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>beam 1, 4</td>
<td>-6969</td>
<td>-13940</td>
<td>-20909</td>
</tr>
<tr>
<td>beam 2, 3</td>
<td>-4489</td>
<td>-8979</td>
<td>-13469</td>
</tr>
<tr>
<td>beam 5, 7</td>
<td>-5331</td>
<td>-10663</td>
<td>-15995</td>
</tr>
<tr>
<td>beam 6</td>
<td>-3077</td>
<td>-6155</td>
<td>-9232</td>
</tr>
</tbody>
</table>

The effect of the varying thermal axial force on the actuator eigenfrequency has been evaluated. The first three eigenfrequencies have been found for each set of thermal axial forces (see Table 4) using the new FGM beam finite element for modal analysis. Only one our new finite element was used for each actuator’s part. The same problem has been solved using a fine mesh – 1400 of BEAM3 elements (each element has different constant material properties) of the FEM program ANSYS [19]. The average relative difference $\Delta [%]$ between eigenfrequencies calculated by our method and the ANSYS solution has been evaluated.

**Table 4:** Eigenfrequencies of the FGM actuator

<table>
<thead>
<tr>
<th>Eigen-frequency [Hz]</th>
<th>T = 20 °C ($N'' = 0$)</th>
<th>T = 40 °C</th>
<th>T = 60 °C</th>
<th>T = 80 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>new finite element</td>
<td>ANSYS</td>
<td>$\Delta [%]$</td>
<td>new finite element</td>
</tr>
<tr>
<td>1$^{st}$</td>
<td>471059</td>
<td>472690</td>
<td>0.35</td>
<td>437598</td>
</tr>
<tr>
<td>2$^{nd}$</td>
<td>595574</td>
<td>592030</td>
<td>0.60</td>
<td>551678</td>
</tr>
<tr>
<td>3$^{rd}$</td>
<td>610157</td>
<td>604660</td>
<td>0.91</td>
<td>571254</td>
</tr>
</tbody>
</table>

The results, obtained for the thermal loading free state ($T_{ref} = 20$, $N'' = 0$) which are presented in Table 1 and Table 4, show the effect of mixture of the both components on the eigenfrequency. The eigenfrequencies of FGM actuator lies between the eigenfrequencies of actuator made of only one constituent.
The first three vibration eigenforms for thermal axial forces evaluated for temperature \( T = 60 ^\circ C \) are shown in Figure 6.

![Figure 6: The first three vibration eigenforms: T=60°C](image)

The effect of thermal axial forces evaluated for different temperatures (Table 3) on the eigenfrequencies is shown in Figure 7. As expected, the eigenfrequency decreases with increasing thermal load.

![Figure 7: The effect of thermal axial forces evaluated for different temperature on the eigenfrequencies](image)

### 4 CONCLUSIONS

Modal analysis of chosen actuator which is built of the FGM beams has been done by our new 2D beam finite element. Effects of the material properties and thermal axial forces on the eigenfrequency were analyzed. The temperature loads are lower than the critical buckling temperature.

The obtained results by this new finite element have been studied and compared with results obtained by a fine mesh of the BEAM3 finite element of the program ANSYS. The main additions of our new approach are:

- Eigenfrequency of the actuator is dependent on the operating load, which is caused by internal compressive axial forces in the respective beams;
- Eigenfrequencies of the system can be optimized by functional gradation of its material properties;
- Our new FGM finite element can be used in very efficient modal and buckling analysis of 2D mechatronic beam-like structures.
ACKNOWLEDGEMENT
This paper has been supported by Grant Agency VEGA, project No. 1/0534/12.
This paper has been supported by Grant Agency KEGA, project No. 015STU - 4/2012.
This paper has been supported by APVV 0450 – 10.
This paper has been supported by CRISIS ITMS 26240220060.

REFERENCES
[19] ANSYS Swanson Analysis System, Inc., 201 Johnson Road, Houston, PA 15342/1300, USA.
MODEL IDENTIFICATION AND ON-LINE OPTIMAL CONTROL OF FOOD PROCESSES

E. BALSA-CANTO, A. A. ALONSO, L. T. ANTELO, A. ARIAS-MÉNDEZ, E. LÓPEZ QUIROGA, D. RIVAS, C. VILAS

(Bio)Process Engineering Group, IIM-CSIC
C/ Eduardo Cabello 6, 36208 Vigo, Spain
e-mail: ebalsa@iim.csic.es, web page: http://www.iim.csic.es/~gingproc/

Key words: Food processing, coupled systems, model identification, reduced order modelling, dynamic optimisation, real time optimisation

Abstract. Food processes are coupled systems that involve heat, mass and momentum transfer together with kinetic processes related to quality and safety. This work is devoted to illustrate how model-based techniques offer the possibility to rationally optimise processes even in real time. The contribution is mainly based on our group experience and illustrates concepts with several examples such as the refrigeration of fruits, the deep-fat frying of potato chips, the freeze-drying of dairy products and the thermal processing of packaged foods. Coupled Problems 2013 Conference.

1 INTRODUCTION

Computer-aided simulation and model-based optimisation offer a powerful and systematic way to design and operate food processes. However, current industrial applications usually rely on simplified stationary models, the so called response surfaces, which are insufficient to describe the dynamic and distributed nature of food processing.

As an alternative, in recent decades there has been a growing interest in the development of rigorous models, based on first principles, that enable not only to perform experiments in silico, but to design and to optimise operation policies.

However, several problems arise: i) the complexity of food processes that include physical, chemical and biological phenomena on a wide range of time and space scales calls for advanced numerical methods; ii) the lack of information about specific processes and food-related thermo-physical and kinetic constants requires the implementation of identification loops incorporating parameter estimation, identifiability analyses and optimal experimental design; iii) the necessity of adapting the models to be used in real time decision making, calls for the development of techniques to obtain accurate and efficient
reduced order models and iv) the usual multimodality of the associated optimisation problems demands the development of robust and efficient global optimisation methods.

This contribution presents and describes methods to face such problems. Concepts are illustrated with the following examples: the refrigeration of fruits, the deep-fat frying of potato chips, the freeze-drying of dairy products and the thermal processing of packaged foods.

2 MODELING AND SIMULATION OF FOOD PROCESSING

2.1 Mathematical models

The mathematical modelling of food processing requires identifying and describing the physical, chemical, and biological changes experienced by the product during processing and storage. This means that, besides the heat, mass and momentum transfer mechanisms, the kinetics of quality and safety have also to be considered. As a conclusion, models of food processing typically consist of sets of coupled non-linear ordinary and partial differential (PDEs) equations:

\[ \Psi(z, z_\xi, z_t, u, \theta, t) = 0 \]  
\[ z(\xi, t_0) = \Psi_0(z(\xi, t_0), u(t_0), \theta, t_0) \quad ; \quad B(x, x_\xi, u, \theta, \xi, t) = 0 \]  

where \( \xi \in \Omega \subset \mathbb{R}^3 \) are the spatial variables, \( z(\xi, t) \in Z \subset \mathbb{R}^\nu \) are the state variables (temperature, water content, etc.), \( z_\xi = \partial z / \partial \xi \), \( z_{\xi \xi} = \partial^2 z / \partial \xi^2 \), \( z_t = \partial x / \partial t \), \( u \in U \subset \mathbb{R}^\sigma \) are the control variables and \( \theta \in \Theta \subset \mathbb{R}^\eta \), time independent parameters (thermo-physical or kinetic related constants). Eqs. (2) represent the initial and boundary conditions.

2.2 Numerical simulation

The finite element method (FEM) in combination with a suitable implicit time integration scheme is possibly the standard to handle complex geometries, and non-linear sets of PDEs in food processing simulation. In fact, commercial general purpose implementations, such as COMSOL\textsuperscript{©}, are the most popular choice [1].

These software tools allow testing “what-if” scenarios to predict food product characteristics given particular operation conditions. However, the associated computational cost may be unreasonable for the purpose of design, optimisation and control. This calls for the development of methods to derive reduced order models (ROMs) which preserve the predictive capabilities of the full model and can be solved very efficiently.

2.3 Reduced order models

Reduced order models based on the projection of the original PDE system over a set of global spatial basis functions have emerged as efficient alternatives to classical discretisation techniques. In particular, the proper orthogonal decomposition (POD) was successfully applied in the context of food processing [2–4]. In this approach, each of the state variables is approximated by a truncated series of the form:
\[ z(\xi,t) \simeq \tilde{z}(\xi,t) = \sum_{j=1}^{p} m_j(t) \phi_j(\xi) \] (3)

and the terms \( \phi_j \) are obtained by solving the following eigenvalue problem:

\[ \lambda_j \phi_j(\xi) = \int_V K(\xi,\xi') \phi_j(\xi') d\xi' \quad K = \frac{1}{k} \sum_{n=1}^{k} Z_n Z_n^T \] (4)

being \( \{Z_n\}_{n=1}^{k} \) a set of measurements (snapshots) representative of the system’s behavior. POD basis functions can be different for each state variable (separate basis approach) or can be unique for all state variables (joint basis approach) [4]. The joint basis approach is particularly useful when the states are highly coupled.

The coefficients \( m_j \) are computed as the solution of the ODE system resulting from the projection of Eqs. (1)-(2) over the POD basis. Details on how the FEM structure can be exploited to numerically compute projections can be found in [5].

3 MODEL IDENTIFICATION

Thermo-physical properties vary considerably among food products, within products (anisotropic, heterogeneous,...) or even with the process itself. However thermo-physical properties are only a part of the story. We also need to know the rates of reactions for biochemical (nutrient, color, flavor, etc.) or microbial changes.

Even though some thermo-physical parameters may be found in the literature, most of the model parameters are not known and can not be measured. In this regard, their values must be estimated from experimental data in the so call parametric model identification which comprises the following steps: parameter estimation, identifiability analysis and optimal experimental design.

3.1 Parameter estimation

The parameter estimation problem is usually formulated as a nonlinear optimisation problem where the objective is to find the parameters that minimise \( J_{wlsq} \), i.e. the sum of the weighted least squares of the residuals between the model predictions and the experimental data. The weights can be related to experimental error.

The problem is particularly challenging due to the usual presence of multiple suboptimal solutions (multimodality), or multiple equivalent solutions (lack of or poor identifiability) [6]. Multimodality may be overcome by the use of global optimisation methods. Lack or poor practical identifiability may be reduced (or even eliminated) by model-based optimal experimental design.
3.2 Identifiability analysis

In order to assess the quality of the parameter estimates, several possibilities exist. Bootstrap or jack-knife approaches allow to compute robust confidence intervals. However, the associated computational cost makes it difficult to use these methods for large scale models. Alternatively, the confidence interval of $\theta_i^*$ may be obtained through the covariance matrix:

$$\pm t_{\gamma}^{*} \sqrt{C_{ii}}$$

(5)

where $t_{\gamma}^{*}$ is given by Students t-distribution, $\gamma = N_d - \eta$ degrees of freedom and $(1 - \alpha)100\%$ is the confidence interval selected, typically 95%.

For non-linear models, there is no exact way to obtain $C$. Therefore, the use of approximations has been suggested. Possibly the most widely used is based on the Cramér-Rao inequality which establishes, under certain assumptions on the number of data and non-linear character of the model, that the covariance matrix may be approximated by the inverse of the Fisher information matrix (FIM) which is formulated as follows:

$$F = E_{y_m,\theta^*} \left\{ \left[ \frac{\partial J_{wlsq}(\theta)}{\partial \theta} \right] \left[ \frac{\partial J_{wlsq}(\theta)}{\partial \theta} \right]^T \right\}$$

(6)

3.3 Optimal experimental design

In order to improve the quality of parameter estimates it is possible to use the model to define new experiments. The idea is to formulate a dynamic optimisation problem where the objective is to find those experimental conditions which result in maximum information content as measured by, for example, the FIM, subject to the system dynamics Eqs. (1-2) plus experimental constraints. The problem can be solved by a combination of the control vector parameterisation (CVP) method and a suitable optimiser enabling the simultaneous design of several dynamic experiments with optimal sampling times [7] and optimal sensor locations [5].

AMIGO (Advanced Model Identification using Global optimisation) [8], a MATLAB based toolbox that covers all model identification steps, was used in this work.

4 DYNAMIC optimisation

The dynamic optimisation (DO) problem can be formulated as: Find the controls $u(t)$ subject to the system dynamics Eqs. (1-2) so as to minimize (or maximize) a given objective functional, that can be related to final product quality, energy consumption, etc. State and control variables may be also subject to constraints which force the satisfaction of safety or environmental regulations, proper operation conditions, etc. These constraints may be point constraints, that must be satisfied at certain time points during process and path constraints, that must be satisfied throughout the process.

The DO problem can be solved using the CVP method. The control variables are discretised and approximated using low order polynomials. The coefficients of these poly-
nominals become the decision variables in a non-linear optimisation problem whose solution involves the simulation of the system dynamics and the assessment of the constraints.

5 ILLUSTRATIVE EXAMPLES

5.1 Reduced order modeling: fruit refrigeration

Refrigeration is probably the most critical process in fruit storage and transport. A poor temperature/pressure control inside the storage chamber would imply a quality loss of the fruit. In this context, model-based process design may help us to compute optimal operation policies in real time. However rigorous models take into account several processes, such as hydrolysis, cellular respiration or mass and heat transport, to name a few. The result is a set of coupled nonlinear PDEs whose solution is computationally demanding. Thus preventing, its application for real time purposes.

Eight state variables were considered: temperature ($T$), concentration of starch ($C_S$), middle lamella ($C_L$), hexose ($C_H$), water ($C_W$), oxygen ($C_{O_2}$), carbon dioxide ($C_{CO_2}$) and nitrogen ($C_{N_2}$). The complete set of equations reads as follows [9]:

$$\frac{\partial C_i}{\partial t} = r_{C_i}; \quad i \in \{S, L, H\}$$ (7)

$$\alpha_{C_i} \frac{\partial C_i}{\partial t} + \nabla(\bar{u}C_i) = \nabla(D_C \nabla C_i) + r_{C_i}; \quad i \in \{O_2, CO_2, N_2\}$$ (8)

$$\frac{\partial C_W}{\partial t} = \nabla(D_W \nabla C_W) + r_W; \quad \rho c_T \frac{\partial T}{\partial t} = \nabla(K_T \nabla T)$$ (9)

where $r_{C_i}$ denotes the reaction terms which are in general highly nonlinear on the corresponding states (for a detailed description see [4]).

For the solution with the FEM, linear Lagrange elements were considered and a spatial mesh of 1818 points (see Figure 1a) was used. Note that since the model consists of eight state variables, around 14500 ODEs need to be solved with this scheme.

The snapshots required for the computation of PODs in Eq. (4) were obtained by running 27 simulations of the PDE model (7)-(9) in different conditions (chamber temperature, humidity and air composition). For validation purposes a new experiment, not included in the 27 simulations, was performed. As shown in the Figures 1b&c, results obtained with ROMs are in good agreement with the ones obtained by FEM, with two orders of magnitude reduction in computational cost.

It should be also mentioned that the joint basis approach, which consists of 18 ODEs, is more efficient than the separated basis approach (31 ODEs). The solution of the ROM requires around 2 min in a PC Intel R CoreTM i7 PC, being therefore suitable for real time applications.
5.2 Dynamic optimisation: freeze-drying of dairy products

Freeze-drying is an attractive dehydration process for preserving nutritional and organoleptic properties of valuable food goods since it helps to maintain the biological activity of their thermosensitive components. Unfortunately, it is also known to be a high demanding process in terms of time and energy, which calls for efficient tools capable of minimizing costs while attaining market quality requirements.

Classical models for freeze-drying processes are usually large-scale and computationally involved, thus unsuitable for real time applications. In order to lighten the computational efforts involved, an accurate yet simplified distributed model has been recently developed [3]. After performing a time-scale analysis of process dynamics, modelling tasks have been focused on the leading scale (the one related to the temperature within the porous matrix). The resulting 1D model for a dairy product sample describing primary and secondary drying reads:

\[
\frac{\partial T_I}{\partial t}(x, t) = \alpha_I \frac{\partial^2 T_I}{\partial x^2}(x, t) ; \quad \frac{\partial T_{II}}{\partial t}(x, t) = \alpha_{II} \frac{\partial^2 T_{II}}{\partial x^2}(x, t) \quad (10) \\
\frac{\partial}{\partial x} \left( -\rho_v(x, t) \frac{K}{\mu} \frac{\partial P_v}{\partial x}(x, t) \right) = 0, \; \forall x \in (0, S(t)) \quad (11)
\]

with the following boundary conditions:

\[
k_I \frac{\partial T_I}{\partial x}(0, t) = \sigma e f \left( T_c^4 - T_I^4(0, t) \right) ; \quad k_{II} \frac{\partial T_{II}}{\partial x}(L, t) = h_L \left( T_L - T_{II}(L, t) \right) \quad (12) \\
T_{I_s} = T_{II_s} = T_s(P_v(S(t))) ; \quad k_{II} \frac{\partial T_{II}}{\partial x}(S(t), t) - k_I \frac{\partial T_I}{\partial x}(S(t), t) = \Delta H_s(\rho_{II} - \rho_I) \frac{\partial S(t)}{\partial t} \quad (13)
\]
\[ P_v(0, t) = P_c \quad ; \quad -\rho_v(S(t), t) \frac{K}{\mu} \frac{\partial P_v}{\partial x}(S(t), t) = \frac{\partial S(t)}{\partial t} (\rho_{II} - \rho_I) \] (14)

where, \( T \) regards temperature, \( P \) the pressure, \( L \) the sample length, \( S(t) \) the moving front position, \( C_b \) the water content and \( T_g \) the glass transition temperature. The subscript \( I \) refers to the dried region, \( II \) to the frozen region, \( c \) to the chamber, \( L \) to the shelf, \( v \) to the vapour and \( s \) to sublimation. The secondary drying is only governed by the Fourier equation in the dried region, while boundary conditions are defined by fluxes in Eq. (12).

For a complete description of the time-scale analysis and the derivation of this simplified model (including the definitions of model parameters and the GAB-based desorption model) readers are referred to [3]. The system of Eqs. (10-14) was solved by FEM with an Arbitrary Lagrangian-Eulerian method so as to track the moving front [11]. An adaptive mesh of 73 nodes was used and the simulation takes around 20 s in COMSOL®.

The model was used to compute the temperature in the shelf (223K ≤ \( T_L \) ≤ 323K) and the pressure in the chamber (10 Pa ≤ \( P_c \) ≤ 60 Pa) that minimise the freeze-drying cycle time \( t_f \), while satisfying process dynamics in Eqs. (10)-(14) simultaneously with the product stability specifications \( \langle C_{ave}^T \rangle(t_f) = 0.02, T(x, t) \leq T_g \).

Different configurations in the equipment were considered, leading to different operational scenarios. The CVP approach was used in combination with a scatter search based global optimiser [10] to solve the associated problems. Optimal profiles (Figure 2a&b)) resulted in a 25% reduction of the cycle time: from \( t_f = 38.4 \) h to \( t_f = 28.6 \) h achieved with constant and dynamic \( P_c \) and \( T_L \) optimal profiles, respectively.

Figure 2: a) & b) Optimal operation profiles, c) & d) Satisfaction of operation constraints
5.3 Identification and dynamic optimisation: frying of potato chips

In deep-fat frying foodstuff is immersed into oil at high (constant) temperature. This induces water evaporation and the formation of a thin crust. As the temperature increases and moisture is lost, the typical deep-frying sensory characteristics (colour, flavour, texture) are developed. However, the use of high temperatures results in the production of acrylamide, a carcinogen compound. Thus model-based optimisation may assist in the design of those operating conditions that provide the best compromise between quality and safety.

A multiphase porous media based model was formulated to describe heat, mass and momentum transfer and acrylamide kinetics within a potato chip [12]. The model consists of a set of coupled nonlinear PDEs describing the evolution of the saturation of water, oil and vapor ($S_w, S_o, S_g$), product temperature ($T$), moisture content ($M$), pressure ($P$), water vapour mass fraction ($\omega_v$) and acrylamide content ($c_{AA}$). The potato chip is assumed to be cylindrical and heated from outside therefore axi-symmetry can be assumed.

The model was solved in COMSOL©. The Convection and Diffusion module was used to solve for water, oil and acrylamide mass conservation while Maxwell-Stefan Diffusion and Convection was used to gas mass fraction and Darcy’s Law and Convection and Conduction were used to solve for pressure and temperature respectively. The selected mesh consists of $20 \times 10$ rectangular elements. The simulation of 1.5 min frying takes around 40 s in a standard PC 3.25GB RAM and 2.83GHz.

Unknown parameters, the heat transfer coefficient ($h$) and the surface oil saturation $S_{o,surf}$, were identified from experimental data using AMIGO [8]. The final model exhibits good predictive capabilities enabling the possibility to analyse alternative operating conditions.

The objective was then to compute the oil temperature profile ($T_{oil,min} \leq T_{oil} \leq T_{oil,max}$) that guaranties the desired moisture content ($M(t_f) \leq 2$) while minimizing final acrylamide content subject to the process dynamics. The problem was solved by means of a combination of the CVP approach and a scatter search based global optimiser [10].

Results show that using two heating zones significantly reduces the final acrylamide content with respect to typical constant operating profiles (Figure 3a). Under constant optimally designed $T_{oil}$, the minimum acrylamide content would correspond to a longer process at a lower temperature. However, a longer process leads to an increase of oil uptake as well as to a reduction of the production rates. The optimal profile corresponds to the use of a higher temperature at the beginning of the process, this helping to satisfy the constraint on the moisture content, followed by a lower temperature to minimise the final acrylamide content (Figure 3b).

5.4 Real time optimisation: thermal sterilisation of packaged foods

The thermal processing of packaged foods is intended to inactivate possible spores, microorganisms or enzymes present in the foodstuff which may have a negative impact
on consumer’s health or product quality. To that purpose, the product is introduced in a steam retort where it is subjected to a given heating-cooling cycle so as to get a pre-specified degree of inactivation indicated by the microbiological lethality. However, some organoleptic properties or nutrients can be negatively affected by the heat action. The objective is, therefore, to optimise operation conditions to maximise quality while guaranteeing safety. In this example, we go a step further, and propose a real time optimisation (RTO) architecture to handle the optimisation during processing and in the presence of uncertainty or sudden disturbances. The performance of the proposed RTO architecture was experimentally validated for tuna paté at the pilot plant in the IIM-CSIC.

The dynamic representation of the plant couples the description of the temperature and pressure inside the retort, the temperature distribution inside the food product and the corresponding distribution of nutrients and microorganisms:

**Retort dynamics**

\[
\frac{dz}{dt} = f(z; \theta) + g(z, u; \theta),
\]  
(15)

here \( f \) and \( g \) are nonlinear vector fields of appropriate dimensions; \( z \) denotes the temperature and pressure in the retort \([T_R, P_R]\); \( u \) stands for the control variables: valve positions for input and output streams. Finally, \( \theta \) denotes the vector of unknown parameters.

**Temperature distribution inside the food product**

\[
\frac{\partial T_{\text{prod}}}{\partial t} = \alpha \Delta T_{\text{prod}} , \quad n(k \Delta T_{\text{prod}}) = h(T_R - T_{\text{prod}})
\]  
(16)

where \( T_{\text{prod}} \) is the temperature of the food stuff and \( h, k, \alpha \) stand for the heat transfer coefficient of the package and the food thermal conductivity and diffusivity, respectively.

**Quality and safety models**

\[
\frac{dC_i(t)}{dt} = -\left( \frac{\ln 10}{D_{i,\text{ref}}} \right) C_i(t) \exp\left( \frac{T_{\text{prod}}(\xi, t) - T_{\xi,\text{ref}}}{z_{i,\text{ref}}} \right)
\]  
(17)
where subindex $C_i$ refers to the concentration of either microorganisms or nutrients.

The unknown parameters of the model, the functional dependencies of fluxes on valves openings and the valves related constants were identified by means of parameter estimation, identifiability analysis and multi-experimental optimal design, using AMIGO toolbox [8].

For the case of the evolution of temperature inside the retort, the resulting model presents excellent predictive capabilities taking into account that a maximum error of around 3% is observed in fast transitions.

The product was packed in glass containers with metal top. The corresponding geometry and the FEM mesh used for simulation purposes are depicted in Figure 4. Selected mesh consists of 184 nodes which translates into 553 ODEs. Three model parameters were estimated from the temperature measurements, namely, the product thermal conductivity, and the glass/steam and the metal/steam heat transfer coefficients. After the model identification, the differences between model predictions and experimental data are lower than 1%.

Once a satisfactory model became available, a POD-based ROM model was developed to be used within the RTO scheme, it should be noted that each simulation of the ROM takes less than 1 s. In addition, the optimal operating conditions were computed off-line using the CVP and scatter search [10] methods.

Real time implementation of the optimal control needs to consider the effect of unmeasured disturbances not being part of the prediction model. To that purpose, feedback was implemented by regularly measuring the current retort variables and observing the relevant variables of the packaged product to compute efficient on-line optimisation. Optimal operation conditions are then re-computed any time a difference between predicted value and off-line optimal solution is detected. A combination of a local optimiser and SSm was designed so as to guarantee feasibility and optimality of the solution even in the presence of significant perturbations or plant/model mismatch (see details in [14]).

Figures 5 illustrate the performance of the RTO architecture in an experimental case were large perturbations occur. The implementation of the optimal off-line heating profile leads to a product that does not fulfil the lethality requirement ($F_c = 8$ min). The RTO architecture proposed in the work was able to drive the system to feasibility and optimality by means of re-computing optimal profiles on-line and slightly extending the duration of the heating phase.
6 CONCLUSIONS

This work presented an overview of recent developments of our group in the context of food process modeling, model identification and reduction and real time dynamic optimisation through a number of cases of interest such as food refrigeration, freeze-drying, deep-fat frying and packaged food thermal processing.

Mechanistic models of the processes were identified for specific food products and simulated via the finite element method; reduced order versions were obtained by means of the proper orthogonal decomposition approach; quality and safety optimal operation conditions were computed using advanced global optimisation techniques and, for some cases, a real time optimisation architecture was designed and tested in pilot-plant.

Results revealed that the use of model-based process design largely improves process performance in terms of final quality and reductions on process time and energy consumption with respect to traditional operations. Proposed methodologies are general therefore opening new venues for the design of emergent or minimal food processing techniques.

Acknowledgements

Funding: EU [CAFE FP7-KBBE-2007-1(212754)], Spanish Ministry of Science and Innovation [SMART-QC AGL2008-05267-C03-01], Xunta de Galicia [IDECOP 08DPI007402PR] and CSIC [PIE 201270E075]. A. Arias-Mendez and E. Lopez-Quiroga acknowledge financial support through JAE-pre and FPI programmes, respectively.

REFERENCES


NETWORK MODELS FOR THE NUMERICAL SOLUTION OF COUPLED ORDINARY NON-LINEAL DIFFERENTIAL EQUATIONS

IVÁN ALHAMA*, JOSÉ L. MORALES† AND FRANCISCO ALHAMAϒ

* Network Simulation Research Group
Universidad Politécnica de Cartagena (UPCT)
Campus Alfonso XIII, ETSIA. 30202 Cartagena, Spain
e-mail: iam3@alu.upct.es

† Department of Structures and Construction, ETSII.
Campus Muralla del Mar
Universidad Politécnica de Cartagena, 30203 Cartagena, Spain
email: josel.morales@upct.es

ϒ Applied Physic Department, ETSIA
Campus Alfonso XIII
Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
e-mail: paco.alhama@upct.es

Key words: Coupled Problems, Mechanics, Network method, Numerical Simulation, Chaos, Electrical Analogy.

Abstract. Many apparently simple problems in mechanics or mechanical engineering, particularly problems related to chaotic systems, are governing by coupled differential equations, generally non-lineal, that have to be solved numerically by specialists in this field. The network model, a tool very used in the last decades for numerical problems in different fields of science and engineering, allows that non-specialists, and even students familiarized with circuits theory, to design networks whose governing equations are just those of the engineering phenomenon, assuming a suitable or formal equivalence between electrical and physical variables. The design of the model, which is composed of a principal network, which implements a balance between the addends of the differential equations, and auxiliary networks to implement the derivative terms, follows a standard procedure. Non-lineal terms of the differential equations are implemented by a controlled source, a kind of device whose operation is quite intuitive. In this communication the models of two characteristic non-lineal mechanical problems are designed step by step with a detailed explanation: the elastic pendulum and the chaotic double pendulum: Solutions are presented graphically by using MATLAB.
1 INTRODUCTION

A large variety of mechanical problems with two or more freedom degrees are formulated by coupled, ordinary differential equations with time as the independent variable. For example: two masses jointed by a spring, a system of masses and pulleys, elastic pendulum, pendulum over a slide bar, double pendulum, etc. Most of these problems are non-linear or even chaotic, since they contain addends with time harmonic (trigonometric) dependencies or with potential functions of the dependent variables. As a consequence, numerical procedures are required for the solution.

In this paper, we solve this kind of problems by using the network method, a kind of analogy between the real process and one formulated by electric circuits whose equations are formally equivalent to those of the problem. The application of this analogy, per se, can be considered an important goal since it relates phenomena mathematically equivalent. Its use to describe physical processes – even though it has been applied in many areas of science and engineering [1,2], especially in heat conduction [3] – remains under-exploited for problems formulated by coupled ordinary differential equations. The network method has demonstrated to be an efficient tool that provides reliable and computationally fast numerical solutions for a large variety of problems formulated by coupled partial differential equations such as transport through membranes [4], heat transfer [5], inverse problems and fluid flow and solute transport [6], among others. One of the main advantages of using this method is that, if the models are correctly designed, their simulation in suitable software provides (almost) the exact solution of the problem due to the powerful mathematics algorithms implemented in the circuit simulation codes.

The proposed analogy is based on the following [7]. On the one hand, the addends of each differential equation are considered as currents (branches in one of the main circuits) that enter (or leave) the only node of this main circuit, according to their sign; the unknown variable of each differential equation is the voltage at that node. There are as many independent main circuits (even they are couple) as equations define the mathematical model. The first derivative term (dy/dt), one of the branches of the main circuit, is simply the current flowing through a capacitor according to the constitutive equation ic=C(dVc/dt). The successive derivatives are obtained by auxiliary circuits formed by new capacitors, whose capacitance is the coefficients of the term, and a special kind of device contained in the libraries of the software, named controlled source. Once obtained, these derivative terms are transported to the main circuit, where the terms of the equation are balanced in the common node, and again implemented by controlled current-sources according to their sign. The rest of the terms of the equations, such as those depending on the unknown variable and/or its powers (integer or fractional), coupled and independent terms, which must be also balanced at the common node of the main circuit, are implemented by controlled current sources. No restrictions are assumed as regards the order and degree of the equation as well or the kind of non-linearity involved. The model is completed by fixing the initial voltages at the capacitors which are defined by the initial conditions. Once the model is designed no mathematical manipulation is needed; the code Pspice [8] does this work with its powerful computational algorithms.
Two advantages mentioned as regards the network method: (i) no mathematical manipulation (inherent to most of the numerical and analytical methods) is required; the computer code used in this work, Pspice, does the calculations with sophisticated mathematical algorithms (Nagel [9]), and (ii) a few programming rules are necessary to elaborate the network text file, since the number of different electrical devices or components involved is very small.

The sections of the documents are organized as following: a detailed explanation of the network design method followed by its application to two illustrative, selected problems. The section conclusions resume the advantages of the method for this kind of applications.

2 MATHEMATICAL AND NETWORK MODELS

The mathematical model is a system of coupled, ordinary differential equations, with as many equations as dependent variables exist. The equations can be of any order and any grade, with power of real numbers, and may contain coupled terms as well as arbitrary functions of the dependent variables and independent terms.

2.1 Auxiliary circuits

Firstly, we describe the design of the auxiliary networks that implement the derivative terms. Pspice, or any bother code for circuit simulation, contains a group of ideal controlled sources capable of assuming any kind of non-linearity; these, suitably connected with capacitors, provide the auxiliary circuits that implement any derivative term. Four different sources can be used, Figure 1: E is a voltage source whose output is defined (by programming) as an arbitrary function of the voltage at any node (or voltages of any nodes) of the network, while H is a voltage source whose output is proportional to the current of a time-independent voltage source. The other two current-sources, G and F, have similar meanings.

Now, if we call $V_j$ the voltage at node $j$, the auxiliary network of Figure 2 (a) formed by a capacitor (of capacitance $C_a = a_1$) and a voltage-controlled voltage-source (whose output voltage is $v_{E1} = V_j$, the same as the input voltage) is able to provide the value $a_1(dV_j/dt)$ since the current through $C_a$ is defined as $i_{Ca} = C_a(dV_j/dt) = a_1(dV_j/dt)$. A new auxiliary loop, formed by the current-controlled voltage-source $H_1$ and the resistor $R_1$ (of resistance $a_1$), provides the first derivative of $V_j$.

![Figure 1: Controlled sources: E: voltage-controlled voltage-source, G: voltage-controlled current-source H: current-controlled voltage-source, and F: current-controlled current-source](image-url)
The output of $H_1$ is a voltage whose value is the input current $i_{zero,1}$, i.e. the current of the ammeter $V_{zero,1}$ which, in turn, is the current of the capacitor $C_a$; consequently, the voltage through $R_1$ is $(1/a_1)a_1(dV_j/dt)=dV_j/dt$, the first derivative function of $V_j$. Resistor $R_{zero,1}$ is included to satisfy the continuity criteria required by Pspice. Also, the use of $V_{zero}$ as ammeter is prescribed by the requirements of Pspice: the input current of the controlled sources of type $H$ must be specified as a current coming from a constant voltage source.

In the same way, the second derivative of $V_j$, $a_2(d^2V_j/dt^2)$, is provided by the auxiliary network of Figure 2 (b). The output of the current-controlled voltage-source $H_b$, $V_{Hb} = i(V_{zero,1}) = a_1(dV_j/dt)$, defines the current through the capacitor $C_b$ (of capacitance $a_2/a_1$) as $i_{Cb} = C_b(a_1(d^2V_j/dt^2)) = a_2(d^2V_j/dt^2)$. In addition, $H_2$ and $R_2$ (of resistance $a_2/a_1$) provide the second derivative of $V_j$ (the voltage through the resistor $R_2$). The following derivative terms are implemented in the same way; Figure 2 (c) shows the network of the third derivative term. Quantities between brackets always denote the control variables that determine the input of the sources.

2.2 Main circuit loops

There are as many main loops as equations or dependent variables in the mathematical model. In turn, each main network is formed by as many branches in parallel as terms of
the differential equations. One of the nodes of the main loop is the solution of the related
dependent variable while the other is the common reference voltage (earth). Each branch
drives a current (whose value is that of the term) that comes in or out of the common
node, according to the sign of the term in the equation. The term related to the first
derivative term, if it exists, is implemented by a capacitor, while the rest of the derivative
terms are implemented by voltage-controlled current-sources; these read their value from
their respective auxiliary circuits and introduces them as the output of the controlled
source by software. When the derivative term of any order has a degree different from
unity (or it is a real number), it is also possible to introduce it by software, as the output of
the related controlled source.

The rest of the terms of the differential equation (coupled, independent or other kind of
non-lineal terms, such as terms with arbitrary dependencies on the dependent variable) can
be implemented in the model by controlled or independent sources. Finally, the
independent term is simply implemented by a constant source. A resistor of very high
value that does not influence the solution is also located in parallel in the main circuit to
satisfy the continuity requirements imposed by the code Pspice.

Whatever be the initial conditions, they are implemented in the model by giving initial
voltages at the capacitors. The solutions \( y_i(t), 1 \leq i \leq n \), are read at the only node of the main
circuit (as a consequence of the balance between the currents of the branches, Kirchhoff’s
law), while the solutions of successive derivatives can be read at the nodes of the auxiliary
circuits.

The network file can be designed by a text editor (text file) or directly by a graphic
ambient by means of the option ‘schematics’ contained in the code Spice.

3 APPLICATIONS

3.1 The elastic pendulum

This is the system whose physical scheme is shown in Figure 3. A small body of mass
\( m_0 \), fixed to a spring of negligible mass, is hanging to the ceiling. The system has degree
of freedom, the angle that forms the spring line with the vertical or equilibrium line, \( \theta \),
and the distance from the mass to the fixing ceiling point, \( r \).

The governing equations are

\[
\frac{d^2r}{dt^2} - r \left( \frac{d\theta}{dt} \right)^2 - g(\cos \theta) + \frac{k}{m_0}(r-r_o) = 0
\]

(1)

\[
r \left( \frac{d^2\theta}{dt^2} \right) + 2 \left( \frac{dr}{dt} \right) \left( \frac{d\theta}{dt} \right) + g(\sin \theta) = 0
\]

(2)

where \( t \) is time, \( k \) the constant of the spring, \( g \) the gravitational acceleration, and \( r_o \) the initial
location of the mass with respect to the fixing ceiling point. To simplify, we will assume that
\( r_o = l_0 + d_o \), being \( l_0 \) the length of the rest spring and \( d_o \) the solution of the equilibrium
equation \( d_o = (m_0g)/k \).
**Figure 3:** Physical scheme of the elastic pendulum

**Figure 4:** Network model of the elastic pendulum. (a): Equation (1), (b): equation (2)

Figure 4 shows the network model formed by two main loops (one per equation) plus the auxiliary circuits. Controlled sources G1a to G4a implement the four terms of equation
(1), the $r$ equation, while $G_{1b}$ to $G_{3b}$ balance the three terms of equation (2), the $\theta$ equation. The control voltage of each source is shown between brackets. Auxiliary circuits $E_{1a}$, $H_{1a}$ and $H_{2a}$, together with the associated capacitors and ammeters implement the necessary electrical components required to form the successive derivative terms of the equation (1), while $E_{1b}$, $H_{1b}$ and $H_{2b}$ plays this role for the equation (2).

The solution for $r(t)$, $\frac{dr}{dt}$ and $\frac{d^2r}{dt^2}$ can be read at the nodes 10 (the node of the main loop of the $r$ circuit), 30 and 40, respectively, while the solution for $\theta(t)$, $\frac{d\theta}{dt}$ and $\frac{d^2\theta}{dt^2}$ can be read at the nodes 50 (the node of the main loop of the $\theta$ circuit), 70 and 80, respectively. Couple terms are directly written by software when specifying the control voltage of the associated source. Initial conditions for $r$, $\theta$, and their derivatives $\frac{dr}{dt}$ and $\frac{d\theta}{dt}$ are implementing by fixing the initial voltage of the capacitors $C_{1a}$, $C_{1b}$ and $C_{2a}$, $C_{2b}$, respectively.

For the following values of the parameters:
$m_0 = 1$, $k = 10$, $r_0 = 2.2$, $g=2$, $\theta_{\text{initial}} = 0.01$, $(\frac{dr}{dt})_{\text{initial}} = (\frac{d\theta}{dt})_{\text{initial}} = 0$
the solution, represented in the output graphic ambient of Pspice, is shown in Figure 5. Computational time is of the order of 5 s in a portable PC.

![Figure 5: Solution for $r(t)$, $\frac{dr}{dt}$ and $\frac{d^2r}{dt^2}$ (up) and for $\theta(t)$, $\frac{d\theta}{dt}$ and $\frac{d^2\theta}{dt^2}$ (down)]
3.2 The chaotic double pendulum

Two small masses are hanging as shown in Figure 6. When one or both masses are displaced from their equilibrium location, the system oscillates in a chaotic form.

\[ 2 \left( \frac{d^2 \theta_1}{dt^2} \right) + \left( \frac{d^2 \theta_2}{dt^2} \right) + 2 \left( \frac{g}{l_1} \right) \theta_1 = 0 \] (3)

\[ \left( \frac{d^2 \theta_2}{dt^2} \right) + \left( \frac{d^2 \theta_1}{dt^2} \right) + \left( \frac{g}{l_2} \right) \theta_2 = 0 \] (4)

Figure 7 shows the network model which can be understood following the explanation of the elastic pendulum. Auxiliary circuits are the same as the former application. Controlled sources G1a to G3a implement the terms of equation (3), the \( \theta_1 \) equation, while G1b to G3b the terms of equation (4), the \( \theta_2 \) equation. The solution for \( \theta_1(t) \), \( d\theta_1/dt \) and \( d^2\theta_1/dt^2 \) can be read at the nodes 10, 30 and 40, while that of \( \theta_2(t) \), \( d\theta_2/dt \) and \( d^2\theta_2/dt^2 \) at the nodes 50, 70 and 80, respectively.

For the following values of the parameters:

\[ m_1 = 1, \ m_2 = 1, \ l_1 = 1, \ l_2 = 1, \ g = 1, \ \theta_{1,\text{initial}} = 0.001, \ \theta_{2,\text{initial}} = 0.005, \ (d\theta_1/dt)_{\text{initial}} = (d\theta_2/dt)_{\text{initial}} = 0 \]

the solution is shown in Figures 8 and 9. For a better appreciation of the chaotic movement the phase diagrams \( \theta_1 = \theta_1(d\theta_1/dt) \) and \( \theta_2 = \theta_2(d\theta_2/dt) \) are shown in Figures 10 and 11. Total computing time is of the order of 6 s.
Figure 7: Network model of the double pendulum. (a): equation (3), (b): equation (4)

Figure 8: Solution for $\theta_1(t)$, $d\theta_1/dt$ and $d^2\theta_1/dt^2$
Figure 9: Solution for $\theta_2(t)$, $d\theta_2/dt$ and $d^2\theta_2/dt^2$

Figure 10: Phase diagrams. $\theta_1 = \theta_1(d\theta_1/dt)$

Figure 11: Phase diagrams. $\theta_2 = \theta_2(d\theta_2/dt)$
4 CONCLUSIONS

Network method has demonstrated to be an efficient tool for the numerical simulation of mechanical problems whose mathematical model is formed by coupled, non-linear ordinary differential equations of any order and any grade. The design of the model is relatively simple since very few electrical components are required which, in turn, makes that very few programming rules are needed. Thanks to the so named ‘controlled sources’, a special device contained in the libraries of the circuit simulation computing codes, whose output are specified by software, any kind of nonlinearity as well as coupled terms can be easily implemented in the model. A same protocol define the design of the network; this contains as many main loops as governing equations, where the solution of the dependent variables emerges at the common nodes, and as many auxiliary loops as derivative terms contain the equations, where the derivative terms can be determined. The power mathematical algorithms used in modern codes make the computational time quite negligible and the numerical solution reliable.

REFERENCES

1 INTRODUCTION

The proliferation of mobile devices over the last years provides opportunities and challenges for solving problems in science and engineering. Among other novel features, mobile devices contain global positioning sensors, wireless connectivity, built-in web browsers and photo/video/voice capabilities that allow providing highly localized, context aware applications. Mobile phones have become as powerful as any desktop computer in terms of applications they can run. However, the software development in mobile computing is still not as mature as it is for desktop computer and the whole potential of mobile devices is wasted [7, 8].

Although mobile technologies present new opportunities for services and business, they also present development and implementation challenges. Various authors describe challenges of mobile software development, for example, in [7] authors highlight creating user interfaces for different kinds of mobile devices, providing reusable applications across multiple mobile platforms, designing context aware applications and handling their complexity and, specifying requirements uncertainly. To ensure that the application provides sufficient performance while maximizing battery life is remarked in [18]. Some mobile applications also must determine the user location before offering the service and then track the location to adapt services and information accordingly. Besides, an additional challenge is to achieve the required level of security, reliability and quality of mobile services. Accepted rules for the design of traditional interfaces can not be fully implemented in the design of mobile interfaces [7].

A current problem in the engineering community is the modernization of legacy software. Software modernization, understood as technological and functional evolution of legacy systems, provides principles, methods, techniques and tools to support the transformation from an existing software system to a new one that satisfies different requirements. To meet new demands, existing systems must be constantly evolved. Many of the existing systems may be written for technology which is expensive to maintain and which may not be aligned with current organizational politics. However, they resume key knowledge acquired over the life of an organization and there is a high risk to replace them because they are generally business-critical systems. A number of solutions have been proposed to deal with this problem such as redevelopment, which rewrites existing applications, or migration, which moves the
existing system to a more flexible environment while retaining the original system data and functionality. A good solution should be to restore the value of the existing software, extracting knowledge and exploiting investment in order to migrate to new software that incorporates the new technologies.

On the one hand, traditional reverse engineering techniques can help in the software migration to mobile applications. They are related to the process of analyzing available software with the objective of extracting information and providing high-level views on the underlying code [5,19].

On the other hand, the rapid proliferation of different mobile platforms has forced developers to make applications tailored for each type of device. Within the mobile development, many companies have different development teams redoubling the software engineering efforts for functionally similar mobile applications. To achieve interoperability with multiple platforms the migration needs of novel technical frameworks for information integration and tool interoperability such as the initiative of the Object Management Group (OMG) called Model Driven Architecture (MDA) [11]. The outstanding ideas behind MDA are separating the specification of the system functionality from its implementation on specific platforms, managing the software evolution from abstract models to implementations, increasing the degree of automation of model transformations, and achieving interoperability with multiple platforms. Models play a major role in MDA which distinguishes at least Platform Independent Model (PIM) and Platform Specific Model (PSM). An MDA forward engineering process focuses on the creation of PIMs which are automatically transformed by tools to PSMs which are next transformed to specific code.

The essence of MDA is the Meta Object Facility Metamodel (MOF) that allows different kinds of software artifacts to be used together in a single project [14]. MOF provides two metamodels: EMOF (Essential MOF) and CMOF (Complete MOF). EMOF favors simplicity of implementation over expressiveness. CMOF is a metamodel used to specify more sophisticated metamodels. Transformations are expressed in the MOF 2.0 Query, View, Transformation (QVT) metamodel [16].

OMG is involved in the definition of standards to successfully modernize existing information systems. The OMG Architecture-Driven Modernization Task Force (ADMTF) is developing a set of specifications and promoting industry consensus on modernization of existing applications [1].

The objective of this paper is to describe a reengineering process that allow moving existing desktop applications for solving engineering problems of multidisciplinary character to mobile platforms. Our research aims to simplify the creation of applications for mobile platforms by integrating traditional reverse engineering techniques, such static and dynamic analysis, with MDA. It is worth considering that mobile applications are not different applications but are mainly intend to complement the existing desktop systems in the organization to make them mobile. We validated our approach by using the open source application platform Eclipse, EMF (Eclipse Modeling Framework), EMP (Eclipse Modeling Project) and the Android platform [2, 9].

This paper is organized as follows. Section 2 describes a reengineering process for adapting existing object-oriented software applications to mobile platforms. In Section 3 we summarizes reverse engineering techniques such as static and dynamic analysis. Section 4 describes metamodeling techniques in the context of the reengineering process. Particularly,
this section specifies how to transform models by using transformation languages aligned with the MDA standards. Also, it describes how to obtain a target application in the mobile platform. Finally, Section 5 presents conclusions and challenges in the modernization of legacy systems to mobile technologies.

2 A REENGINEERING PROCESS: FROM DESKTOP TO MOBILE APPLICATIONS

We propose a reengineering process for modernizing desktop applications to mobile platforms (Fig. 1). Reengineering process can be summarized into three steps: reverse engineering, model transformation and implementation. Reverse engineering extracts out higher level views of the system expressed by different kind of artifacts that allow creating a model of the source application called PIM in the MDA context. The transformation for one PIM to several PSMs is at the core of MDA. The objective of the model transformation step is to transform the source model (a PIM), into target models (PSMs linked to different mobile platforms). Finally, during the implementation step, target applications for different mobile platforms are generated from the PSMs.

The proposed process starts from a source application and the application of reverse engineering techniques to support the understanding of it. We consider that only the source code is the repository of information for recovering the system design. Because of this, the first stage of this process consists of applying different techniques of reverse engineering that are based on two main types of analysis: structural or static analysis, and behavioral or dynamic analysis.

Static analysis extracts static information that describes the software structure reflected in the documentation (e.g., the source code text) and is supported by CASE tools. Dynamic analysis information describes the structure of the run-behavior and can be extracted by using debuggers, event recorders and general tracer tools. Then, the first stage of the reengineering process allows extracting artifacts in a high abstraction level that describe the application being analyzed.

At this point, it is necessary to consider the dependencies that have the recovered software artifacts with the technologies applied to implement the system under analysis. These dependencies should not impact to the artifacts that describe the new system to be implemented. To avoid these situations, the integration of reverse engineering techniques with MDA is proposed. MDA aims interoperability between platforms and technology independence proposing that all devices involved in a development process are represented from MOF. MOF allows different kinds of software artifacts to be used together in a single project. The transformation between models allows representing the new system to be implemented.

There are different ways to achieve model transformations, for example by using a programming language or metamodeling techniques. There exists specific transformation languages that provides a way to specify how generate a target model that conforms to a target metamodel from a source model that conforms to a source metamodel, for example, we can mention QVT or ATL transformation languages. As a result of this step, models of target applications related to different mobile platforms are created.
Next, in the implementation step, target applications are generated. To ensure the success of the steps of the reengineering process different tools known as CASE (Computer Aided Software Engineering) are needed. Each of these tools presents different features and support for the techniques involved in the reengineering process. In Section 2.1 we describe characteristics of different CASE tools.

2.1 CASE Tools

The success of MDA depends on the existence of CASE tools that make a significant impact on software processes such as forward engineering and reverse engineering processes. All of the MDA tools are partially compliant to MDA features. CASE MDA are generally extensions of CASE UML and most of them are not aligned with MOF. They provide good support for modeling and limited support for automated transformation in reverse engineering [6].

Many CASE tools support reverse engineering, however, they only use more basic notational features with a direct code representation and produce very large diagrams. Reverse engineering processes are facilitated by inserting annotations in the generated code. These annotations are the link between the model elements and the language.
The Eclipse Modeling Framework (EMF) [9] was created for facilitating system modeling and the automatic generation of Java code. EMF started as an implementation of MOF resulting Ecore, the EMF metamodel comparable to EMOF. EMF has evolved starting from the experience of the Eclipse community to implement a variety of tools and to date is highly related to Model Driven engineering (MDE). For instance, commercial tools such as IBM Rational Software Architect, Spark System Enterprise Architect or Together are integrated with Eclipse-EMF. Blu Age Reverse Modeling recovers from legacy systems some information necessary to build UML 2 models [6].

Few MDA-based CASE tools support any of the QVT languages. As an example, IBM Rational Software Architect support model-to-model and model-to-text transformations but not MOF and QVT. Spark System Enterprise Architect is based on MDA and UML 2.1 and then is compatible with MOF.

Other tools partially support QVT, for instance Together allows defining and modifying transformations model-to-model (M2M) and model-to-text (M2T) that are QVT-Operational compliant [17]. Medini QVT supports partially MOF and implements QVT. It is integrated with Eclipse and allows the execution of transformations expressed in the QVT-Relation language [13]. The MMT (Model-to-Model Transformation) Eclipse project, is a sub-project of the top-level Eclipse Modeling Project that provides a framework for model-to-model transformation languages. Transformations are executed by transformation engines that are plugged into the Eclipse Modeling infrastructure. The main transformation engines developed in the scope of that project are ATL and QVT [3,17]. ATL is a model transformation language and toolkit developed by ATLAS INRIA & LINA research group. In the MDE field, ATL provides ways to produce a set of target models from a set of source models. To date, the QVT declarative component is in its “incubation” phase and provides only editing capabilities to support the QVT language.

Currently, there are no tools supporting a complete reengineering process as proposed by this work. However, various tools are available to deal with it. In this paper we validate our approach by using the open source application platform Eclipse, EMF and EMP. In this section, we select a set of appropriate tools that will be described in the following sections. Besides, we select as a running example an adaptation of a CRM (Customer Relationships Management) desktop application to mobile platforms, Android platform in particular. Next, in Section 2.2, we introduce the running example.

### 2.2 A running example

In the following sections we describe in detail the proposed reengineering process. The different steps are illustrated by using a common example, a CRM (Customer Relationship Management) application. A CRM manages company interactions with current and future customers. Interactions are supported and guided by creating dynamic customer profiles that register information such as contracted services and products, frequent contact channels, and commercial transactions and their associated responses.

The CRM application that will be used to exemplify each step of the proposed process reengineering is called SellWin [17]. The analysis in this examples, will prioritize entities related to managing customer data. The simple client-server architecture of the application follows a component-oriented design separated in different modules: Domain, Data Base,
Server and User Interface. SellWin lacks adequate documentation to understand its design, which allows us to analyze the strengths and weaknesses of the application of reverse engineering techniques for understanding its functionality.

3 REVERSE ENGINEERING: FROM OBJECT-ORIENTED CODE TO MODEL

Reverse Engineering is the process of analyzing available software artifacts such as requirements, design, architectures, code or byte code, with the objective of extracting information and providing high-level views on the underlying system. Reverse engineering does not involve changing the source legacy systems, but understanding them to help reengineering processes that are concerned with their re-implementing. The main traditional techniques related to reverse engineering are static and dynamic analysis.

Static analysis extracts static information that describes the software structure reflected in the software documentation (e.g., the source code text) whereas dynamic analysis information describes the structure of the run-behavior and can be extracted by using debuggers, event recorders and general tracer tools. Static analysis is based on classical compiler techniques and abstract interpretation.

In [10], author provides a comparison of static and dynamic analysis from the point of view of their synergy and duality. He argues that static analysis is conservative and sound. Conservatism means reporting weak properties that are guaranteed to be true, preserving soundness, but not be strong enough to be useful. Soundness guarantees that static analysis provides an accurate description of the behavior, no matter on what input or in what execution environment. Dynamic analysis is precise due to it examines the actual run-time behavior of the program, however the results of executions may not generalize to other executions. Also, author argues that whereas the chief challenge of static analysis is choosing a good abstract interpretation, the chief challenge of performing good dynamic analysis is selecting a representative set of test cases. A test can help to detect properties of the program, but it can be difficult detect whether results of a test are true program properties or properties of a particular execution context. The combination of static and dynamic analysis can enrich reverse engineering process. There are different ways of combination, for instance performing first static analysis and then dynamic one or perhaps, iterating static and dynamic analysis. Likewise, the definition of appropriate heuristics may guide the search for information on the traces generated during the dynamic analysis.

3.1 Example

The first stage focuses on retrieving software artifacts that are useful to understand the design and implementation decisions for the chosen application. The aim is to detect the classes that make up the application and objects involved in the different functionality. With this information, models expressed by UML diagrams are generated [21].

Static analysis allows detecting the classes that compose the application and their relationships. Dynamic analysis is used to detect how they interact to solve the offered functionality. In this case, dynamic information is recovered using two techniques: execution trace and memory snapshot.
3.1.1 Static analysis: Class Diagrams

The initial step had to do with the recovering of class diagrams to detect relationships between the various components that make up the main modules. The explorer tool integrated with the Eclipse development environment, called UML ObjectAid [6], was used in this step. ObjectAid is a free tool for working with class diagrams but, it restricts access to sequence diagrams using a special license.

As an example, we show the class diagram of the Customer Management (Fig. 2). The purpose of this diagram is to visualize the relationships between the various modules. As we can see, the user interface module is unrelated to the database, and the access to data is provided by the server module, with which it maintains a direct association via a defined interface. Moreover, the user interface is the only one that has direct associations with the domain, since both the server and database, have only registered dependencies according to the methods of the interface of each class.

3.1.2. Dynamic analysis: execution traces and memory snapshots

To obtain and analyze execution traces of an application, we select the Eclipse Test and Performance Tools Platform (TPTP). It provides an open platform supplying powerful frameworks and services that allow software developers to build unique test and performance tools. TPTP allows executing instances of the application and registering the invocations. While the result is not a classic sequence diagram (for example, control statements are not detected) it is a good approximation to detect methods involved in each specific functionality and method invocation sequences. Also, the resulting diagram lets see how user interface components interact directly with domain components but not with the database components. Other dynamic analysis technique that was used in the process is memory snapshots. This analysis seeks to recover what is the current value of each of the attributes of the objects created during the execution of the application. This information is important not only to successfully deploy the application on the target platform, but in the modeling stage, as described in the next section. To detect the state of the memory was used a commercial tool that can be freely used for a limited evaluation time called YourKit Java Profiler. This tool allows running the application and capturing the information of the objects that were created in memory.

As a result of the application of static and dynamic analysis techniques it is possible to recover artifacts that allow reconstructing the design of the application under consideration. From this design, the source application can be implemented on the target platform, making the necessary modifications according to the mobile restrictions (memory space, screen size, usage limitations, among others).
4 MODEL TRANSFORMATION

MDA aims at the development of software systems based on the separation of business and application logic from underlying platform technologies, facilitating technology independence and interoperability between platforms. All artifacts involved in a development process are represented by means of metamodeling techniques, MOF metamodeling in particular. MOF metamodels are used to describe the transformations at model level. For each transformation, source and target metamodels are specified. A source metamodel defines the family of source models to which transformation can be applied. A target metamodel characterizes the generated models. In the MDA context, we consider that source models are PIMs and target models are PSMs. Model transformation provides a way to specify how generate a target model that conforms to a target metamodel from a source model that conforms to a source metamodel.

We validate our approach in the Eclipse Modeling Framework. Source and target metamodels conform to Ecore metamodel, which is comparable to EMOF. In this experience, we select ATL as model transformation language. This stage of the translation process, was supported by the Eclipse Modeling Project (EMP) which provides tools for both defining metamodels and transformation rules, and executing the translation process.
4.1 Example

The Android platform provides a version of the Java language that is different to the version provided by environments of standard execution (Java Runtime Environment). One of main differences of this version of Java is the way of constructing graphic interfaces. It does not provide frameworks such as Swing or AWT but its own component libraries called widgets. Considering the above-mentioned, we present examples of translation centered on the components of the user interface module, which require substantial changes.

Fig. 3 shows a simplified Java/JSwing metamodel that includes classes (and attributes) used for the construction of client management screen. On the other hand, Fig. 3 also shows a simplified Java/Android target metamodel to implement screens of client management.

The main difference between the source and target metamodels is that interface controls do not provide the same functionality for all cases. In some cases, due to technological constraints and characteristics of the target platform, it is necessary create equivalent functionality using different widgets. One such case may be the JTable class, which

Figure 3. Java/JSwing source metamodels and Java/Android target metamodel
module SwingToAndroid;
create OUT : JavaAndroid from IN : JavaSwing;

helper context JavaSwing!Component def:
getVisibility(): JavaAndroid!Visibility =
if self.visible = true then
  #VISIBLE
else
  #INVISIBLE
endif;

helper context JavaSwing!Component def:
getWidth(s: JavaSwing!Dimension): Integer =
if s.oclIsUndefined() then
  0
else
  s.width
endif;

helper context JavaSwing!Component def:
getHeight(s: JavaSwing!Dimension): Integer =
if s.oclIsUndefined() then
  0
else
  s.height
endif;

rule ComponentToView {
  from
  jc: JavaSwing!Component
to
  tv: JavaAndroid!View ( visibility <- jc.getVisibility(), id <- jc.name, enabled <- jc.enabled, width <- jc.width, height <- jc.height, mLeft <- jc.x, mTop <- jc.y, mMinHeight <- jc.getHeight(jc.minimumSize), mMinWidth <- jc.getWidth(jc.minimumSize))
}

rule ContainerToViewGroup extends ComponentToView {
  from
  jc: JavaSwing!Container
to
  tv: JavaAndroid!ViewGroup ( mChildren <- jc.component, mChildrenCount <- jc.ncomponents )
}

rule JComponentToViewGroup extends ContainerToViewGroup {
  from
  jc: JavaSwing!JComponent
to
  tv: JavaAndroid!ViewGroup
}

rule JLabelToTextView extends JComponentToViewGroup {
  from
  jc: JavaSwing!JLabel
to
  tv: JavaAndroid!TextView( mText <- jc.text )
}

rule JTextFieldWithColumnsToEditText extends JComponentToViewGroup {
  from
  jc: JavaSwing!JTextField(jc.columns > 0)
to
  tv: JavaAndroid!EditText( enabled <- jc.editable, filters: JavaAndroid!LengthFilter ( mMax <- jc.columns ) )
}

rule JTextFieldToEditText extends JComponentToViewGroup {
  from
  jc: JavaSwing!JTextField(jc.columns = 0)
to
  tv: JavaAndroid!EditText ( enabled <- jc.editable )
}

Figure 4. From Swing to Android: ATL rules

implements a data table, which has no equivalent functionality in Android and will be implemented by combining other controls. In other cases, we can also see restrictions that are configured from attributes of a control, becoming associations between widgets. For example, to set a maximum size for the number of characters that can be entered in an edit control (for class JTextField, attribute column), it is represented in Android by means of the association between the class EditText with a filter of input of length (class LengthFilter and
Figure 5. The original and the resulting screen of Client Management

The configuration of its attribute $nMax$). These considerations will be present at the moment of establishing translation rules in ATL. In Fig. 4, we present some of the ATL rules that allow the translation between the two metamodels. For example, the first rule describes how to transform the parent metaclass of the source class Component into the parent metaclass of the target class View. The transformation is performed for each attribute in an almost direct way, except for attributes that need to be invoked from the previously defined helpers. The main difficulties in the new implementation are associated with the particular features of platforms, primarily the size of the screen available to build the user interface, and methods of use of the input devices available which differ significantly from those found in a classical computer. Fig. 5 shows the original screen of client management and the resulting screen on the mobile device.

6 CONCLUSIONS

In this paper we have described a model driven reengineering process to adapt software systems for mobile platforms. Our proposal intends to improve the productivity of development teams taking into account the following strategies: working at a higher abstraction level focusing on design rather than on implementations, encapsulating mobile
platform model knowledge in a metamodel specification that will be reused in different applications, linking different CASE tools to the different activities, and particularly linking models to full code generators.

To propose a development process that considers platform-independent models is a very important practice to prevent future duplication of effort when trying to deploy the application to a new target platform. However, we detect some inconveniences. When the only information is the code, the success of the reverse engineering process depends largely on the availability of assistance and automation tools. This is one of the most important complications when attempting to migrate legacy system logic into a new application.

Beyond these difficulties, we show the acceptable feasibility of the proposed reengineering process by integrating different CASE tools and highlight the importance of having tools that assist during each stage of the proposed process. Our approach focuses on important problems in mobile development: creating user interfaces and enabling software reuse across multiple mobile platforms.

REFERENCES

SEMI-ANALYTICAL APPROACH FOR RATCHETING STRAIN AND WEAR PREDICTION IN CASE OF LINE ROLLING CONTACT

M. SOFER*, R. HALAMA†

* Centre of Excellence IT4Innovations; Department of mechanics of materials; VŠB – Technical University of Ostrava
17. Listopadu 15/2172 Ostrava-Poruba, Czech Republic
email: michal.sofer@vsb.cz, www.vsb.cz

† Centre of Excellence IT4Innovations; Department of mechanics of materials; VŠB – Technical University of Ostrava
17. Listopadu 15/2172 Ostrava-Poruba, Czech Republic
email: radim.halama@vsb.cz, www.vsb.cz

Key words: Rolling contact fatigue, Wear, Coupled problems, Mazzu, Inverse algorithm

Abstract. The most studies in the field of rolling contact fatigue and wear consider that these phenomena are independent, but this paper shows strong coupling of both in the case, when ratcheting mechanism is the driving one. There is presented the application of semi-analytical approach, introduced by A. Mazzu [1], for two dimensional rolling contact case. The model is based on non-linear kinematic and isotropic hardening rule of Chaboche and Lemaître [2]. Mazzu’s approach has been modified to correctly predict ratcheting and wear rate for different values of contact pressure. A correlation between the parameter \( \sigma_L \) (yield stress) and the value of the maximum contact pressure \( p_0 \) has been found. All performed simulations are based on rolling contact fatigue tests realized on author’s department [3].

1 INTRODUCTION

Rolling contact is in practice relatively common phenomenon, no wonder, that this phenomenon is investigated worldwide. Some authors have in recent years published papers, which are focused on evolution of surface and subsurface ratcheting or wear due to rolling contact. This problem can be to a certain degree solved by numerical methods, but the main disadvantage is their computational demand, related to impossibility of simulating thousands or millions of cycles. On the other side, analytical or semi-analytical methods are giving us a possibility of simulating a high number of cycles.

Wear and rolling contact fatigue are one of the most common damage phenomena, which are occurring on components subjected to cyclic and repeated loading. On the basis of realized twin-disk experiments on author’s department [3], there was a possibility to explore the correlation between the parameter \( \sigma_L \) (yield stress) and maximum contact pressure for the case of line contact, concretely for maximum pressure values of 800, 1000 and 1200 MPa. For these purposes, there were utilized semi-analytical model of A. Mazzu [1] and inverse algorithm [4, 5].
2 WEAR ASSESSMENT WITH AIM OF MAZZU APPROACH

Mazzu’s approach [1] is based on non-linear kinematic hardening rule, introduced by Lemaitre and Chaboche [2] and is subsequently used for ratcheting prediction under contact surface in case of rolling/sliding two-dimensional contact. The basis of the model is the assumption that $\tau_{xz}$ is the only stress component, which is responsible for ratcheting and plastic flow in contact.

![Coordinate system](image)

**Figure 1:** Coordinate system

This hypothesis is, according to [1], reasonable in the material layer of subsurface flow, where the kinematic hardening process is mainly governed by shear stress $\tau_{xz}$, referring to the coordinate system presented in Fig.1. The reason is in the variation of the shear stress component $\tau_{xz}$, which varies within larger range compared to other deviatoric stress components whatever the friction coefficient is. In the Mazzu’s model [1], elastic domain can be expressed in the following way:

$$ F = \sqrt{3} \cdot \tau_{xz} - X_{xz} - (R + \sigma_L) \leq 0 \tag{1} $$

where:

- $\sigma_L$ – initial tensile yield stress [MPa],
- $X_{xz}$ – back-stress expressing kinematic hardening [MPa],
- $R$ – isotropic hardening variable [MPa]

Occurrence of plastic strain will induce variation in the isotropic and kinematic hardening term respectively. The back-stress variation principle is given by following equation:

$$ dX_{xz} = C \cdot \frac{d\gamma_{xz}}{\sqrt{3}} - \gamma \cdot X_{xz} \bigg| \frac{d\gamma_{xz}}{\sqrt{3}} \bigg| \tag{2} $$

where:

- $C$ – material parameter [MPa],
- $\gamma$ – material parameter [-]

The isotropic term variation is expressed by:

$$ R = R_{a1} \left[ 1 - \exp \left( -b_1 \cdot \frac{\gamma_{xz}}{\sqrt{3}} \right) \right] + R_{a2} \left[ 1 - \exp \left( -b_2 \cdot \frac{\gamma_{xz}}{\sqrt{3}} \right) \right] \tag{3} $$
where:

\[ R_{\infty 1}, \quad R_{\infty 2} \] – material parameters [MPa], which states the limit value of \( R \):

\[ R_{\infty} = R_{\infty 1} + R_{\infty 2} \]

\( b_1, b_2 \) – material parameters [-]

In general, \( R_{\infty} \) will be greater than zero if material hardens and less than zero if material softens. If the plastic flow if sufficiently large, the isotropic term \( R \) tends asymptotically to \( R_{\infty} \) and therefore can be considered constant. We can get the solution in closed-form for the backstress variable by integration of equation (2), thus:

\[
X_{xz} = \nu \cdot \frac{C}{\gamma} \left( X_{xz}^0 - \nu \cdot \frac{C}{\gamma} \right) \exp \left[ -\frac{\nu \cdot \gamma}{\sqrt{3}} \left( \gamma_{xz} - \gamma_{xz}^0 \right) \right]
\]

from where we can finally derive relation for plastic shear strain increment:

\[
\Delta \gamma_{xz} = -\frac{\sqrt{3}}{\nu \cdot \gamma} \ln \left( \frac{X_{xz} - \nu \cdot \left( C/\gamma \right)}{X_{xz}^0 - \nu \cdot \left( C/\gamma \right)} \right)
\]

where:

\( \nu \) – multiplier, which is defined as \( \nu = 1 \) for loading and \( \nu = -1 \) for unloading

\( X_{xz}^0 \) – initial value of backstress variable [MPa],

\( \gamma_{xz}^0 \) – initial value of plastic shear strain component [-]

Finally, from equation (1) can be expressed relation for shear stress as a function of backstress and isotropic variable, respectively:

\[
r_{xz} = \frac{1}{\sqrt{3}} \left[ X_{xz} + \nu \cdot \left( R + \sigma_I \right) \right]
\]

### 2.1 Shear band cracking mechanism

Shear band cracking mechanism, published by Donzella and Mazzu [6], describes wear as an independent simultaneous phenomenon, which is progressively removing material layers from surface and influences plastic strain accumulation under the contact surface. If is in particular depth reached the critical strain, then all material layers above will be considered as removed material, see Fig.2.
According to [6] there are four layers in material subjected to contact loading: wear layer, layer with fatigue cracks, layer with ratcheting presence and layer with shakedown presence. These layers are moving towards the depth and thus simulating the wear process.

3 INVERSE ALGORITHM FOR MATERIAL PARAMETERS IDENTIFICATION

Identification process of material parameters introduced by Donzella [6] is time and financially challenging. Therefore authors of presented paper were trying to find more effective procedure, which will also utilize experimental dates gained so far.

In our case, the identification of material parameters for finding the correlation between the maximum contact pressure and yield stress has been realized with aim of inverse algorithm. The algorithm is due to the nature of the task incorporating the random number generator. In each step there is, according to initial settings of the algorithm, generated a set of material parameters, which is then used for calculating the worn layer after specified number of cycles. The user can specify the number of reference steps on which has to be calculated worn layer after specified number of cycles. There are usually used three or four reference values, gained from proven linear relationship between diameter loss and realized cycles [6] from the experiment.

According to input parameters such as initial material parameters, defined scatter of material parameters, selected number of reference values of diameter loss with corresponding number of cycles for each cycle, there is made specific number of computations per one cycle, which is dependent on chosen number of reference steps. In each reference step, the program is calculating diameter loss after reference number of cycles and we are also taking into account the number of cycles, if the reference diameter loss has been reached earlier than after reference number of cycles.

After each cycle, which consists of prescribed number of reference steps, the error value is computed and if it is smaller than its previous value, the generated parameters are stored. The error value is computed by following way:

\[
error = \sum_{k=1}^{4} \left( \frac{2 \left| N_{ref} - N_{k} \right| + \frac{h_{W_{ref}} - h_{W_{k}}}{h_{W_{ref}}} \right),
\]

where:
- \(N_{ref}\) – reference number of cycles of k-th step [-]
- \(h_{W_{ref}}\) – reference value of diameter loss for k-th step [mm]
- \(N_{k}\) – number of cycles, needed for reaching \(h_{W_{k}}\) [-]
- \(h_{W_{k}}\) – diameter loss after for \(N_{ref}\) number of cycles [mm]

It is worth to mention, that through the weight of two is taken greater emphasis on the algorithm’s calculated number of cycles to achieve \(N_{k}\), which is needed to achieve diameter loss \(h_{W_{k}}\). The Fig. 3 simply shows flowchart of inverse algorithm.
4 PERFORMED COMPUTATIONAL STUDY

The main aim was to analyze correlation between the parameter $\sigma_L$ (yield stress) and maximum contact pressure with aim of Mazzu/Shear band cracking algorithm and inverse algorithm. Experimental data came from three realized twin-disc experiments with R7T wheel material. The experimental conditions are stated in table 1.
Table 1: Summary of parameters of realized experiments with R7T wheel material

<table>
<thead>
<tr>
<th>Experiment no.</th>
<th>Maximum contact pressure $p_0$ [MPa]</th>
<th>Average value of friction coefficient $f$ [-]</th>
<th>Slip ratio $s$ [%]</th>
<th>Test duration/Number of cycles $N$ [-]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>800</td>
<td>0.24</td>
<td>0.75</td>
<td>100000</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>0.35</td>
<td>0.75</td>
<td>100000</td>
</tr>
<tr>
<td>3</td>
<td>1200</td>
<td>0.33</td>
<td>0.75</td>
<td>100000</td>
</tr>
</tbody>
</table>

On the basis of proven linear relationship between wear and realized cycles [5], authors were allowed to apply the linear extrapolation in order to get the wear data for cycles in range from 250000 to 600000. Reason of choice the reference cycles in mentioned range consisted in detected stabilized deformation profile under the contact surface for stated conditions under which the experiments were carried out. The table 2 contains reference values of diameter loss with corresponding number of cycles, gained with help of linear extrapolation from performed experiments.

Table 2: Reference values of diameter loss with corresponding number of cycles for performed computations

<table>
<thead>
<tr>
<th>$p_0$ = 800 MPa</th>
<th>$h_{W_{ref,1}}$ [mm]</th>
<th>0, 2</th>
<th>$N_{ref,1}$ [-]</th>
<th>250000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_{W_{ref,2}}$ [mm]</td>
<td>0.36</td>
<td>$N_{ref,2}$ [-]</td>
<td>450000</td>
</tr>
<tr>
<td></td>
<td>$h_{W_{ref,3}}$ [mm]</td>
<td>0.48</td>
<td>$N_{ref,3}$ [-]</td>
<td>600000</td>
</tr>
<tr>
<td>$p_0$ = 1000 MPa</td>
<td>$h_{W_{ref,1}}$ [mm]</td>
<td>0.25</td>
<td>$N_{ref,1}$ [-]</td>
<td>235000</td>
</tr>
<tr>
<td></td>
<td>$h_{W_{ref,2}}$ [mm]</td>
<td>0.45</td>
<td>$N_{ref,2}$ [-]</td>
<td>420000</td>
</tr>
<tr>
<td></td>
<td>$h_{W_{ref,3}}$ [mm]</td>
<td>0.6</td>
<td>$N_{ref,3}$ [-]</td>
<td>560000</td>
</tr>
<tr>
<td>$p_0$ = 1200 MPa</td>
<td>$h_{W_{ref,1}}$ [mm]</td>
<td>0.25</td>
<td>$N_{ref,1}$ [-]</td>
<td>225000</td>
</tr>
<tr>
<td></td>
<td>$h_{W_{ref,2}}$ [mm]</td>
<td>0.45</td>
<td>$N_{ref,2}$ [-]</td>
<td>400000</td>
</tr>
<tr>
<td></td>
<td>$h_{W_{ref,3}}$ [mm]</td>
<td>0.61</td>
<td>$N_{ref,3}$ [-]</td>
<td>540000</td>
</tr>
</tbody>
</table>

In the first step, there have been identified material parameters for experiment no. 1, i.e. for maximum contact pressure 800 MPa. Following tables show us initial and optimized material parameters for numerical calculation with $p_0 = 800$ MPa as well as output data (Table 4).

Table 3: Initial and optimized material parameters for experiment no. 1($p_0 = 800$ MPa)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial material parameters</th>
<th>Optimized material parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_0$ [MPa]</td>
<td>150000</td>
<td>174636</td>
</tr>
<tr>
<td>$\gamma_0$ [-]</td>
<td>8</td>
<td>5.54</td>
</tr>
<tr>
<td>$\sigma_{LV}$ [MPa]</td>
<td>150</td>
<td>172</td>
</tr>
<tr>
<td>$R_{\infty 10}$ [MPa]</td>
<td>60</td>
<td>98</td>
</tr>
<tr>
<td>$b_{10}$ [-]</td>
<td>0.03</td>
<td>0.0129</td>
</tr>
<tr>
<td>$R_{\infty 20}$ [MPa]</td>
<td>30</td>
<td>47</td>
</tr>
<tr>
<td>$b_{20}$ [-]</td>
<td>0.0003</td>
<td>0.0005</td>
</tr>
</tbody>
</table>
Table 4: Output data obtained from inverse algorithm for experiment no. 1: \((p_0 = 800 \text{ MPa})\)

<table>
<thead>
<tr>
<th>(h_{W ref 1} [\text{mm}])</th>
<th>0.2</th>
<th>(h_{W 1} [\text{mm}])</th>
<th>0.2</th>
<th>(N_{ref 1} [-])</th>
<th>250000</th>
<th>(N' [-])</th>
<th>236477</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h_{W ref 2} [\text{mm}])</td>
<td>0.36</td>
<td>(h_{W 2} [\text{mm}])</td>
<td>0.36</td>
<td>(N_{ref 2} [-])</td>
<td>450000</td>
<td>(N' [-])</td>
<td>387633</td>
</tr>
<tr>
<td>(h_{W ref 3} [\text{mm}])</td>
<td>0.48</td>
<td>(h_{W 3} [\text{mm}])</td>
<td>0.48</td>
<td>(N_{ref 3} [-])</td>
<td>600000</td>
<td>(N' [-])</td>
<td>529633</td>
</tr>
<tr>
<td>(\text{error} [-])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,2066</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the next two computations, concretely for \(p_0 = 1000 \text{ MPa}\) and \(p_0 = 1200 \text{ MPa}\), the amount of optimized parameters has been restricted to only one parameter, which was the yield stress \(\sigma_l\). Despite the optimization of only one parameter, concretely \(\sigma_l\), there has been still achieved relatively small error value. Table 5 and 6 below contain gained results and also output data from inverse algorithm. Graph 1 shows us the dependance of yield stress on maximum contact pressure.

Table 5: Output data obtained from inverse algorithm for experiment no. 2 \((p_0 = 1000 \text{ MPa})\)

<table>
<thead>
<tr>
<th>Yield stress (\sigma_l [\text{MPa}])</th>
<th>Initial value</th>
<th>Optimized value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h_{W ref 1} [\text{mm}])</td>
<td>0.25</td>
<td>(h_{W 1} [\text{mm}])</td>
</tr>
<tr>
<td>(h_{W ref 2} [\text{mm}])</td>
<td>0.45</td>
<td>(h_{W 2} [\text{mm}])</td>
</tr>
<tr>
<td>(h_{W ref 3} [\text{mm}])</td>
<td>0.6</td>
<td>(h_{W 3} [\text{mm}])</td>
</tr>
<tr>
<td>(\text{error} [-])</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Output data obtained from inverse algorithm for experiment no. 3 \((p_0 = 1200 \text{ MPa})\)

<table>
<thead>
<tr>
<th>Yield stress (\sigma_l [\text{MPa}])</th>
<th>Initial value</th>
<th>Optimized value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h_{W ref 1} [\text{mm}])</td>
<td>0.25</td>
<td>(h_{W 1} [\text{mm}])</td>
</tr>
<tr>
<td>(h_{W ref 2} [\text{mm}])</td>
<td>0.45</td>
<td>(h_{W 2} [\text{mm}])</td>
</tr>
<tr>
<td>(h_{W ref 3} [\text{mm}])</td>
<td>0.61</td>
<td>(h_{W 3} [\text{mm}])</td>
</tr>
<tr>
<td>(\text{error} [-])</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graph 1: Dependence of yield stress on maximum contact pressure
CONCLUSION

A semi-analytical approach, introduced by A. Mazzu [1], has been used to assess the correlation between the yield stress and the value of maximum contact pressure for the case of line contact. It was presented in the previous work [7], that the cyclic plasticity model including simple Armstrong-Frederick kinematic hardening rule can correctly describe the ratcheting behavior of the wheel steel. On the other hand, such simple constitutive model cannot capture correctly cyclic strain curve of the material. Based on incremental theory of plasticity [2], it is necessary to incorporate a memory surface to the constitutive model to better describe the stress-strain behaviour of the material. Results of performed computations show, that the parameter $\sigma_L$ corresponding to yield stress have to increase in Mazzu approach with increasing the Herzian contact pressure, which corresponds to level of loading of a material point. The memory surface seems to be the simplest way to improve the models capability, while preserving the coupling of ratcheting and wear in the calculations.

Acknowledgment

This chapter has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state budget of the Czech Republic. Authors would like to appreciate the assistance of BONATRANS GROUP a.s. Bohumín company during the preparation and realization of experiments.

REFERENCES


TWO DIMENSIONAL SOLUTION OF THE ADVECTION-DIFFUSION EQUATION USING TWO COLLOCATION METHODS WITH LOCAL UPWINDING RBF

JULIAN M. GRANADOS*, CARLOS A. BUSTAMANTE†, HENRY POWER‡ AND WHADY F. FLOREZ†

*Institució Universitaria de Envigado (IUE)  
055420 Envigado, Colombia  
e-mail: jmgranados@correo.iue.edu.co, web page: http://www.iue.edu.co/

†School of Engineering  
Universidad Pontificia Bolivariana  
Campus Laureles, 050031 Medellíin, Colombia  
e-mail: carlos.bustamante@upb.edu.co; whady.florez@upb.edu.co - Web page: http://www.upb.edu.co/

‡School of Mechanical, Materials and Manufacturing Engineering  
University of Nottingham  
University Park, Nottingham NG7 2RD, UK  
e-mail: henry.power@nottingham.ac.uk - Web page: http://www.nottingham.ac.uk

Key words: Advection-Diffusion, Coupled Problems, Collocation Methods, RBF

Abstract. The two-dimensional advection-diffusion equation is solved using two local collocation methods with Multiquadric (MQ)Radial Basis Functions (RBFs). Although both methods use upwinding, the first one, similar to the method of Kansa, approximates the dependent variable with a linear combination of MQs. The nodes are grouped into two types of stencil: cross-shaped stencil to approximate the Laplacian of the variable and circular sector shape stencil to approximate the gradient components. The circular sector opens in opposite to the flow direction and therefore the maximum number of nodes and the shape parameter value are selected conveniently. The second method is based on the Hermitian interpolation where the approximation function is a linear combination of MQs and the resulting functions of applying partial differential equation (PDE) and boundary operators to MQs, all of them centred at different points. The performance of these methods is analysed by solving several test problems whose analytical solutions are known. Solutions are obtained for different Peclet numbers, Pe, and several values of the shape parameter. For high Peclet numbers the accuracy of the second method is affected by the ill-conditioning of the interpolation matrix while the first interpolation
method requires the introduction of additional nodes in the cross stencil. For low $Pe$ both methods yield accurate results. Moreover, the first method is employed to solve the two-dimensional Navier-Stokes equations in velocity-vorticity formulation for the lid-driven cavity problem moderate $Pe$.

1 INTRODUCTION

Recently, the RBFs have been used as the base of meshless collocation approaches for solving PDEs. The use of RBF interpolation technique has become the foundation of the RBF meshless collocation methods for the solution of PDEs, since the pioneer work on the Unsymmetric method by Kansa [1]. Kansa used the MQ function to obtain an accurate meshless solution to the advection-diffusion and Poisson equations without employing any special treatment for the advective term (upwinding), due to the high order of the resultant scheme and the intrinsic relationship between governing equations and the interpolation.

With the aim of improving the Kansa’s Method, Fasshauser [2] used Hermite interpolation to construct an RBF interpolating function which gives a non-singular symmetric collocation matrix. He concluded that the Hermitian (Symmetric) method performs slightly better than the Kansa (Unsymmetric) method. Jumarhon et al. [3] obtained a similar improvement using the Symmetric method and, more recently, Power and Barraco [4] attained better results by employing the Symmetric method for a variety of problems including advection-diffusion equation.

Although full-domain RBF methods are highly flexible and exhibit high order convergence rates, the fully-populated matrix systems they produce lead to the problem described by Shaback [5] as the uncertainty relation; better conditioning is associated with worse accuracy, and worse conditioning is associated with improved accuracy. As the system size is increased, this problem becomes more pronounced. Many techniques have been developed to reduce the effect of the uncertainty relation, such as RBF-specific preconditioners and adaptive selection of data centres. However, at present the only reliable method of controlling numerical ill-conditioning and computational cost as problem size increases is through domain decomposition. One of the first attempts in this direction was made by Lee et al. [6] who proposed the local MQ approximation in which only the nodes inside the influence subdomain of one central node are used in the Unsymmetric method for solving the Poisson equation. Divo and Kassab [7] solved the non-isothermal flow problem by implementing a localized Radial Basis method based on the formulation proposed by Sarler and Vertnik [8] and using a sequential algorithm. Afterwards, Stevens et al. [9] implemented the Local Hermitian Interpolation (LHI) method to solve transient and non-linear diffusion problems. Unlike the method proposed in [7], Stevens et al. solved accurately the two- and three-dimensional advection-diffusion equation by LHI method including the PDE operator and PDE centres in the approximation of the
solution field for each local domain.

On the other hand, many authors have implemented upwinding schemes to avoid oscillations in the solution when dealing with the advective term. As the most common strategy, the stencil or subdomain employed to approximate the value of the variable at a given point contains points that are selected based on the flow direction. In the solution of advection-diffusion problems with Finite Volume Method (FVM) different upwinding schemes such as Upwind Differencing (UD) and QUICK have been widely used [10]. Several authors have reported the use of that kind of upwinding applied to mesh-based methods. Lin and Atluri [11] developed two upwinding schemes, USI and USII, for the Meshless Local Petrov Galerkin (MLPG) method applied to stationary advection-diffusion problems in one- and two-dimensions. In USI scheme, Lin and Atluri propose a circular stencil in which the central node moves a certain distance in the opposite way to the flow direction while in USII the complete stencil is moved. In both cases, the distances are previously defined and they depend on the local Peclet numbers. The authors report best results for USII scheme in most of the cases, mainly at high Reynolds numbers.

The RBF interpolation method has been used to solution different formulations of the Navier-Stokes equations. Several numerical techniques have been reported in the literature to solve viscous flow problems in terms of their velocity-vorticity formulation (Skerget and Rek [12] use a Boundary Element Method (BEM), Huang and Li [13] a Finite Difference Method (FDM) and Young, Liu and Eldho [14] a Finite Element Method (FEM)-BEM coupled scheme). More recently, Hribersek and Skerget [15] deal with complex geometry situations by the Boundary Domain Integral Method (BDIM) for high Reynolds numbers. Zunic et al. [16] use the scheme implemented by Young et al. in [14] for three-dimensional domains. With a similar formulation Pascazio and Napolitano [17] solve the Navier-Stokes equations for transient flow in staggered grids, where velocities are known at the volume faces and the vorticities at the nodes. Qian and Vezza [18] apply the Control Volume Method (CVM) to solve the kinetics equation and the Bio-Savart Law to compute velocities in an iterative time marching algorithm. They also used an additional scheme to compute vorticity values at boundaries. Among others, these are some examples of previously works published in the literature using the velocity-vorticity formulation.

In this paper, two RBF collocation methods with local upwinding to solve advection-dominated problems are implemented. The first method uses circular sector shape stencil for the advective term approximation. The second method is based on the Hermitian interpolation in which the approximation functions are enforced at different locations and the PDE and boundary operators are employed at local level. Both methods are detailed in section 2 and are tested with problems in one- and two-dimensions in section 3.

2 LOCAL COLLOCATION METHOD FOR A SYSTEM OF EQUATIONS

With the vorticity vector, $\vec{\omega}$, understood as the curl of the velocity field,

$$\vec{\omega} = \nabla \times \vec{v},$$
the dimensionless kinematics and kinetics equations for an incompressible fluid in a two-dimensional domain, $\Omega$, are given by

$$\nabla^2 v_1 = -\frac{\partial \omega}{\partial x_2}, \quad \nabla^2 v_2 = \frac{\partial \omega}{\partial x_1}$$

(1)

and

$$\nabla^2 \omega - Re \vec{v} \cdot \nabla \omega = 0,$$

(2)

where $\vec{v} = [v_1, v_2]$. The Reynolds number is defined, in terms of the density, $\rho$, dynamic viscosity, $\mu$, relative velocity, $U$, and relative length, $L$, as $Re = \frac{U L}{\mu}$ and this is equivalent to $Pe$ in problems of section 3. Equation (1) is obtained by applying the curl operator to the vorticity definition and by considering the mass conservation equation, $\nabla \cdot \vec{v} = 0$. Similarly, the curl operator applied to the Navier-Stokes equations in terms of the primitive variables leads to the vorticity transport equation (2). The equations (1)-(2) form the velocity-vorticity formulation of the Navier-Stokes system of equations for two-dimensional steady state flows. On the boundary, $\partial \Omega$, the variables $v_1, v_2$ and $\omega$ satisfy Dirichlet conditions. PDEs linear operators

$$L_1(\cdot) : \nabla^2(\cdot) \quad \text{and} \quad L_2(\cdot) : \nabla^2(\cdot) - Re \vec{v} \cdot \nabla(\cdot),$$

respectively, are associated to (1) and (2) and, if $\phi$ is used instead of $v_1, v_2$ and $\omega$, they are represented by $\mathcal{J}$. Namely,

$$\mathcal{J}(\phi(\vec{x})) = \begin{cases} L_1(\phi(\vec{x})) & \text{if } \vec{x} \in \Omega, \\ B(\phi(\vec{x})) & \text{if } \vec{x} \in \partial \Omega, \end{cases} \quad \text{and} \quad \mathcal{L}(\phi) = \begin{cases} L_1(\phi) & \text{if } \phi = v_i, i = 1, 2, \\ L_2(\phi) & \text{if } \phi = \omega, \end{cases}$$

where $B$ represents the boundary operator, which in this case is the identity operator.

An RBF, $\psi$, is a symmetric function respect to a source point, $\vec{\xi}$, and it is defined in terms of the Euclidean distance, $r$, between the source point and an evaluation point, $\vec{x}$. There are different types of RBF widely referenced in the literature, but in this case only the multiquadric, $\psi(r) = (r^2 + c^2)^{m/2}$, with $m = 1$ is considered. The real constant $c$ is the shape parameter.

The Kansa’s method builds the solution approximation, $\hat{\phi}(\vec{x})$, as the linear combination (3), consisting of two kinds of continuous functions: a polynomial of degree $m - 1$, with
The constant coefficients $\alpha_i$ satisfy the following homogeneity conditions

$$\sum_{i=1}^{N} \alpha_i P_{k}^{m-1}(\vec{x}_i) = 0, \; 1 \leq k \leq NT. \quad (4)$$

For each equation, this approach leads to a system of linear equations with a matrix made of four blocks resulting from the application of the PDE and the boundary operators to equation (3), one block generated by conditions (4) and a last block of zeros. Regarding the fact that this matrix suffers ill-conditioning problems as $N$ becomes very large and the calculations require a large computational cost, the local formulation of the methods is a valid alternative to avoid these problems.

The local collocation consist in selecting a suitable number of points that form a stencil whereby the value of the solution is approximated in one of the selected nodes, main node. In the first method, that we call Modified Stencil RBF (MSRBF) collocation method, $\hat{\phi}$ is calculated with two kinds of stencil: cross shape and circular sector shape (Figure 1). The main node is the one who is at once cross center and sector vertex. The stencil in cross shape is used to approximate the diffusive term, stencil for diffusion, SD, while the advective term is approximated with the stencil in sector circular shape, stencil for advection, SA. The SA opens in upwind direction and the maximum number of nodes within is conveniently selected. If the SA contains only one point then both terms are approximated with SD. If $N_i$ represents the number of nodes in the stencil $i$, $1 \leq i \leq N$, wherein $\vec{x}_i$ is the main node, the interpolation (3) is given by
Evaluation of the interpolation function at all nodes in the stencil allows to express the left side of (1), (2) and the boundary condition at $\vec{x}_i$ in terms of the solution at the remaining nodes of stencil $i$, as shown in the following expression

$$
\hat{\phi}(\vec{x}_i) = \sum_{j=1}^{N_i} \alpha_j \psi_j(r) + \beta_i.
$$

(5)

where $\hat{\phi}_i$, $\Psi_{ij}$ and $\mathcal{J}(\psi^i)$ represent, respectively, a vector formed by unknowns variables at stencil nodes, the interpolation matrix on the stencil $i$ and a vector whose components are obtained by applying the operator $\mathcal{J}$ to the RBFs associated to the node $i$ evaluated at all nodes in the stencil $i$. By applying equation (6) to every stencil, $1 \leq i \leq N$, it is obtain a sparse linear system that is solved to find the solutions $\hat{\phi}(\vec{x}_i)$.
3 Numerical Results

In this section we use MSRBF and LHI methods to solve one- and two-dimensional advection-dominated problems. The equation (2) with unidirectional velocity is solved in a rectangle in section 3.1 and the equation with a source term and skew velocity is solved in a unit square in section 3.2. The results obtained with MSRBF method in the mentioned sections suggest the parameter values employed to solve the lid-driven cavity flow problem at $Re = 100$ and 200 in section 3.3. The Picard iteration is used in order to solve the coupled equations (1)-(2). From an initial vorticity on $\Omega$, a velocity field is obtained from equation (1). Then, this velocity field is used in equation (2) to get a vorticity solution on $\Omega$ with which the next iteration starts. The Root Mean Square (RMS) error \[ \epsilon_{rms} = (\phi_{max} \sqrt{N})^{-1} \left( \sum_{i=1}^{N} [\phi(\vec{x}_i) - \hat{\phi}(\vec{x}_i)]^2 \right)^{1/2} \] is selected to assess the solution error.
3.1 One-dimensional advection-dominated problem

Let’s consider the equation (2) with  $\vec{u} = [u_1, 0]$ where $u_1 = 50, 100, 200$ and $400$ in $\Omega = (0,1) \times (0,0.2)$. The following boundary conditions are imposed

$$\phi = 1, \quad x_1 = 0, \quad 0 < x_2 < 0.2,$$
$$\phi = 2, \quad x_1 = 1, \quad 0 < x_2 < 0.2,$$

and $\frac{\partial \phi}{\partial n} = 0$ at the remaining walls. The analytic solution is given as

$$\phi(x_1, x_2) = 2 - \frac{1 - e^{u_1(x_1-1)}}{1 - e^{-u_1}}.$$

Solutions obtained with the MSRBF method, with a uniform node distribution of $21 \times 5$, $u_1 = 200$ and $c = 0.05$ on the edge of the rectangle ($x_1 \in [0,1]$ and $x_2 = 0$) and on the central line ($x_1 \in [0,1]$ and $x_2 = 0.1$) oscillate as shown in Figure 2 where solutions and relative errors at each node are presented.

The oscillations increase by increasing $Pe$ or aspect ratio, that is, the ratio between the horizontal and vertical distances from one node to its neighbouring nodes. These difficulties are avoided when using a fictitious node strategy. Figure 3 shows the solutions
and relative errors on the edge of the rectangle and on the central line with a nodal distribution of $41 \times 9$, $c = 0.05$ and $u_1 = 100, 200, 400$. The RMS errors obtained by the two methods are shown in Figure 4. The absence of fictitious node in the LHI method leads to greater error on the side of the rectangle. Because of the ease of deployment of the RBFs and the smaller number of nodes required in the interpolations of the MSRBF method, the similarity in the solutions on the center line (Figure 3) shows that the MSRBF is a good option in the one-dimensional case. According to Figure 4a, the LHI method has a higher order convergence with a chosen value of $c$, but simultaneously it is very sensitive to the shape parameter, as shown in Figure 4b. A much more stable behaviour is observed in the MSRBF in comparison to the LHI in a double-precision arithmetic.

3.2 Two-dimensional advection-dominated problem

Regarding equation (2), in this case we take $\vec{u} = [10^6, -10^6]$, $Re = 1$, $\Omega = (0, 1)^2$ and a source term and the Dirichlet boundary conditions in agreement with the exact solution

$$\phi(x_1, x_2) = x_1 \cos(0.5 \pi x_2).$$

Figure 5 shows the RMS errors as a function of the shape parameter for $Pe = 10^6$. Both methods report similar accuracy for $c < 10^{-7}$. The LHI results are more accurate within the range $10^{-7} < c < 10^{-2}$. However, ill-conditioning problems in the local interpolation matrices produce an unstable behaviour of the solution in terms of the shape parameter, making difficult the selection of a suitable value. On the other hand, the MSRBF method presents better behaviour as the shape parameter is increased in agreement with the typical trend of the RBF direct collocation schemes. Moreover, the MSRBF method shows convergence when comparing results presented in Figures 5a and 5b, where the minimum RMS errors are of order $10^{-1}$ and $10^{-3}$, respectively. For relatively high shape parameter values, the LHI method does not produce accurate results according to the mentioned issues.
3.3 Two-dimensional driven cavity flow problem

In this problem, we consider the system of equations (1)-(2) with $Re = 100$ and 200 in $\Omega = (0, 1)^2$, that describes the movement of an incompressible, isothermal and Newtonian fluid that fills a square cavity. The flow field is due to the motion of the upper wall located at $x_2 = 1$, with a dimensionless prescribed velocity $u_1 = 1$ and $u_2 = 0$. On the remaining walls the velocities are zero, i.e. $u_i = 0$ for $i = 1, 2$. The values of the boundary vorticity are computed from the definition. The initial guess value in the algorithm for $Re = 200$ is set to be equal to the obtained solution at $Re = 100$.

Results obtained with MSRBF method are compared with Ghia et al. [21] at $Re = 100$ and Hou et al. [22] at $Re = 200$. Figure 6 shows the velocities on the cavity central lines where the good agreement with the reference can be observed. A $41 \times 41$ nodal distribution is used in all cases.

The achieved vorticity contours for $Re = 100$, Figure 7a), are qualitatively similar to the results presented by Ghia, Figure 7b), and Bustamante et al [23], Figure 7c). In the solution of this problem, every SA has an angle $120^\circ$, the selected radius guarantees no more than seven nodes per SA and the shape parameter has a value of 1.0. The parameter value is in the stability range shown in figure 5b by $51 \times 51$ nodal distribution.

4 CONCLUSIONS

- The two-dimensional advection-diffusion equation has been solved by means of two different upwinding strategies. By using the MSRBF method, accurate results were found with stencils for advection including 3 to 7 nodes and covering an angle between $90^\circ$ and $100^\circ$. The LHI method presents higher order convergence than the MSRBF when solving a one-dimensional problem. Nevertheless, the ill-conditioning problems makes the LHI method more shape parameter sensitive. The MSRBF method is a suitable option regarding a straightforward implementation and a good behaviour in terms of the shape parameter.

- The MSRBF method has been used to solve the two-dimensional driven cavity flow at Reynolds numbers of 100 and 200. The parameter value used in the RBF interpolations was 1.0 and the maximum number of points in every SA was 7 as in the advection-dominated problems. The numerical results are in good agreement with the reference solution. However, the authors are currently working on the improvement of the solution algorithm with the aim of having better results.

REFERENCES


USE OF A BEM MODEL TO SIMULATE UNSTEADY COUPLED HEAT AND MOISTURE FLOW THROUGH POROUS SOLID

L. ŠKERGET*, A. TADEU†
*Faculty of Mechanical Engineering
University of Maribor
Smetanova 17, SI-2000 Maribor, Slovenia
e-mail: leopold.skerget@um.si/

†Faculty of Civil Engineering, ITeCONS
University of Coimbra
Coimbra, Portugal
e-mail: tadeu@itecons.uc.pt

Key words: Heat and mass transport, Porous materials, Moisture uptake in a semi-infinite region, Boundary element method

Abstract. The problem of unsteady coupled moisture and heat energy transport through porous solid is studied numerically using singular boundary integral representation of the corresponding governing equations. Bench mark example moisture uptake in a semi-infinite region is studied numerically.

1 INTRODUCTION

Building materials are in general porous, composed of solid matrices and pores. Within the pores, moisture can exist in any of three thermodynamic states of matter, i.e. vapor, liquid, and solid [1, 2]. However, moisture transport is possible only in the case of gaseous/vapor and liquid states. The main moisture transport mechanisms are vapor diffusion, capillary suction or a combination of both, depending on the moisture content of the material.

In the paper, the numerical model based on boundary element method (BEM) is considered to solve coupled nonlinear heat energy and moisture transport through porous media [8]. Since the singular integral representation is based on the use of an appropriate fundamental solution which incorporates more or less physics of the transport phenomenon such as accumulation and diffusion of the field function, the accurate description of different time and length scales can be accommodated and treated much more accurately in a physically and mathematically justified manner. One serious drawback of the BEM is
that it leads to a fully populated system of equations. However, this can be overcome efficiently by the sub-domain or macro element approach, yielding a sparse system similar to FEM and FDM schemes [7].

The first part of the present work describes the problem. Relevant governing energy and moisture transport differential equations are considered [1, 2]. Next, the corresponding singular integral representation is developed describing nonlinear heat and moisture diffusion in an integral form. The nonlinearity of the coupled diffusion problem is accomplished using an iterative solution strategy, based on an under-relaxation procedure. One benchmark example, e.g. moisture uptake in a semi-infinite region, is given to demonstrate the efficiency and accuracy of the proposed solution strategy [2, 3].

2 Governing equations for two-phase system

Let us consider a two-phase thermodynamic system in a control volume Ω bounded by a control surface Γ, where the indices l and v represent the two phases. For example, l may refer to the liquid water and v to the vapor water in a liquid/vapor moisture system.

2.1 Moisture transport equation

The mass balance equation describing accumulation within the control volume, mass flux in and out of the control volume and generation of a species via reaction/phase change written for the water vapor and liquid water [1, 2]

\[ \rho_m \frac{\partial Y_v}{\partial t} = - \nabla \cdot \vec{j}_v - \dot{m}_C \quad \text{and} \quad \rho_m \frac{\partial Y_l}{\partial t} = - \nabla \cdot \vec{j}_l + \dot{m}_C, \]

(1)

where the dimensionless field functions \( Y_v = m_v/m_m \) and \( Y_l = m_l/m_m \) represent mass fraction or moisture ratio of water vapour and liquid water, respectively, and \( m_m \) represents mass of the dry porous material. The vector quantities \( \vec{j}_v \) and \( \vec{j}_l \) denote the nonconvective water vapour and liquid water mass fluxes, respectively, and \( \dot{m}_C \) represents moisture condensation/evaporation mass rate. The basic governing conservation equation for the moisture flow through a porous solid can now be derived by adding the individual species conservation eqs.(1), yielding

\[ \frac{\partial W}{\partial t} = - \nabla \cdot \vec{j}_v - \nabla \cdot \vec{j}_l, \]

(2)

where the derived potential field function \( W = \rho_m Y_v + \rho_m Y_l = m_w/V_m \) represents moisture content \( m_w = m_v + m_l \) per volume of dry material \( V_m \). The model is based on the assumption that the water transport can be divided into vapor and liquid flows [1].

Using Fick and Darcy constitutive models for expressing the vapour diffusion mass flux \( \vec{j}_v \) and liquid conduction mass flux \( \vec{j}_l \) due to capillary suction and gravity effect, respectively, one can write

\[ \vec{j}_v = -\delta_p(W,T) \nabla p_v \quad \text{and} \quad \vec{j}_l = -D_l(W,T) \nabla p_l + D_l(W,T) \rho_l \vec{g}, \]

(3)
where the transport coefficients $\delta_p$ and $D_l$ are vapour permeability and liquid conductivity of the material, respectively, and the quantity $\vec{g}$ is the gravity acceleration. The transport coefficient $D_l$ is given as follows

$$D_l(W, T) = \frac{k_l(W) \rho_l}{\eta_l(T)},$$

where $k_l$ is the permeability of the material, $\eta_l$ and $\rho_l$ are the dynamic viscosity and mass density of the liquid water, respectively. The liquid flux $\vec{j}_l$ can be rewritten in the form for the suction pressure, $p_{suc} = p_a - p_l$, where $p_a$ is the atmospheric pressure, yielding

$$\vec{j}_l = D_l \nabla p_{suc} + D_l \rho_l \vec{g},$$

where $p_{suc} = p_{suc}[T, p_v, p_s(T)]$ and $p_s$ is the saturation vapour pressure.

One may apply moisture content $W$ for the driving potential and the relation for the water retention curve $W = W(p_l)$. The moisture content gradient can be written as

$$\nabla W = \frac{dW}{dp_l} \nabla p_l$$

where $dW/dp_l$ is the slope of the water retention curve, yielding

$$\vec{j}_l = -D_l \nabla W + D_l \rho_l \vec{g}$$

The eq.(5) or eq.(7) can be rewritten for the $p_v$ and $T$ field functions. The vapour/liquid water two-phase system is in equilibrium when the pore water pressure $p_l$ or the suction pressure $p_{suc}$ and the relative humidity $\phi$ satisfies the Kelvin moisture state relation [1]

$$p_l \approx \frac{RT \rho_l}{M_w} \ln(\phi)$$

or

$$p_{suc} \approx p_a - \frac{RT \rho_l}{M_w} \ln(\phi)$$

with $M_w$, $R$ and $p_a$ being the water molecular mass, the universal gas constant and atmospheric pressure, respectively. Thus, an expression for suction pressure gradient as a function of temperature $T$ and vapour pressure $p_v$ can be obtained by making use of the partial differentiation of eq.(8) as follows

$$\nabla p_{suc} = -\nabla \left( \frac{RT \rho_l}{M_w} \ln\left(\frac{p_v}{p_s}\right) \right) = -\frac{R \rho_l}{M_w} \ln(\phi) \nabla T + \frac{RT \rho_l}{M_w} \frac{dp_s}{p_s} \nabla T - \frac{RT \rho_l}{M_w} \frac{1}{p_v} \nabla p_v$$

$$= -\frac{R \rho_l}{M_w} \left[ \left( \ln(\phi) - \frac{T}{p_s} \frac{dp_s}{dT} \right) \nabla T + \frac{T}{p_v} \nabla p_v \right],$$

resulting in the following form of the constitutive model

$$\vec{j}_l = -D_l \frac{R \rho_l}{M_w} \left[ \left( \ln(\phi) - \frac{T}{p_s} \frac{dp_s}{dT} \right) \nabla T + \frac{T}{p_v} \nabla p_v \right] + D_l \rho_l \vec{g}.$$
Substituting mass flux eqs. (5) into conservation eq. (2) yields the governing moisture transport equation due to water vapour diffusion, liquid water conduction and gravity

\[
\frac{\partial W}{\partial t} = \nabla \cdot \left( \delta_p \nabla p_v \right) - \nabla \cdot \left( D_l \nabla \rho_s \right) - \nabla \cdot \left( D_i \rho_l \mathbf{g} \right).
\] (11)

The moisture transport eq. (11) is comprised of various moisture driving potentials, e.g. moisture content \( W \), partial water vapour pressure \( p_v \) and the suction pressure \( p_{suc} \). These driving potentials can be expressed in terms of a single transport potential. Let us first formulate the moisture transport eq. (11) for the driving potentials relative humidity \( \varphi \) and temperature \( T \) [2]. The chosen potentials are continuous field functions at the interface of two layers of materials having different moisture storage properties (sorption and retention), therefore they are continuous field functions throughout the solution domain. Consequently, all terms in eq. (11) have to be mathematically transformed using relative humidity \( \varphi \) and temperature \( T \) as primitive driving potentials.

Let us consider first the transient term on the right hand side of the eq. (11), when the following expressions may be written

\[
\frac{\partial W}{\partial t} = \frac{dW}{d\varphi} \frac{\partial \varphi}{\partial t} = \theta \frac{\partial \varphi}{\partial t},
\] (12)

where \( \theta = dW/d\varphi \) is the slope of the sorption isotherm \( W = W(\varphi) \). The vapour pressure gradient can be transformed as

\[
\nabla p_v = \nabla \left( p_s(T)\varphi \right) = \varphi \nabla p_s + p_s \nabla \varphi = \varphi \frac{dp_s}{dT} \nabla T + p_s \nabla \varphi.
\] (13)

Using Kelvin eq. (8) the suction pressure gradient in liquid conduction term can be treated as follows

\[
\nabla p_{suc}(T, \varphi) = \frac{\partial p_{suc}}{\partial T} \nabla T + \frac{\partial p_{suc}}{\partial \varphi} \nabla \varphi = -\frac{R \rho_l}{M_w} \ln(\varphi) \nabla T + \frac{T}{\varphi} \nabla \varphi.
\] (14)

Finally, substituting eqs. (12)-(14) into eq. (11) gives

\[
\theta \frac{\partial \varphi}{\partial t} = \nabla \cdot \left( D_\varphi \nabla \varphi \right) + \nabla \cdot \left( D_T \nabla T \right) - \nabla \cdot \left( D_i \rho_l \mathbf{g} \right),
\] (15)

where the primitive variable in eq. (15) is the relative humidity \( \varphi \) while the second and third term on the right side act as nonhomogeneous nonlinear source terms due to temperature gradient and gravity force. Notice that due to the second term the eq. (15) is coupled to the heat energy transport equation. The transport coefficients \( D_\varphi \) and \( D_T \) are given as:

\[
D_\varphi = \delta_p p_s + D_i \frac{R \rho_l T}{M_w} \varphi \quad \text{and} \quad D_T = \delta_p \varphi \frac{dp_s}{dT} + D_i \frac{R \rho_l}{M_w} \ln(\varphi).
\] (16)
If a moisture content $W$ is used as a flow driving potential, using relations
\[ W = W(\varphi) \quad \text{and} \quad \nabla W = \frac{dW}{d\varphi} \nabla \varphi = \theta \nabla \varphi, \]
the eq.(15) can be rewritten as follows
\[ \frac{\partial W}{\partial t} = \nabla \cdot \left( D \nabla W \right) + \nabla \cdot \left( D_T \nabla T \right) - \nabla \cdot \left( D_l \rho_l g \right) \quad \text{and} \quad D = D_\varphi / \theta. \]

2.2 Heat energy transport equation

The heat energy balance equation considers accumulation within the control volume, energy flux (sensitive and latent) in and out of the control volume and heat source/sink term [1, 2, 3], as follows
\[ (c_{pm} \rho_m + c_{pl} W) \frac{\partial T}{\partial t} = -\nabla \cdot \tilde{q} = -\nabla \cdot \tilde{q}_{sens} - \nabla \cdot \tilde{q}_{lat} + I, \]
where the specific capacities $c_{pm}$ and $c_{pl}$ per mass refer to the dry porous material and to liquid water, respectively. The heat capacity of the air/water vapour mixture in the pores is neglected. The sensible heat energy flux $\tilde{q}_{sens}$ can be given by Fourier model
\[ \tilde{q}_{sens} = -\lambda_{eff}(W) \nabla T, \]
while for the latent heat flux $\tilde{q}_{lat}$ one writes the following expression
\[ \tilde{q}_{lat} = [h_e + T(c_{pv} - c_{pl})] \tilde{j}_v = -h_{lat} \delta_p \nabla p_v, \]
where the quantities $h_{lat}$, $h_e$, $c_{pv}$, and $c_{pl}$ denote specific latent enthalpy, specific latent enthalpy of evaporation or condensation, specific heat of water vapour and specific heat of liquid water, respectively. Substituting heat flux eq.(20) and (21) into conservation eq.(19) results in the governing heat energy transport equation
\[ c_{eff} \frac{\partial T}{\partial t} = \nabla \cdot (\lambda_{eff} \nabla T) + \nabla \cdot (h_{lat} \delta_p \nabla p_v) + I, \]
where coefficients $c_{eff}$ and $\lambda_{eff}$ are the effective specific heat per unit volume and effective thermal conductivity, respectively.

3 Boundary-domain integral equations
3.1 Integral representation of heat energy kinetics

Let us first consider the nonlinearity in unsteady energy transport eq.(22) caused by nonlinear transport properties, e.g. specific heat per volume $c_{eff} = c_{eff}(W, T)$ and heat conductivity $\lambda_{eff} = \lambda_{eff}(W, T)$ [6, 9]. They can be partitioned into a constant and a
variable part, e.g. $c_{eff} = c_o + \tilde{c}$ and $\lambda_{eff} = \lambda_o + \tilde{\lambda}$, enabling to partition the eq.(22) into a linear and nonlinear part in the following manner

$$\frac{\partial T}{\partial t} = a_o \frac{\partial^2 T}{\partial x_j \partial x_j} + \frac{1}{c_o} \left[ \frac{\partial}{\partial x_j} \left( \lambda \frac{\partial T}{\partial x_j} + h_{lat} \delta_p \frac{\partial p_v}{\partial x_j} \right) - \frac{c}{\partial T} + I \right], \quad (23)$$

with $a_o = \lambda_o/c_o$. The integral representation of the nonlinear energy diffusion equation can now be derived considering the linear parabolic diffusion differential operator [6]

$$L[T] + b = a_o \frac{\partial^2 T}{\partial x_j \partial x_j} - \frac{\partial T}{\partial t} + b = 0, \quad (24)$$

with the corresponding integral equation written for a time step $\Delta t = t_F - t_{F-1}$

$$c(\xi) T(\xi, t_F) + a_o \int_{t_{F-1}}^{t_F} T q^* dtd\Gamma = a_o \int_{t_{F-1}}^{t_F} q u^* dtd\Gamma + \int_{t_{F-1}}^{t_F} b u^* dtd\Omega + \int_{\Omega} T_{i,F-1} u^*_{F-1} d\Omega, \quad (25)$$

where $q = \partial T/\partial n = q_j n_j$ and $u^*$ is the parabolic diffusion fundamental solution [8]. The domain integral of the pseudo-body forces includes the effects nonlinear transport properties, latent heat and the heat source term, namely

$$b = \frac{1}{c_o} \left[ \frac{\partial}{\partial x_j} \left( \lambda q_j + h_{lat} \delta_p \frac{\partial p_v}{\partial x_j} \right) - \frac{c}{\partial T} + I \right] = \frac{1}{c_o} \left[ \frac{\partial b_j}{\partial x_j} - \frac{c}{\partial T} + I \right], \quad (26)$$

therefore the following integral representation can be obtained

$$c(\xi) T(\xi, t_F) + a_o \int_{t_{F-1}}^{t_F} T q^* dtd\Gamma = a_o \int_{t_{F-1}}^{t_F} q u^* dtd\Gamma + \frac{1}{c_o} \int_{t_{F-1}}^{t_F} b_j n_j u^* dtd\Gamma - \frac{1}{c_o} \int_{\Omega} q_j^* d\Omega + \frac{1}{c_o} \int_{t_{F-1}}^{t_F} \left( -\frac{c}{\partial T} + I \right) u^* dtd\Omega + \int_{\Omega} T_{i,F-1} u^*_{F-1} d\Omega. \quad (27)$$

For the solution of eq.(27), the boundary $\Gamma$ and domain $\Omega$ are discretized into a series of boundary elements and series of internal cells, respectively. Furthermore, functions and their derivatives are assumed to vary within each element or cell and each time step according to the space $\{ \Phi \}$ or $\{ \phi \}$ and time $\{ \Psi \}$ interpolation functions such that

$$T(S, t) = \{ \Phi \}^T \{ \Psi \} \{ T \}^n_m, \quad q(S, t) = \{ \Phi \}^T \{ \Psi \} \{ q \}^n_m, \quad b_j(S, t) = \{ \Phi \}^T \{ \Psi \} \{ b_j \}^n_m, \quad b_j(s, t) = \{ \phi \}^T \{ \Psi \} \{ b_j \}^m, \quad etc., \quad (28)$$

where index $n$ refers to the number of nodes within each element or cell, and the index $m$ refers to the degree of variation of the function $\{ \Psi \}$.
In the discretized eq. (31) there are the following boundary and domain integrals, which

\[ U_1^* = a_o \int_{t_{F-1}}^{t_F} \Psi_1 u^* dt = \frac{1}{4\pi} \left[ \exp(-x_{F-1}) - x_{F-1}E_1(x_{F-1}) \right], \]

\[ U_2^* = a_o \int_{t_{F-1}}^{t_F} \Psi_2 u^* dt = \frac{1}{4\pi} \left[ E_1(x_{F-1}) - \exp(-x_{F-1}) + x_{F-1}E_1(x_{F-1}) \right], \]

\[ Q_1^* = a_o \int_{t_{F-1}}^{t_F} \Psi_1 q^* dt = \frac{x_i(\xi) - x_i(s)}{8\pi a_o \tau} \left[ 1 - \exp(-x_{F-1}) - E_1(x_{F-1}) \right], \]

\[ Q_2^* = a_o \int_{t_{F-1}}^{t_F} \Psi_2 q^* dt = \frac{x_i(\xi) - x_i(s)}{8\pi a_o \tau} \left[ 1 - \exp(-x_{F-1}) - E_1(x_{F-1}) \right]. \]

where \( E_1 \) is the exponential integral function and \( x_{F-1} = r^2/4a_o \tau; \) the eq. (27) can be rewritten as follows

\[ c(\xi)T_2(\xi) + \sum_{m=1}^{2} \sum_{n=1}^{E} \left[ \int_{\Gamma_e} \{\Phi\}^T Q_{nm}^* d\Gamma \right] [T]_m^n = \sum_{m=1}^{2} \sum_{n=1}^{E} \left[ \int_{\Gamma_e} \{\Phi\}^T U_{nm}^* d\Gamma \right] [q]_m^n \]

\[ + \sum_{m=1}^{2} \sum_{n=1}^{E} \left[ \int_{\Gamma_e} \{\Phi\}^T n_j d\Gamma \right] \left\{ \frac{b_j}{\lambda_0} \right\}_m^n - \sum_{m=1}^{2} \sum_{n=1}^{C} \left[ \int_{\Omega_c} \{\phi\}^T Q_{jm}^* d\Omega \right] \left\{ \frac{b_j}{\lambda_0} \right\}_m^n \]

\[ + \sum_{c=1}^{C} \left[ \int_{\Omega_c} \{\phi\}^T u_{F-1}^* d\Omega \right] [T]_{F-1}^n. \]

In the discretized eq.(31) there are the following boundary and domain integrals, which are the functions of geometry, time step and material properties

\[ h_{em}^n = \int_{\Gamma_e} \{\Phi\}^T Q_{nm}^* d\Gamma, \quad g_{em}^n = \int_{\Gamma_e} \{\Phi\}^T U_{nm}^* d\Gamma, \quad c_{em}^n = \int_{\Gamma_e} \{\Phi\}^T n_j d\Gamma, \]

\[ d_{cm} = \int_{\Omega_c} \{\phi\}^T Q_{jm}^* d\Omega, \quad b_c = \int_{\Omega_c} \{\phi\}^T u_{F-1}^* d\Omega, \]

yielding the following discretized form of eq.(31)

\[ c(\xi)T_2(\xi) + \sum_{m=1}^{2} \sum_{n=1}^{E} \left\{ h_{em}^n \right\}_m [T]_m^n = \sum_{m=1}^{2} \sum_{n=1}^{E} \left\{ g_{em}^n \right\}_m [q]_m^n + \sum_{m=1}^{2} \sum_{n=1}^{E} \left\{ c_{em}^n \right\}_m \left\{ \frac{b_j}{\lambda_0} \right\}_m^n \]

\[ - \sum_{m=1}^{2} \sum_{n=1}^{C} \left\{ d_{cm} \right\}_m \left\{ \frac{b_j}{\lambda_0} \right\}_m^n + \sum_{c=1}^{C} \left\{ b_c \right\}^T [T]_{F-1}^n. \]
If the above statement is now applied, using the collocation method, to all boundary and domain nodes, and applying the notation, e.g. \( [H] = [c(\xi)] + [\hat{H}] \) and \( [E_j] = [C_j] - [D_j] \), yielding a nonlinear system of equations

\[
[H]_2\{T\}_2 + [H]_1\{T\}_1 = [G]_2\{q\}_2 + [G]_1\{q\}_1 + \{E_j\}_2\{\frac{b_j}{\lambda_o}\}_2 + \{E_j\}_1\{\frac{b_j}{\lambda_o}\}_1 + [B]\{T\}_1. \tag{34}
\]

### 3.2 Integral representation of moisture kinetics

The integral representation of the moisture parabolic diffusion transport eq.(18) can be derived in a manner as it was obtained for the parabolic diffusion heat energy transport equation, e.g. using linear diffusion differential operator with \( D = D_o + D \), therefore eq.(18) may be reformulated as

\[
L [W] + b = D_o \frac{\partial^2 W}{\partial x_j \partial x_j} - \frac{\partial W}{\partial t} + b = 0, \tag{35}
\]

with the following integral representation

\[
c(\xi) W(\xi, t_F) + D_o \int_\Gamma \int_{t_{F-1}}^{t_F} Wq^* dtd\Gamma = D_o \int_\Gamma \int_{t_{F-1}}^{t_F} \frac{\partial W}{\partial n} u^* dtd\Gamma + \int_\Omega \int_{t_{F-1}}^{t_F} bu^* dtd\Omega + \int_\Omega W_{i,F-1} u_{F-1}^* d\Omega. \tag{36}
\]

The domain integral incorporates two terms, e.g. the first one describes the mass flux due to nonlinear transport diffusivity while the second one describes mass flux due to temperature gradient

\[
b = \frac{\partial}{\partial x_j} \left( D \frac{\partial W}{\partial x_j} + D_T \frac{\partial T}{\partial x_j} \right) = \frac{\partial b_j}{\partial x_j}, \tag{37}
\]

yielding the following integral representation

\[
c(\xi) W(\xi, t_F) + D_o \int_\Gamma \int_{t_{F-1}}^{t_F} Wq^* dtd\Gamma = D_o \int_\Gamma \int_{t_{F-1}}^{t_F} qu^* dtd\Gamma + \int_\Omega \int_{t_{F-1}}^{t_F} bu^* dtd\Omega - \int_\Omega \int_{t_{F-1}}^{t_F} b_j n_j u^* dtd\Omega + \int_\Omega W_{F-1} u_{F-1}^* d\Omega. \tag{38}
\]

#### 3.2.1 Linear time interpolation

Assuming linear variation of all field functions within the individual time increment, the eq. (38) is written as follows

\[
c(\xi) W_2(\xi) + \sum_{m=1}^{2} \sum_{e=1}^{E} \{h\}_m^T \{W\}_m^n = \sum_{m=1}^{2} \sum_{e=1}^{E} \{g\}_m^T \{q\}_m^n + \sum_{m=1}^{2} \sum_{e=1}^{E} \{c\}_m^T \{\frac{b_j}{D_o}\}_m^n - \sum_{m=1}^{C} \sum_{e=1}^{C} \{d\}_m^T \{\frac{b_j}{D_o}\}_m^n + \sum_{c=1}^{C} \{b\}_1^T \{W\}_1^n. \tag{39}
\]
Based on the development of the final discretized eq. (34) for the energy transport, one may state the following discretized representation of the moisture transport

$$\left[H_2\{W\}_2 + H_1\{W\}_1 = [G_2\{q\}_2 + G_1\{q\}_1 + [E_j_2\left\{ \frac{b_j}{D_o} \right\}_2 + [E_j_1\left\{ \frac{b_j}{D_o} \right\}_1 + [B\{W\}_1. \right) (40)$$

4 Numerical algorithm

To decrease storage and CPU time requirements of the single domain BEM we employ the macro element approach [5, 7]. The idea is to use a collocation scheme for integral equation for each domain cell separately and require that the field functions and their normal derivatives must obey some restriction conditions over the domain cell boundaries. Since every domain cell is neighbour only to a few cells we end up with a sparse system of equations. In a nutshell, we are using single domain BEM on every domain cell separately and connect them via compatibility and equilibrium conditions. The heat energy transport equation and moisture transport equation represent coupled nonlinear system of equations which can only be solved iteratively.

5 Numerical example - moisture uptake in a semi-infinite region

The benchmark test example, shown in fig. (1), deals with a $L = 6.0 \text{m}$ or $L = 18.0 \text{m}$ long and $H = 2.0 \text{m}$ width single homogeneous material in equilibrium with a constant surrounding environment [3]. The material is perfectly air tight. At a certain moment the temperature and the relative humidity undergoes a step change.

$$T_S = 30^\circ \text{C} \quad \varphi_S = 0.95$$

$$T_0 = 20^\circ \text{C} \quad \varphi_0 = 0.5$$

One non-uniform non-symmetric mesh of $M = 400 \times 2$ macro elements is used, with the ratio $R_x = 10$ between the largest and the smallest boundary element. Convergence criterion was selected as $\epsilon = 10^{-7}$. The time-dependent analysis was performed by running the simulation from the initial state with a time step value of $\Delta t = 1.0$. The objective is to calculate the moisture and temperature distribution after $t = 7, 30$ and $365$ days. The short solution domain, e.g. $L = 6.0 \text{m}$, is used to simulate time dependent behaviour for 7 and 30 days, while the long solution domain, e.g. $L = 18.0 \text{m}$ is applied for the 365
days simulation. The initial hygrothermal conditions of the structure are temperature \( T_0 = 20^\circ C \) and relative humidity \( \varphi_o = 0.50 \). After the step change the left surface of the structure is exposed to \( \varphi_S = 0.95 \) relative humidity and temperature \( T_S = 30^\circ C \), respectively, while on the right surface the normal derivatives of the corresponding field functions are prescribed to be zero. Therefore, the following boundary conditions of the first kind can be prescribed on the left boundary at \( x = 0 \)

\[
W_S = W(\varphi_S) \text{ and } T_s = T_S \quad \text{on } x = 0 \text{ and } 0 \leq y \leq H \text{ for } t > 0,
\]

(41)

and on all other boundaries the zero boundary conditions of the second kind are prescribed

\[
\frac{\partial W}{\partial n} = 0 \text{ and } \frac{\partial T}{\partial n} = 0 \quad \text{for } t > 0,
\]

(42)

while the initial conditions are

\[
W = W_o(\varphi_o) \text{ and } T = T_o \quad \text{in } \Omega \text{ at } t = 0,
\]

\[
\frac{\partial W}{\partial n} = \frac{\partial W_o}{\partial n} \text{ and } \frac{\partial T}{\partial n} = \frac{\partial T_o}{\partial n} \quad \text{on } \Gamma \text{ at } t = 0.
\]

(43)

The sorption isotherm is given by an expression [3]

\[
W(\varphi) = \frac{146}{[1 + (-8.0 \cdot 10^{-8} R_w T \rho_l \ln(\varphi))^{1.6}]}^{0.375}.
\]

(44)

The vapour diffusivity/permeability \( \delta_p \) and liquid water diffusivity/conductivity \( D_l \) transport coefficients are given by the expressions [3]

\[
\delta_p = \frac{M_w}{RT} \cdot 26.1 \cdot 10^{-6} \cdot \frac{1 - W}{146} \left[ \frac{1}{0.503(1 - \frac{W}{146})^2 + 0.497} \right].
\]

(45)

\[
D_l = \exp \left( -39.2619 + 0.0704(W - 73) - 1.7420 \cdot 10^{-4}(W - 73)^2 \\
-2.7952 \cdot 10^{-6}(W - 73)^3 - 1.1566 \cdot 10^{-7}(W - 73)^4 + 2.5969 \cdot 10^{-9}(W - 73)^5 \right).
\]

(46)

Effective thermal conductivity and specific heat per volume for dry material, respectively, are given by the relations [3]

\[
\lambda_{eff} = 1.5 + \frac{15.8}{1000} W \quad \text{and} \quad c_{pm} \rho_m = c_m = 1.824 \cdot 10^6.
\]

(47)

The numerical simulation results are shown in fig.(2) and fig.(3) for the moisture and temperature distributions, respectively, for a time increment \( \Delta t = 1.0 \text{ day} \) for the moisture and temperature distributions in 7, 30 and 365 days. The numerical results obtained are in excellent agreement with a benchmark results given with the limit values [3] which are marked with \( \circ, \Box, \triangledown \) in the figures.
Figure 2: The moisture distribution at 7, 30 and 365 days: mesh $M = 400 \times 2$; time step $\Delta t = 1.0$ day; ○, □, ▽ - limits of validity for numerical results [3]

6 Conclusions

The boundary element method has been formulated and implemented to solve two-dimensional time dependent coupled nonlinear heat and moisture transfer through porous solid. Quadratic basis functions to approximate field functions and constant interpolation for fluxes are used, while linear variations of all functions over each individual time step is considered. The test benchmark example is highly nonlinear coupled moisture and heat transport, numerical solution of which is severe due to very different time and length scales caused by the difference of several order in the magnitude of heat and mass diffusion coefficients. The linear variation of functions over time step model proved to be very accurate and stable numerical model.

REFERENCES


Figure 3: The temperature distribution at 7, 30 and 365 days: mesh $M = 400 \times 2$; time step $\Delta t = 1.0$ day; $\circ, \square, \triangledown$ - limits of validity for numerical results [3]


OPTIMAL ACTUATOR AND SENSOR PLACEMENT WITH REGARD TO COUPLED ELECTRO-MECHANICAL BEHAVIOUR OF SMART STRUCTURES

TAMARA NESTOROVIC* AND MIROSLAV TRAJKOV†

* e-mail: tamara.nestorovic@rub.de
† email: miroslav.trajkov@rub.de

Key words: Optimal Placement, Smart Structures, Norms.

Abstract. In this paper the problem of optimal actuator and sensor placement for active large flexible structures is considered. The proposed placement optimization method is based on balanced reduced models. It overcomes disadvantages arising from demanding numeric procedures related with high order structural models. Optimization procedure relies on $H_2$ and $H_{\infty}$ norms, as well as on controllability and observability Gramians, related with structural eigenmodes of interest. The optimization procedure is documented by examples, which show a good agreement between the results obtained using different placement indices.

1 INTRODUCTION

The study and development of piezoelectric smart structures involves a very important investigation of optimal actuator and sensor placement. Especially for piezoelectric smart structures and systems, the placement once applied cannot be changed easily and it is often related with the need to build a new structure in order to perform another placement constellation for actuators and sensors. Development of appropriate and reliable optimization procedures, which can be applied prior to real structure or a prototype building, is therefore the task of a great significance. In this paper we have proposed a reliable method for determining appropriate actuator/sensor positions, based on structural models developed using the finite element (FE) approach. Model based approach represents an indispensable tool in the optimization procedure due to requirement for iterative problem solution.

Optimization problem was treated by several authors and investigated for different structures. An overview of the optimization criteria for optimal placement of piezoelectric sensors and actuators on a smart structure was given in a technical review by Gupta et al. [1]. In [2] based on the modal approach, optimal geometrical conditions were obtained for several cases of active beams with different boundary conditions. Optimization criterion for finding optimal actuator/sensor positions for piezoelectric beams in [3] is the performance of an optimal LQR controller. In [4] efficiency indices based on the mode shapes for a clamped piezoelectric beam were determined for typical eigenmodes.

Kumar and Narayanan [5] have applied the LQR controller based criteria to find optimal
location of piezoelectric actuators/sensors for vibration control of plates and used genetic algorithm (GA) for solving a zero-one optimization problem. Peng et al. [6] involved maximizing of the controllability Gramian as the optimization criterion for optimal placement on a clamped plate using GA. Similar approach with modal controllability and observability Gramians and GA were also used in [7].

In this paper we present a general approach to optimal actuator and sensor placement applicable both for beam and plate structures, but also for other complex geometries of structures. The optimal placement procedure is based on the method for balanced model reduction, which assumes models with equally controllable and observable retained modes. The method has advantage over modal truncation and mathematical criteria for controllability and observability, since the retaining of the modes of interest is founded on their equal controllability and observability expressed in terms of appropriate Gramians. Further the paper deals with optimization criteria based on the $H_2$ and $H_\infty$ norms, which are calculated for all possible candidate locations. In this way the fulfillment of the criteria is not limited to a narrow set of selected assumed favorable locations, but it relies on verification through all candidate positions by finding the placement indices with largest values.

2 MODELS AND OBJECTIVE FUNCTIONS FOR OPTIMAL PLACEMENT

The procedure for finding optimal placement of actuators and sensors relies on the state space models of smart structures, which are obtained through the finite element (FE) modeling procedure and model order reduction.

2.1. FE based state space models

Applying general FE modeling procedure the model of a smart structure can be represented as a set of equations of motion in matrix form (1) obtained by assembling all finite elements of the structure (more details on FE modeling of piezoelectric structures can be found in [8,9]).

$$M \ddot{q} + D_0 \dot{q} + Kq = F$$

(1)

Vector $q$ contains all degrees of freedom and it can be formed e.g. by node-wise arranging of degrees of freedom for all elements. For modeling of piezoelectric materials besides mechanical degrees of freedom, electric voltage or charge is included as additional degree of freedom to model electro-mechanical behavior.

The total load vector $F$ is split, for the purpose of the control design later, into the vector of external forces $FE$ and the vector of control forces $FC$:

$$F = F_e + F_c = \tilde{E}\tilde{f}(t) + \tilde{B}\tilde{u}(t) = B_0u$$

(2)

The forces are here generalized quantities, which include also electric charges or electric potentials. Matrices $\tilde{E}$ and $\tilde{B}$ describe the positions of generalized external forces $\tilde{f}$ and the control parameters $\tilde{u}$ in the finite element structure, respectively. Matrix $B_0$ represents the input matrix, and vector $u$ includes all model inputs.

For the controller design purposes equation (2) is accompanied by the output equation in the form:
where in a general case $C_{0q}$ represents the output displacement matrix, and $C_{0v}$ the output velocity matrix. In the output equation (3) $q$ represents a generalized displacement vector containing all degrees of freedom defined in the modeling procedure, like in (1). Matrices $C_{0q}$ and $C_{0v}$ are obtained through an FE procedure by defining appropriate sensor locations.

Solution of the equation (1) is determined in the form $q = \phi \omega t$ by solving the eigenvalue problem for a homogeneous case.

The nodal model representation (1) is transformed into a model in modal coordinates applying the following modal transformation:

$$q = \Phi q_m$$

where $q_m$ represents the vector of modal degrees of freedom or generalized modal displacements and $\Phi$ is the modal matrix.

Introducing the modal coordinates (4) into (1) after normalization with respect to mass and appropriate transformations, taking into account the orthogonality properties the modal model is obtained, which after introducing the coordinate transformation in the state space form:

$$\Omega q_m = \Omega q_m$$

can be obtained as a state space realization:

$$\dot{x} = Ax + Bu, \ y = Cx + Du$$

Considering that flexible structures can be described in terms of independent coordinates, the modal state space model can be expressed in terms of state space realizations $(A_{mi}, B_{mi}, C_{mi})$ for each mode $i$ (7). With the coordinate transformation as in (5) corresponding matrices in the realization $(A_{mi}, B_{mi}, C_{mi})$ are determined by [10]:

$$A_{mi} = \begin{bmatrix} \omega_i & 0 \\ -\omega_i & -2\zeta_i \omega_i \end{bmatrix}, \ B_{mi} = \begin{bmatrix} 0 \\ b_{mi} \end{bmatrix}, \ C_{mi} = \begin{bmatrix} c_{mi} \omega_i \\ c_{mi} \end{bmatrix}$$

with natural eigenfrequencies $\omega_i$ and dampings $\zeta_i$ of the eigenmodes. The elements of the realization $(A_{mi}, B_{mi}, C_{mi})$ are used for assessing the optimal actuator/sensor locations based on candidate input/output transfer functions relating corresponding actuators and sensors.

### 2.2. Norms – objective functions for optimal placement

Optimization of the actuator/sensor placement in this work is based on the properties of the $H_2$ and $H_{\infty}$ norms and approximations for their determining, which enables norm calculation in cases of large structures with high model orders. Exact calculation of the norms in such cases would require high computational effort and computational time. Proposed approach represents a suitable basis for optimal actuator and sensor placement in large structures due to reduced required computational time. The norms and their properties, which are considered and implemented in optimization procedure, are defined for a single mode, for a structure and for a system including a set of actuators and sensors [10], [11]. The main norm properties are summarized below. The proofs are derived in [10].
**H₂ norm of a single mode.** For a transfer function \( G_i(\omega) = C_{mi}(j\omega I - A_{mi})^{-1}B_{mi} \) of the \( i \)th mode obtained from the realization (7), the \( H₂ \) norm of the mode is estimated as:

\[
\|G_i\|_2 \approx \sqrt{\frac{\|B_{mi}\|_2 \|C_{mi}\|_2}{2\sqrt{\zeta_i\omega_i}}} = \sqrt{\frac{\|B_{mi}\|_2 \|C_{mi}\|_2}{2\Delta\omega_i}} \approx \sigma_i \sqrt{2\Delta\omega_i}
\]

where \( B_{mi}, C_{mi} \) represent the input and the output matrices of the modal state space model defined in (7), \( \zeta_i \) is the damping of the \( i \)th mode, \( \sigma_i \) the Hankel singular value corresponding to the \( i \)th mode, and \( \Delta\omega_i = 2\zeta_i\omega_i \) is a frequency segment at the \( i \)th resonance for which the value of the power spectrum is one half of its resonance value.

**H∞ norm of a single mode.** For an \( i \)th mode given by its modal realization \((A_{mi}, B_{mi}, C_{mi})\) or by the parameters \((\omega_i, \zeta_i, b_{mi}, c_{mi})\) the \( H∞ \) norm of the mode is estimated as:

\[
\|G_i\|_\infty \approx \frac{\|B_{mi}\|_2 \|C_{mi}\|_2}{2\zeta_i\omega_i} = \frac{\|B_{mi}\|_2 \|C_{mi}\|_2}{2\zeta_i\omega_i}
\]

**H₂ norm of a structure.** Given a modal state space realization \((A_m, B_m, C_m)\) of a structure, the \( H₂ \) norm of the structure can be determined approximately as the root mean square of the modal norms:

\[
\|G\|_2 \approx \sqrt{\sum_{i=1}^{n} \|G_i\|_2^2}
\]

where \( n \) represents the number of the modes, and \( G \) and \( G_i \) are the transfer function (or the transfer matrix) of the structure and of the \( i \)th mode respectively.

**H∞ norm of a structure.** Since the modes are almost independent, the norm \( H∞ \) norm of a structure is approximately determined as the largest of the mode norms:

\[
\|G\|_\infty \equiv \max_i \|G_i\|_\infty, \quad i = 1, \ldots, n
\]

For a system including a set of actuators and sensor, for the \( H₂ \) and \( H∞ \) norms an additive property both for a single mode and for a structure is valid and can be used in the approximated calculation or the norms.

**H₂ and H∞ norms of a system with a set of actuators and sensors**

For a single mode:

\[
\|G_i\|_{2,\infty} \equiv \sqrt{\sum_{j=1}^{s} \|G_{ij}\|_{2,\infty}^2}, \quad i = 1, \ldots, n
\]

for a structure:

\[
\|G\|_{2,\infty} \equiv \sqrt{\sum_{j=1}^{s} \|G_{ij}\|_{2,\infty}^2}
\]

with \( s \) representing the number of actuators or the number of sensors, which may be different in a general case.

For a given structure the actuator/sensor placement problem requires the selection of
optimal locations as a subset from a given set of possible candidate locations with regard to the specified objective function. The set of possible candidate locations consists of a larger number of elements then the subset of locations to be optimized.

In the first approach the placement is performed based on the placement indices and matrices, where the actuator and sensor placements are solved independently using similar procedures. Definition of placement indices and matrices is based on the additive properties of modal norms on the structural level.

For a flexible structure represented by a modal state space model, the norms of any mode \( i \) are determined based on appropriate input (\( B_{mi} \)) and output (\( C_{mi} \)) matrices of the corresponding mode, (8), (9). If \( s \) represents the total number of defined inputs (actuators) \( j = 1,\ldots,s \), and \( r \) the total number of outputs (sensors) \( k = 1,\ldots,r \), then the corresponding input and output matrices are:

\[
B_{mi} = \begin{bmatrix} B_{mi}^1 & B_{mi}^2 & \cdots & B_{mi}^j & \cdots & B_{mi}^s \end{bmatrix}, \quad C_{mi}^T = \begin{bmatrix} C_{mi}^1 & C_{mi}^2 & \cdots & C_{mi}^j & \cdots & C_{mi}^r \end{bmatrix}
\tag{14}
\]

where each of the matrices \( B_{mi}^j \) represents the \( 2 \times 1 \) block of the \( j \)th actuator and \( C_{mi}^k \) represents the \( 1 \times 2 \) block of the \( k \)th sensor, both having the form as in (7). Then according to the additive properties of the \( H_2 \) and \( H_\infty \) norms, the norm of a mode with a set of actuators (sensors) can be approximated by the root mean square sum of the norms of this mode with a single actuator (sensor), which can be expressed as for actuators and sensors respectively as in (15), (16):

\[
\|G_i\|_{(2,\infty)}^2 \approx \sum_{j=1}^s \|G_{ij}\|_{(2,\infty)}^2
\tag{15}
\]

\[
\|G_i\|_{(\infty,\infty)}^2 \approx \sum_{k=1}^r \|G_{ik}\|_{(\infty,\infty)}^2
\tag{16}
\]

Here the \( H_2 \) norms of the \( i \)th mode with a single actuator corresponding to the \( j \)th position, and of the \( i \)th mode with a single sensor corresponding to the \( k \)th position are given respectively by:

\[
\|G_{ij}\|_{2} = \frac{\|B_{mj}\|_{2} \|C_{mi}\|_{2}}{2\sqrt{\zeta_i \omega_i}}, \quad \|G_{ik}\|_{\infty} = \frac{\|B_{mj}\|_{\infty} \|C_{mi}\|_{\infty}}{2\sqrt{\zeta_i \omega_i}}.
\tag{17}
\]

Similarly the \( H_\infty \) norms of the \( i \)th mode with a single actuator corresponding to the \( j \)th position, and of the \( i \)th mode with a single sensor corresponding to the \( k \)th position are expressed as:

\[
\|G_{ij}\|_{\infty} = \frac{\|B_{mj}\|_{\infty} \|C_{mi}\|_{\infty}}{2\zeta_i \omega_i}, \quad \|G_{ik}\|_{\infty} = \frac{\|B_{mj}\|_{\infty} \|C_{mi}\|_{\infty}}{2\zeta_i \omega_i}.
\tag{18}
\]

Placement indices are defined in terms of \( H_2 \) or \( H_\infty \) norms for an actuator or a sensor placement. Each index \( \eta_{i,j}^{(2,\infty)} \) evaluates the \( k \)th actuator (or sensor) in the \( i \)th mode in terms of the \( H_2 \) or \( H_\infty \) norm and it is defined with respect to all modes \( i = 1,\ldots,n \) and all admissible actuators \( k = 1,\ldots,s \) (or sensors \( k = 1,\ldots,r \)).
Here the norms $\|G_i^k\|_{(2,2)}$ are determined accordingly as in (17) or (18), and $G$ is the transfer function of the system with all candidate actuators (or sensors). Placement indices determined according to (19) can be arranged in the form of matrix, where each row corresponds to the $i^{th}$ mode and each column to the $k^{th}$ actuator or sensor. Actuator and sensor placement indices are then obtained from the placement matrix by performing column-wise appropriate operations on the elements over all modes. For the objective function in terms of the $H_2$ norm, actuator (subscript $a$) or sensor (subscript $s$) placement indices are determined as the root mean square sum of the column-wise elements:

$$\eta_i^{k(a,s)} = \sqrt{\sum_{j=1}^{n} (\eta_i^{k})^2}, \quad k = 1, \ldots, p \tag{20}$$

and $p = s$ (for $s$ actuators) or $p = r$ (for $r$ sensors). For the objective function in terms of the $H_\infty$ norm, the actuator/sensor placement index is the largest index over all modes:

$$\eta_i^{k(a,s)} = \max_i (\eta_i^{k}), \quad i = 1, \ldots, n, \quad k = 1, \ldots, p \tag{21}$$

where again $p = s$ (for $s$ actuators) or $p = r$ (for $r$ sensors). The placement indices $\eta_i^{k(a,s)}$ determined in this way characterize the importance of the $k^{th}$ actuator or sensor, and represent therefore a criterion for the actuator/sensor placement in the presented approach, which treats the actuator and sensor placement individually.

Placement index for simultaneous actuator/sensor placement is defined as

$$\eta_{ij}^{k} = \frac{\|G_{ij}^{k}\|_{(2,2)}}{\|G_{i}^{k}\|_{(2,2)}}, \quad i = 1, \ldots, n \tag{22}$$

for each mode $i$, where $G_{ij}^{k}$ characterizes the $i^{th}$ mode in the presence simultaneously of the actuator placed at the $j^{th}$ candidate location and of the sensor at the $k^{th}$ candidate location.

Besides the introduced placement indices, for the comparison purposes, the controllability index is introduced as an objective function for the optimal placement as well. The influence of the actuators to structural eigenforms is determined by the term $B_m = \Phi^T \hat{B}$, see Eq. (2). Different actuator configurations and their influence on the controllability of the $i^{th}$ mode $\varphi_i$ are investigated by determining the value of $\tau_i(j) = \varphi_i^T \hat{B}_j$ for the $j^{th}$ actuator location. The controllability index is calculated based on the squared value of $\tau_i$ and divided by the scalar product of the eigenvectors, in order to obtain the controllability index as a measure which is independent of the sign influenced by placement and independent of the eigenvector scaling. The controllability index can thus be determined as [19]:

$$\mu_i(j) = \frac{\varphi_i^T \hat{B} \hat{B}_j^T \varphi_i}{\varphi_i^T \varphi_i} \tag{23}$$
In a similar way the influence of the sensor placement can be considered through appropriate observability indices for the kth sensor location:

\[ v_j(k) = \frac{\phi_j^T C_k^T C_k \phi_j}{\phi_j^T \phi_j} \]  

(24)

3 APPLICATION RESULTS

To illustrate the optimization of the actuator/sensor placement, the results of the placement for a clamped piezoelectric beam and plate are presented in this section.

3.1 Clamped beam

In this example a steel beam clamped on both sides is considered. It is modeled as a 2D beam using the ANSYS software. As a result of the modal analysis, the eigenfrequencies and eigenvectors are determined, which represent an input to the algorithms for the optimal actuator/sensor placement procedures. Meshing the beam along its length results in 101 nodes, and possible candidate positions for this analysis are represented schematically in Figure 1 with pointed nodes 10, 20, … , 90.

For the comparison purpose the optimal placement procedure was performed applying the algorithms for separate and simultaneous placement as well as the controllability/observability indices. Several representative examples are presented below.

Figure 1. Candidate locations for actuator/sensor placement along the beam clamped on both sides

Figure 2. Placement indices calculated based on the \( H_2 \) norm for the first five eigenmodes
Qualitative representations of the curves presenting the values of the placement indices for different positions along the beam are similar for separate placement based on the $H_2$ and $H_\infty$ norms. Depending on the number of eigenmodes, which should be considered (sensed or actuated) at the same time, the positions for optimal actuator/sensor placement may differ. Figure 2 shows different possible candidate positions with largest placement indices calculated based on the $H_2$ norm under consideration five bending eigenmodes of interest.

![Figure 2](image.png)

**Figure 2.** Placement indices based on the $H_2$ norm for separate and parallel consideration of the eigenmodes

Placement indices determined based on the $H_\infty$ norm are represented in Figure 3. Left hand side plot in represents the placement indices for individually considered eigenmodes 1 to 5. In the right hand side plot the placement indices were calculated based on parallel consideration of several eigenmodes of interest (here 1 to 5). Locations with largest placement indices indicate the candidates for optimal placement, depending on the number of employed actuators/sensors and on the number of considered modes of interest. Figures 2 and 3 represent the sensor placement indices. The forms of the placement indices curves for actuators are qualitatively the same and for the reason of brevity are omitted here. For the comparison, the method based on the controllability/observability indices is also applied. The results regarding the first five eigenmodes of the beam are summarized by the controllability index representation in Figure 4.

![Figure 3](image.png)

**Figure 3.** Placement indices based on the $H_\infty$ norm for separate and parallel consideration of the eigenmodes

![Figure 4](image.png)

**Figure 4.** Controllability indices calculated for eigenmodes 1 to 5 for different candidate locations
The results of the three methods applied to the beam clamped on both ends are summarized in Table 1. It can be seen that all three methods provide identical results, when considering eigenmodes individually. For parallel consideration of several eigenmodes of interest, optimal candidate locations depend on the performance index which was adopted as a criterion for placement.

Table 1. Candidate locations with largest placement indices (beam clamped on both sides)

<table>
<thead>
<tr>
<th>Modes</th>
<th>Separate placement</th>
<th>Simultaneous placement</th>
<th>Controllability/observability indices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_2$</td>
<td>$H_{\infty}$</td>
<td>$H_2$</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>29, 71</td>
<td>29, 71</td>
<td>29, 71</td>
</tr>
<tr>
<td>3</td>
<td>21, 79</td>
<td>21, 79</td>
<td>21, 79</td>
</tr>
<tr>
<td>4</td>
<td>16, 84</td>
<td>16, 84</td>
<td>16, 84</td>
</tr>
<tr>
<td>5</td>
<td>13, 87</td>
<td>13, 87</td>
<td>13, 87</td>
</tr>
<tr>
<td>1, 2</td>
<td>43 to 57</td>
<td>50</td>
<td>34, 35, 36, 64, 65, 66</td>
</tr>
<tr>
<td>1 to 3</td>
<td>48 to 52</td>
<td>50</td>
<td>26, 74</td>
</tr>
<tr>
<td>1 to 4</td>
<td>47 to 53</td>
<td>50</td>
<td>21, 79, 41, 59</td>
</tr>
<tr>
<td>1 to 5</td>
<td>49 to 51</td>
<td>50</td>
<td>17, 83, 69, 31</td>
</tr>
</tbody>
</table>

3.2 Clamped plate

The plate structure in this example was modeled as a 3D plate in ANSYS software and corresponding eigenvectors of interest were obtained through modal analysis. The meshing of the plate, i.e. the nodes which correspond to candidate locations for actuator/sensor placement are represented in Figure 5. Here the corresponding rows and columns are numerated for a better preview.

Figure 5. Candidate locations for the plate denoted by corresponding row and column numbers

Due to a very high number of nodes, i.e. candidate locations for the plate, the simultaneous placement procedure would not give a clear representation and therefore it is omitted from this analysis. The results of other two methods, separate placement and controllability index, are compared and summarized in Table 2. Besides, several representative results of the
actuator/sensor placement for the clamped plate are shown in the figures below. Complete agreement of the results is available for individual consideration of the eigenmodes. For parallel consideration of several structural eigenmodes of interest, the arising differences are based on the calculation, i.e. on the definition of the placement indices for the structure. Qualitative representations of the placement indices based on $H_2$ and $H_\infty$ norms as well as of the controllability index for individually considered modes are the same. Actuator placement indices based on the $H_2$ norm for selected individual modes are represented in Figure 6.

**Table 2.** Candidate locations with largest placement indices (plate)

<table>
<thead>
<tr>
<th>Modes</th>
<th>Separate placement</th>
<th>$H_2$</th>
<th>$H_\infty$</th>
<th>Controllability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(10,15)</td>
<td>(10,15)</td>
<td>(10,15)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(10,8), (10,9),</td>
<td>(10,8), (10,9),</td>
<td>(10,8), (10,9),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10,21), (10,22)</td>
<td>(10,21), (10,22)</td>
<td>(10,21), (10,22)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(14,15), (6,15)</td>
<td>(14,15), (6,15)</td>
<td>(14,15), (6,15)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(10,6), (10,15),</td>
<td>(10,6), (10,15),</td>
<td>(10,6), (10,15),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10,24)</td>
<td>(10,24)</td>
<td>(10,24)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(6,8), (6,22),</td>
<td>(6,8), (6,22),</td>
<td>(6,8), (6,22),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(14,8), (14,22)</td>
<td>(14,8), (14,22)</td>
<td>(14,8), (14,22)</td>
<td></td>
</tr>
<tr>
<td>1, 2</td>
<td>(10,9), (10,10),</td>
<td>(10,15)</td>
<td>(10,10), (10,20)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10,21)</td>
<td></td>
<td>(10,21)</td>
<td></td>
</tr>
<tr>
<td>1 to 3</td>
<td>(9,9), (9,10),</td>
<td>(10,15)</td>
<td>(7,11), (7,19),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9,20), (9,21)</td>
<td></td>
<td>(13,11), (13,19)</td>
<td></td>
</tr>
<tr>
<td>1 to 4</td>
<td>(10,7), (10,22)</td>
<td>(10,15)</td>
<td>(12,8), (8,8),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(12,22), (8,22)</td>
<td></td>
</tr>
<tr>
<td>1 to 5</td>
<td>(7,8), (13,8),</td>
<td>(10,15)</td>
<td>(13,8), (7,8),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7,22), (13,22)</td>
<td></td>
<td>(13,22), (7,22)</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.** Placement indices based on $H_2$ norm for individually considered selected eigenmodes of the plate

Figure 7 represents the values of the placement indices calculated for all selected candidate locations based on the $H_2$ norm under parallel consideration of several eigenmodes of interest (*left*: modes 1 and 2; *right*: modes 1 to 5).
Figure 7. Placement indices for the plate based on the $H_2$ norm (parallel consideration of eigenmodes of interest)

4 CONCLUSIONS

In this paper the optimization methods for actuator/sensor placement for large flexible structures are presented, based on balanced reduction of structural models. Balanced modal reduction of the model orders for structures with large numbers of degrees of freedom is proposed as an efficient modeling procedure, which results in a realization with equally controllable and observable retained states. Optimal placement procedure is based on the properties of the $H_2$ and $H_\infty$ norms and approximations for their determining. Proposed approach represents a suitable basis for optimal actuator and sensor placement in large structures due to reduced required computational time.

Optimization procedure is proven by showing examples of a beam clamped on both sides and clamped plate. For these examples an extensive analysis was conducted and systematized results of separate and simultaneous placement procedures for individual and parallel consideration of the structural modes are shown. The efficiency of the proposed method is also proven by the comparison with the optimization results based on controllability and observability indices. This analysis has shown a compete agreement of the results. The method suggested in this paper also covers a broad spectrum of possible problems, which do not have to be necessarily limited only to piezoelectric actuators and sensors, but can be
extended more generally to systems with integrated actuators and sensors, whose effect may be considered through actuation forces or moments.

ACKNOWLEDGEMENT
The authors gratefully acknowledge the funding by the German Research Foundation (DFG) within Collaborative Research Center under grant SFB-837/A2.

REFERENCES
PROBABILITY AND VARIANCE-BASED STOCHASTIC DESIGN OPTIMIZATION OF A RADIAL COMPRESSOR CONCERNING FLUID-STRUCTURE INTERACTION

DIRK ROOS*, KEVIN CREMANNS & TIM JASPER

Institute of Modelling and High-Performance Computing
Niederrhein University of Applied Sciences
Reinarzstr. 49, D-47805 Krefeld, Germany
e-mail: dirk.roos[@]hs-niederrhein.de

Key words: robust design optimization, robustness evaluation, reliability analysis, fluid-structure interaction, surrogate models, adaptive design of experiment, importance sampling, directional sampling

Abstract. Since the engineering of turbo machines began the improvement of specific physical behaviour, especially the efficiency, has been one of the key issues. However, improvement of the efficiency of a turbo engine, is hard to archive using a conventional deterministic optimization, since the geometry is not perfect and many other parameters vary in the real approach.

In contrast, stochastic design optimization is a methodology that enables the solving of optimization problems which model the effects of uncertainty in manufacturing, design configuration and environment, in which robustness and reliability are explicit optimization goals. Therein, a coupling of stochastic and optimization problems implies high computational efforts, whereby the calculation of the stochastic constraints represents the main effort. In view of this fact, an industrially relevant algorithm should satisfy the conditions of precision, robustness and efficiency.

In this paper an efficient approach is presented to assist reducing the number of design evaluations necessary, in particular the number of nonlinear fluid-structure interaction analyses. In combination with a robust estimation of the safety level within the iteration and a final precise reliability analysis, the method presented is particularly suitable for solving reliability-based structural design optimization problems with ever-changing failure probabilities of the nominal designs.

The applicability for real case applications is demonstrated through the example of a radial compressor, with a very high degree of complexity and a large number of design parameters and random variables.
1 INTRODUCTION

1.1 Stochastic design optimization

In engineering problems, randomness and uncertainties are inherent and may be involved in several stages, for example in the system design with material parameters and in the manufacturing process and environment. Stochastic optimization, also referred to as reliability-based and variance-based optimization is known as the most adequate and advantageous methodology for system or process design and aims at searching for the best compromise design between design improvement and robustness or reliability assurance, considering uncertainties of geometry, material, manufacturing and process. Herein, the optimization process is carried out in the space of the design parameters and the robustness evaluation and reliability analysis are performed in the space of the random variables. Consequently, during the optimization process the design variables are repeatedly changed, whereby each design variable vector corresponds to a new random variable space. Therefore usually, a high number of numerical calculations are required to evaluate the stochastic constraints at every nominal design point. This repeated search becomes the main problem, especially when numerical nonlinear multi-domain simulations and CAD models are involved.
Unfortunately, in real case applications of the virtual prototyping process, it is not always possible to reduce the complexity of the physical models to obtain numerical models which can be solved quickly. Usually, every single numerical simulation takes hours or even days. Although progress has been made in identifying numerical methods to solve stochastic design optimization problems and high performance computing, in cases such as those that have several nested numerical models, as shown in Fig. 1, the actual costs of using these methods to explore various model configurations for practical applications is too high. Therefore, methods for efficiently solving stochastic optimization problems based on the introduction of simplifications and special formulations for reducing the numerical efforts are required.

1.2 Application to aerodynamic optimization

In comparative studies on the application of the deterministic optimization for aerodynamic optimization (see e.g. Müller-Töws, 2000, Sasaki et al., 2001, Shahpar, 2000) usually stochastic programming algorithms or response surface methods (see e.g. Pierret and van den Braembussche, 1999) are used in turbomachinery design, for example in the development of engine components, such as at Vaidyanathan et al. (2000). In Shyy et al. (2001) a comprehensive overview is represented.

An example of an applied aerodynamic deterministic optimization using a genetic algorithm is published in Trigg et al. (1997) and the optimized design of transonic profiles also using genetic algorithms is given in Oyama (2000). Another very comprehensive study of the use of the combination of genetic algorithms and neural networks for two-dimensional aerodynamic optimization of profiles is presented in Dennis et al. (1999) combine a genetic algorithm with an gradient-based optimization method.

Furthermore, an increasing application of stochastic analysis on turbo machinery (e.g. at Garzon, 2003, Garzon and Darmofal, 2003, Lange et al., 2010, Parchem and Meissner, 2009) underlines the importance of integrating the uncertainty analysis into the aerodynamic design process.

2 RELIABILITY AND VARIANCE-BASED DESIGN OPTIMIZATION

2.1 Deterministic optimization

Optimization is defined as a procedure to achieve the best outcome of a given objective function (sometimes also called cost function) while satisfying certain re-
Figure 2: Different solution points \( \tilde{d}_i \) or \( d_i \) as result of a deterministic vs. stochastic design optimization in the space of given randomly distributed design parameters.

Figure 3: Comparison of the deterministic optimal point \( \tilde{d}_i \) and the solution of a stochastic optimization \( d_i \) with corresponding most probable failure point \( x_j^* \) in the space of the randomly distributed von Mises stress and the yield stress.

restrictions. The deterministic optimization problem

\[
\begin{align*}
f(d_1, d_2, \ldots, d_n) & \rightarrow \min \\
e_l(d_1, d_2, \ldots, d_n) & = 0; \quad l = 1, n_e \\
u_m(d_1, d_2, \ldots, d_n, \gamma) & \geq 0; \quad m = 1, n_u \\
d_l & \leq d_i \leq d_u \\
d_i & \in [d_l, d_u] \subset \mathbb{R}^n_d
\end{align*}
\]

is defined by the objective function \( f : \mathbb{R}^n_d \rightarrow \mathbb{R} \) subject to the restrictions, defined as equality and inequality constraints \( e_l \) and \( u_m \). The variables \( d_1, d_2, \ldots, d_n \) are the optimization or design variables and the vector of the partial safety factors \( \gamma \) ensures the system or design safety within the constraint equations \( u_m \), for example defining a safety distance \( u(d, \gamma) = y_g / \gamma - y_d \geq 0 \) between a defined limit state value \( y_g \) and the nominal design value \( y_d \) of a physical response parameter \( y = f(d) \). In structural safety assessment, a typical constraint for the stress is given as

\[
u(d, \gamma) = \sigma_{y,k} / \gamma - \sigma_d \geq 0
\]
ensuring the global safety distance

\[
\Delta_\gamma = \sigma_{y,k} - \sigma_d = \sigma_{y,k} - \frac{\sigma_{y,k}}{\gamma} = \sigma_{y,k} \left(1 - \frac{1}{\gamma}\right)
\]

between the defined quantile value \(\sigma_{y,k}\) of the yield stress and the nominal design stress \(\sigma_d\) with the global safety factor \(\gamma\), as shown in Fig. 3. Whereby, in the real approach with given uncertainties, \(\sigma_d\) corresponds to the mean von Mises equivalent stress \(\bar{\sigma}_e\) at the current design point.

### 2.2 Stochastic chance-constrained optimization

Stochastic optimization algorithms use the quantification of uncertainties to produce solutions that optimize the expected performance of a process or design, ensuring the target variances of the model responses and failure probability. So, the deterministic optimization problem (1) can be enhanced by additional stochastic restrictions. For example, the expression for system reliability

\[
1 - \frac{P(F)}{P^t(F)} \geq 0 \tag{3}
\]

denotes that the probability of failure

\[
P(F) = P\{\{X : g_k(x) \leq 0\}\} = \int_{g_k(x) \leq 0}^{n_r} \int f_X(x) dx
\tag{4}
\]

cannot exceed a given target probability \(P^t(F)\), considering the vector of all random influences

\[
X = [X_1, X_2, ..., X_{n_r}]^T
\tag{5}
\]

with the joint probability density function of the random variables \(f_X(x)\) and \(k = 1, 2, ..., n_y\) limit state functions \(g_k(x) \leq 0\).

These enhancements of the problem (1) are usually referred to reliability-based design optimization, in which we ensure that the design variables \(d_i\) satisfy the given constraints (3) to some specified probabilities. As a consequence, now the design parameters

\[
d = E[X] \tag{6}
\]

are the means of the \(n_r\) random influences \(X\) with every changing density function during the optimization process. As a result of the random influences, now the objective and the constraints are non-deterministic functions.
2.3 Reliability analysis using adaptive response surface method

For an efficient probability assessment of \( P(\mathcal{F}) \), according to Eq. (4), a multi-domain adaptive design of experiment in combination with directional sampling (see e.g. Ditlevsen et al., 1990) is introduced in Roos (2011) to improve the accuracy and predictability of surrogate models, commonly used in applications with several limit state conditions. Furthermore, the identification of the failure domains using the directional sampling procedure, the pre-estimation and the priori knowledge of the probability level is no longer required. Therefore this adaptive response surface method is particularly suitable to solve reliability-based design optimization problems considering uncertainties with ever-changing failure probabilities of the nominal designs.

However, a reliability analysis method based on surrogate models, is generally suitable for a few random variables only. In case of the proposed probability assessment method, an efficient application is given up to \( n_r = 10, ..., 25 \), depending on the number of relevant unsafe domains. Therefore, a variance-based sensitivity analysis should be used to find a reduced space of the important random influences.

2.4 Global variance-based sensitivity analysis

In general, complex nested engineering models, as shown in Fig. 1 contain not only first order (decoupled) influences of the design parameters or random variables but also higher order (coupled) effects on the response parameter of a numerical model. A global variance-based sensitivity analysis, as introduced by Saltelli et al. (2008), can be used for ranking variables \( X_1, X_2, \ldots, X_{n_r} \) with respect to their importance for a specified model response parameter

\[ Y = f(X_1, X_2, \ldots, X_{n_r}) \]

depending on a specific surrogate model \( \hat{Y} \). In order to quantify and optimize the prognosis quality of these meta models, in Most and Will (2008) and Most (2011) the so-called coefficient of prognosis

\[ \text{COP} = \left( \frac{\mathbb{E}[Y_{\text{Test}} \cdot \hat{Y}_{\text{Test}}]}{\sigma_{Y_{\text{Test}}} \sigma_{\hat{Y}_{\text{Test}}}} \right)^2 ; \quad 0 \leq \text{COP} \leq 1 \quad (7) \]

of the meta model is introduced. In contrast to the commonly used generalized coefficient of determination \( R^2 \) based on a polynomial regression model, in Eq. (7) variations of different surrogate models \( \hat{Y} \) are analyzed to maximize the coefficient of prognosis themselves. This procedure results in the so-called meta model of optimal
value \( y \) of the random response \( Y \)

Figure 4: Relationship between density function \( f_Y(y) \) of a model response, sigma level and exceedance probability, depending on chosen limit state conditions \( y_u^d \).

Figure 5: Convergence of a sequential stochastic chance-constrained optimization with successive interpolation of the nominal response limit \( y_d \) to ensure a target sigma level \( \sigma_L^t \).

prognosis, used as surrogate model \( \tilde{Y} \) with the corresponding input variable subspace which gives the best approximation quality for different numbers of samples, based on a multi-subset cross validation obtained by latin hypercube sampling (see e.g. Huntington and Lyrintzis, 1998).

The single variable coefficients of prognosis are calculated as follows

\[
\text{COP}_i = \text{COP} \cdot \tilde{S}_T_i
\]

with the total sensitivity indices

\[
\tilde{S}_T_i = \frac{E(V(\tilde{Y}|X_{\sim i})))}{V(\tilde{Y})} (9)
\]

which have been introduced by Homma and Saltelli (1996), where \( E(V(\tilde{Y}|X_{\sim i})) \) is the remaining variance of \( \tilde{Y} \) that would be left, on average, if the parameter of \( X_i \) is removed from the model. In Eq. (9) \( X_{\sim i} \) indicates the remaining set of input variables.

2.5 Probability estimation based on moments

For an accurate calculation of the reliability it would be interesting to expand the probability density function of the model responses about a critical threshold. Unfortunately, the density functions are unknown, especially close to the unsafe domain with high failure probability. Existing methods such as polynomial expansions,
maximum entropy method or saddlepoint expansion, as reviewed in Hurtado (2008), are frequently used within the reliability-based structural optimization replacing the expensive reliability analysis.

A more simple, non-intrusive approach for a rough estimation of the failure probability is the calculation of the minimal sigma level \( \sigma_L \) for a performance-relevant random response parameter \( Y \) defined by an upper and lower limitstate value \( y_{g}^{u,l} := \{ Y | g(X) = 0 \} \) as follows

\[
E[Y] \pm \sigma_L \cdot \sigma_Y \lesssim y_{g}^{u,l}
\]

The sigma level can be used in conjunction with standard deviation to measure the deviation of response values \( Y \) from the mean \( E[Y] \). For example, for a pair of quantiles (symmetrical case) and the mean value we obtain the assigned sigma level

\[
\sigma_L = \frac{y_{g} - E[Y]}{\sigma_Y}
\]  

(10)

of the limit state violation, as explained in Fig. 4. Therewith, the non-exceedance
probability results in

\[ P(\mathcal{E}) = P(\{Y|Y \leq y^u_y\}) = f(\sigma_L) \]

as a function of the sigma level, depending on the current distribution type of \( Y \). In
the same manner failure probability

\[ P(\mathcal{F}) = P(\{Y|Y > y_g\}) = f(\sigma_L) \] (11)

is given as a function of the sigma level. For example, assuming a normal distribution
of the random response \( Y \) with \( \mu_Y = 0 \) and \( \sigma_Y = 1 \), as shown in Fig. 6, the failure
probability is given as a nonlinear function

\[ P(\mathcal{F}) = \Phi(-\sigma_L) = \Phi(-y_g) \]

of the sigma level, as illustrated in Fig. 7. Therewith, a probability of \( P^i(\mathcal{F}) = 3.4 \cdot 10^{-6} \) is achieved when the performance target \( \sigma_L^i \) is 4.5 \( \sigma \) away from the mean value.

Other values of acceptable annual probabilities of failure \( P^i(\mathcal{F}) \) depending on the
consequence of failure, significance warning or without warning before occurrence of
failure and (non-)redundant structures can be found in engineering standards, e.g.

2.6 Methods solving stochastic optimization problems

In general, problem (1) to (6) is solved as a combination of a deterministic op-
timization in the \( n_d \)-dimensional design space and a stochastic analysis in the \( n_r \)-
dimensional random space. Derivative-free global optimization methods are typically
recommended to solve the sequential deterministic optimization problem, according
to Eq. (1) for highly nonlinear numerical models, especially fluid-structure interaction
models with probability-based constraints, whose objective and constraint function
value may be computed with some noise or are non-computable in any design points.

Evolutionary computation, as a special class of global optimization strategies, imi-
titates the natural processes like biological evolution or swarm intelligence. Based on
the principle “survival of the fittest” a population of artificial individuals searches
the design space of possible solutions in order to find a better solution for the opti-
mization problem. In this paper an evolution strategy using a class of evolutionary
algorithms is used. This strategy uses normally distributed mutations, recombi-
nation, selection of the best offspring individuals, and the principle of self-adaptation
of strategy parameters, as described in Bäck (1995).

As an alternative derivative-free optimization method, especially useful for large
real-life design optimization problems in which the objectives and constraints are
Figure 8: Basic concept of a decoupled loop of a reliability-based and variance-based stochastic design optimization using global variance-based sensitivity analysis and robustness evaluation to reduce the design parameter and random variable space.

determined as a result of very expensive numerical computations, we use the adaptive response surface methodology, as introduced in Etman et al. (1996), Toropov and Alvarez (1998), Abspoel et al. (1996), Stander and Craig (2002), Kurtaran et al. (2002).

Mainly, there are three methods for solving these kinds of coupled problems (1) to (6). The simplest and most direct solution method is a coupled approach in which a full reliability analysis is performed for every optimization function evaluation (see e.g. Choi et al., 2001). This involves a nesting of two distinct levels of optimization within each other, one at the design level and one at the reliability analysis level. However, despite progress in numerical methods to solve optimization problems and stochastic problems, this coupled procedure leads in general to an inefficient double loop with a large number of design evaluations. Alternatively, one way to overcoming this dilemma would be to use sensitivity analysis to analytically compute the design gradients.

The single-loop method (see e.g. Kharmanda et al., 2002) simultaneously minimizes the objective function and searches for the \( \beta \)-point, satisfying the probabilistic constraints only at the optimal solution, but needs a sensitivity analysis to analyti-
cally compute the design gradients of the probability constraint.

An alternative method, used in the following, is the sequential approach (see e.g. Chen et al., 2003). The general concept is to iterate between optimization and uncertainty quantification, updating the constraints based on the most recent probabilistic assessment results, using safety factors or other approximation methods. This effective iterative decoupled loop approach can be enhanced by updating the constraints during the internal optimization using sigma levels and statistical moments

\[
\frac{\sigma_{L_k}}{\sigma_L} - 1 \geq 0; \quad \sigma_{L_k} = \frac{y_{gk} - E[Y_k]}{\sigma_{Y_k}}; \quad k = 1, n_g
\]

in place of the exceedance probability of the Eq. (3). Essentially, by means of transformation in Eq. (11) of the probability-based highly nonlinear and non-differentiable constraints to linear ones, these functions may be more well conditioned for the optimization approach and we can expect a better performance of the solution process. Of course, the transformation in Eq. (11) can only be used as a rough estimation of the safety level and we have to calculate the probabilities of failure using the reliability analysis, at least at the iteration end.

As shown in Fig. 8, in the initial iteration step a variance-based sensitive analysis identifies the most important multivariate dependencies and design parameters. After this, the deterministic optimization step results in an optimal solution for which the sigma level is calculated using a robustness evaluation, based on a latin hyper-cube sampling. The size of violation of the target sigma level is used to interpolate the constraints using modified safety factors. Whereby, as an important fact, the interpolation order increases continuously with each iteration step, so in practice three or four iteration steps may meet our optimization requirements in terms of robustness and safety. Fig. 5 shows a typical convergence of a sequential stochastic chance-constrained optimization.

Furthermore, the optimization steps and the final reliability analysis run mostly efficiently in the space of the current significant parameters. So every size of problem definition (number of design and random parameters) is solvable within all sigma levels.

The following numerical example with a very high degree of complexity is given to demonstrate the solving power of this sequential stochastic chance-constrained optimization by adapting the constraint \(u_m(d, \gamma)\) depending on interpolated nominal response values \(y_d\).
Figure 9: Parametric CAD model of a one stage radial compressor, consisting of an impeller and returnvane.

3 NUMERICAL EXAMPLE

3.1 Computational fluid dynamics and process integration

Computational fluid dynamics (CFD) is an engineering method in which flow fields and other physics are calculated in detail for an application of interest. ANSYS, which is used for the following example, uses a multidisciplinary approach to simulation in which fluid flow models integrate seamlessly with other types of physics simulation technologies.

Whereas, the CAE integration was carried out with the ANSYS Workbench environment and optiPLug. The defined design and random parameters were modified with the software optiSLang for binary-based CAE process integration, distribution of the parallel Workbench processes and for optimization and stochastic analysis.

3.2 Fluid-structure interaction model

The stochastic optimization method presented here is applied to a CAD and CAE parameter-based design optimization of a radial compressor shown in Fig. 9, including material, process and geometry tolerances. In the example presented the target of the optimization process is to maximize the efficiency of the turbine engine with respect to a limitation of the maximal v. Mises stress. Additional constraints are defined by resonance of any eigen frequency with the rotational velocity of the rotor. In total...
0.84
0.86
0.8889
0.893
0.905
0.91
0.905
0.91
0.91

Figure 10: Overview for the history of the efficiency $\eta$ of all iterations. DO=deterministic optimization, RE=Robustness evaluation, EA=Evolutionary algorithm, ARSM=Adaptive response surface method, SA=Sensitivity analysis

36 optimization parameters and 49 random influences are defined, as collected in Tab. 2.

3.3 Decoupled stochastic optimization loop

In the following the optimization loop presented in Fig. 8 will be shown. The initial step starts with the sensitivity analysis, based on the so called meta models of optimal prognosis. Therefore, the most important multivariate dependencies and design parameters can be identified (see Eq. (9)). With these parameters identified from the sensitivity analysis, the number of parameters related to the optimization problem can be reduced. In our case up to 10 design variables are left over with a relevant coefficient of optimal prognosis. Furthermore the meta models of optimal prognosis can be used as a surrogate model for a pre-optimization. The mean efficiency of the initial radial compressor was 86%. The best design of the latin hypercube sampling with an efficiency of 88.9% is used as start design of an evolutionary optimization based on the surrogate model and gives with one additional
design evaluation an efficiency of 89.3%. The first deterministic optimization in the sensitive design subspace results in the best deterministic design with an efficiency of 90.5%, as shown in Fig. 10.

After this, using a robustness evaluation, we can calculate the sigma level of the constraint violation with Eq. 10, for example regarding the stress. The distance of the design stress to the 5% quantile of the yield strength is a result of the first global safety factor of $\sigma^f = 4.5$. For the first iteration step. The target sigma level is $\sigma^l = 4.5$ to ensure a probability of failure $P(F) = 3.14 \cdot 10^{-6}$. In our case after the first iteration, the sigma level is larger than the desired $\sigma^l = 5.13 > 4.5$, therefore, a further iteration is necessary. In case of lack of prior knowledge, we use “rule of proportion” for the recalculation of the new safety factor (Fig. 14)

$$\sigma^H = \sigma^l \cdot \frac{\sigma^f}{\sigma^l} = 1.32$$
The second deterministic optimization step increases the efficiency to 90.8% and the second robustness evaluation with a new nominal design stress results in a new mean and standard deviation of the violation of the stress limit state as shown in Fig. 12. But with these moments the new sigma level of $\sigma_{II}^L = 3.6$ turns out to be less than the predicted sigma level. Now with the prior knowledge of the first iteration, we can interpolate the new global safety factor to $\gamma_{III} = 1.426$ with a new nominal design stress $\sigma_{III}^d = 1.19 \cdot 10^8$ for the third optimization.

After the third iteration, we obtain a new efficiency of 90.9% and the third robustness evaluation gives the following mean and standard deviation of the limit state violation shown in Fig. 13 with the new sigma level $\sigma_{III}^d = 4.1 < 4.5$ near the target value. In the iteration step III now the interpolation order is a quadratic one and the new safety factor is $\gamma_{IV} = 1.46$. With the resulting new nominal design stress $\sigma_{IV}^d = 1.71 \cdot 10^8$

After the fourth iteration, we obtain the very best efficiency of 91%. In relation
Figure 16: Anthill plot of the analyzed $N = 56$ design evaluations of the reliability analysis within iteration step IV between efficiency $\eta$ and yield stress $\sigma_y$.

Figure 17: Response surface plot of the reliability analysis design IV.

Figure 18: Flow angle of the initial design at the returnvane blades with separations along the blades.

Figure 19: Flow angle of the optimized design at the returnvane blades with a much more uniform flow.

to the initial design we have a better design performance of 5%. The final fourth robustness evaluation confirms the prediction of the sigma level $\sigma_{IV}^L = 4.48 \approx 4.5$ and shows a small deviation of the efficiency as shown in Fig. 15.

Of course, the probability levels of violation of the limit state conditions or of the initial efficiency are only a rough estimation and at least a reliability analysis
of the final design is recommended, especially for small probability levels. With the identification of the random sub domain directional sampling on adaptive moving least square is used for reliability analysis (see Sec. 2.3). The moving least square approximation is based on \( N = 56 \) design evaluations of an adaptive D-optimal design of experiment, as shown in Figs. 16 and 17. A cluster analysis is used to detect only one failure domain with high failure probability. The directional sampling procedure on the surrogate model detects samples in the unsafe domain. The assigned failure probability \( \bar{P}(\mathcal{F}) = 2.5 \cdot 10^{-6} \leq P^t(\mathcal{F}) = 3.4 \cdot 10^{-6} \) indicates an optimized six sigma design.

Finally, the Figs. 18 and 19 show the flow along the return vane blades. It is distinctly and visibly how the separations have been reduced in the optimized design and a more uniform flow is present.

4 CONCLUSIONS

Robust design optimization can provide multiple benefits. It permits the identification of those design parameters that are critical for the achievement of a certain performance characteristic. This can significantly reduce product costs. The effect of variations on the product behaviour and performance can be quantified. Moreover, robust design optimization can lead to a deeper understanding of the potential sources of variations. Consequently, more robust and affordable product designs can be achieved.

In this paper an efficient iterative decoupled loop approach is provided for reducing the necessary number of design evaluations. The applicability of this method for real case applications is demonstrated for a radial compressor. Using the approach presented, it is possible to improve the efficiency by about 5%. In addition we obtain an optimized design which is insensitive to uncertainties and considers the target failure probability.

5 ACKNOWLEDGEMENT

This project is kindly supported through the program “Internal Research Funding”, funded through the Niederrhein University of Applied Sciences. The authors would like to express their thanks to Johannes Einzinger of the ANSYS Germany GmbH for his assistance in formulating the FSI design problem and to Ulrike Adams and Daniela Ochsenfahrt of the DYNARDO GmbH for their support within the method implementation into the optiSLang software package and the collaborative work.
References


## A DESCRIPTION OF THE OPTIMIZATION PROBLEM

<table>
<thead>
<tr>
<th>$o$</th>
<th>Description</th>
<th>Symbol</th>
<th>$Y^o_0$</th>
<th>$Y^o_{IV}$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total temperature ratio $\Theta_T$</td>
<td>$Y_0$</td>
<td>1.1033</td>
<td>1.108</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Static entropy $S$</td>
<td></td>
<td>0.013168</td>
<td>0.0088</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Air mass flow $\dot{m}$</td>
<td>$Y_{IV}$</td>
<td>72.6</td>
<td>72.6</td>
<td>kg/s</td>
</tr>
<tr>
<td>4</td>
<td>Mechanical power $P$</td>
<td></td>
<td>-2.2664 · 10^6</td>
<td>-2.3918 · 10^6</td>
<td>W</td>
</tr>
<tr>
<td>5</td>
<td>Resulting torque $\tau$</td>
<td></td>
<td>-3238.8</td>
<td>-3418.05</td>
<td>J</td>
</tr>
<tr>
<td>6</td>
<td>Flow angle inlet $\alpha_{R1}$</td>
<td>$Y_{IV}$</td>
<td>2.2444</td>
<td>2.44</td>
<td>°</td>
</tr>
<tr>
<td>7</td>
<td>Flow angle between inlet and outlet $\alpha_{RS}$</td>
<td></td>
<td>62.562</td>
<td>67.63</td>
<td>°</td>
</tr>
<tr>
<td>8</td>
<td>Flow angle outlet $\alpha_{S1}$</td>
<td></td>
<td>22.009</td>
<td>17.13</td>
<td>°</td>
</tr>
<tr>
<td>9</td>
<td>Isentropic efficiency $\eta$</td>
<td></td>
<td>0.86</td>
<td>0.9094</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Pressure ratio $\Theta_P = P_{p,inlet}/P_{p,outlet}$</td>
<td></td>
<td>1.347</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Maximal displacement $u_{max}$</td>
<td></td>
<td>0.000111</td>
<td>0.000169</td>
<td>m</td>
</tr>
<tr>
<td>12</td>
<td>Von Mises stress $\sigma_e$</td>
<td></td>
<td>1.3107 · 10^8</td>
<td>1.653 · 10^8</td>
<td>Pa</td>
</tr>
<tr>
<td>13</td>
<td>Eigenfrequencies $f_i$ ($o = 14, ..., 45$)</td>
<td>$Y_{IV}$</td>
<td>1500-3200</td>
<td>1500-3200</td>
<td>Hz</td>
</tr>
<tr>
<td>46</td>
<td>Minimum safety factor $\gamma$</td>
<td></td>
<td>2.4453</td>
<td>1.51</td>
<td></td>
</tr>
</tbody>
</table>

*Table 1: Table of all output parameters*
Random variables $X_j$ and design variables $d_i$ of the initial and optimized Design; $\Phi =$normal- $\Lambda =$log-normal distribution

<table>
<thead>
<tr>
<th>$j$</th>
<th>$i$</th>
<th>Description</th>
<th>Symbol</th>
<th>$d_i^0$</th>
<th>$d_i^{IV}$</th>
<th>$F_X(x)$</th>
<th>$\sigma_X$</th>
<th>$E[X_j]$</th>
<th>$d_u$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>01</td>
<td>Inlet width</td>
<td>$l_{in}$</td>
<td>53</td>
<td>50.35</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>47.7</td>
<td>58.3</td>
<td>mm</td>
</tr>
<tr>
<td>02</td>
<td>02</td>
<td>Exit width</td>
<td>$l_{ex}$</td>
<td>26</td>
<td>26.27</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>23.4</td>
<td>28.6</td>
<td>mm</td>
</tr>
<tr>
<td>03</td>
<td>03</td>
<td>Radius of the impeller</td>
<td>$r_{imp}$</td>
<td>305</td>
<td>305.13</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>285</td>
<td>305</td>
<td>mm</td>
</tr>
<tr>
<td>04</td>
<td></td>
<td>Total temperature inlet</td>
<td>$T_{t,inlet}$</td>
<td>313</td>
<td>313</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>K</td>
</tr>
<tr>
<td>05</td>
<td></td>
<td>Specific heat capacity</td>
<td>$C_p$</td>
<td>1004.4</td>
<td>1004.4</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>J/kg/K</td>
</tr>
<tr>
<td>06</td>
<td></td>
<td>Specific gas constant</td>
<td>$R$</td>
<td>287.1</td>
<td>287.1</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>J/kg/K</td>
</tr>
<tr>
<td>07</td>
<td></td>
<td>Massflow of the air at the inlet</td>
<td>$\dot{m}$</td>
<td>72.6</td>
<td>72.6</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>kg/s</td>
</tr>
<tr>
<td>08</td>
<td></td>
<td>Rotation speed of the impeller</td>
<td>$\Omega$</td>
<td>699.76</td>
<td>699.76</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>radian/s</td>
</tr>
<tr>
<td>09</td>
<td></td>
<td>Total pressure inlet</td>
<td>$P_{p,inlet}$</td>
<td>1724000</td>
<td>1724000</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>Pa</td>
</tr>
<tr>
<td>10</td>
<td>04</td>
<td>Angle variation along hub of the rotor</td>
<td>$\beta_{RHB_1}$</td>
<td>-48</td>
<td>-51.15</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-52.8</td>
<td>-43.2</td>
<td>°</td>
</tr>
<tr>
<td>11</td>
<td>05</td>
<td>Angle variation along hub of the rotor</td>
<td>$\beta_{RHB_2}$</td>
<td>-25</td>
<td>-23.23</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-27.5</td>
<td>-22.5</td>
<td>°</td>
</tr>
<tr>
<td>12</td>
<td>06</td>
<td>Angle variation along hub of the rotor</td>
<td>$\beta_{RHB_3}$</td>
<td>-25</td>
<td>-24.51</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-27.5</td>
<td>-22.5</td>
<td>°</td>
</tr>
</tbody>
</table>

continued on next page ...
Random variables $X_j$ and design variables $d_i$ of the initial and optimized Design; $\Phi =$normal- $\Lambda =$log-normal distribution

<table>
<thead>
<tr>
<th>$j$</th>
<th>$i$</th>
<th>Description</th>
<th>Symbol</th>
<th>$d_i^0$</th>
<th>$d_i^{IV}$</th>
<th>$F_{X_i}(x)$</th>
<th>$\frac{\sigma_X}{E[X_i]}$</th>
<th>$d_1$</th>
<th>$d_u$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>07</td>
<td>Angle variation along shroud of the rotor</td>
<td>$\beta_{RSB_1}$</td>
<td>-55</td>
<td>-65.57</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-60.5</td>
<td>-49</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>14</td>
<td>08</td>
<td>Angle variation along shroud of the rotor</td>
<td>$\beta_{RSB_3}$</td>
<td>-30</td>
<td>-28.02</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-33</td>
<td>-27</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>15</td>
<td>09</td>
<td>Angle variation along shroud of the rotor</td>
<td>$\beta_{RSB_3}$</td>
<td>-45</td>
<td>-47.7</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-49.5</td>
<td>-40.5</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>Thickness of the rotor blades (shroud leading edge)</td>
<td>$t_{R_{SLE}}$</td>
<td>1</td>
<td>0.9</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>0.8</td>
<td>1.2</td>
<td>mm</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>(shroud trailing edge)</td>
<td>$t_{R_{STE}}$</td>
<td>6</td>
<td>5.95</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>5.0</td>
<td>7.0</td>
<td>mm</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>(hub trailing edge)</td>
<td>$t_{R_{HTE}}$</td>
<td>6</td>
<td>5.96</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>5.0</td>
<td>7.0</td>
<td>mm</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>(hub leading edge)</td>
<td>$t_{R_{HLE}}$</td>
<td>1</td>
<td>0.99</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>0.8</td>
<td>1.2</td>
<td>mm</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>Describes the curvature of the vanes (hub)</td>
<td>$\beta_{RVH_1}$</td>
<td>60</td>
<td>71.54</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>54</td>
<td>66</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>Describes the curvature of the vanes (shroud)</td>
<td>$\beta_{RVS_1}$</td>
<td>60</td>
<td>64.92</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>50</td>
<td>70</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>Relative thickness of the returnvane along the shroud</td>
<td>$\beta_{RVS_1}$</td>
<td>45</td>
<td>41.88</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>35</td>
<td>55</td>
<td>mm</td>
</tr>
</tbody>
</table>

continued on next page ...
Random variables $X_j$ and design variables $d_i$ of the initial and optimized Design; $\Phi =$normal- $\Lambda =$log-normal distribution

<table>
<thead>
<tr>
<th>$j$</th>
<th>$i$</th>
<th>Description</th>
<th>Symbol</th>
<th>$d_i^0$ =</th>
<th>$d_i^{IV}$ =</th>
<th>$F_X(x)$</th>
<th>$\frac{\sigma_X}{E[X]}$</th>
<th>$d_l$</th>
<th>$d_u$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>17</td>
<td>Relative thickness of the returnvane along the hub</td>
<td>$\beta_{RVH_1}$</td>
<td>45</td>
<td>43.42</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>45</td>
<td>66</td>
<td>mm</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>Density of the steel material</td>
<td>$\rho$</td>
<td>7850</td>
<td>7850</td>
<td>$\Phi$</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td>kg/m$^3$</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>Youngs modulus of steel</td>
<td>$E$</td>
<td>$2 \cdot 10^{11}$</td>
<td>$2 \cdot 10^{11}$</td>
<td>$\Lambda$</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>Poissons ratio</td>
<td>$\nu$</td>
<td>0.3</td>
<td>0.3</td>
<td>$\Lambda$</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>Coefficient of thermal expansion</td>
<td>$h_c$</td>
<td>$1.2 \cdot 10^{-5}$</td>
<td>$1.2 \cdot 10^{-5}$</td>
<td>$\Phi$</td>
<td>0.04</td>
<td>-</td>
<td>-</td>
<td>C$^{-1}$</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>Reference temperature</td>
<td>$T_{Ref}$</td>
<td>22</td>
<td>22</td>
<td>$\Phi$</td>
<td>0.06</td>
<td>-</td>
<td>-</td>
<td>C</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>Tensile yield strength</td>
<td>$\sigma_y$</td>
<td>$2.5 \cdot 10^{8}$</td>
<td>$2.5 \cdot 10^{8}$</td>
<td>$\Lambda$</td>
<td>0.064</td>
<td>-</td>
<td>-</td>
<td>Pa</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>Compressive yield strength</td>
<td>$\sigma_c$</td>
<td>$2.5 \cdot 10^{8}$</td>
<td>$2.5 \cdot 10^{8}$</td>
<td>$\Lambda$</td>
<td>0.064</td>
<td>-</td>
<td>-</td>
<td>Pa</td>
</tr>
<tr>
<td>31</td>
<td>18</td>
<td>Describes the curvature of the vanes (shroud)</td>
<td>$\beta_{RVS_2}$</td>
<td>0</td>
<td>-0.36</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-1</td>
<td>1</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>32</td>
<td>19</td>
<td>Describes the curvature of the vanes (shroud)</td>
<td>$\beta_{RVS_3}$</td>
<td>0</td>
<td>-0.56</td>
<td>$\Phi$</td>
<td>-0.1</td>
<td>-1</td>
<td>1</td>
<td>$^\circ$</td>
</tr>
<tr>
<td>33</td>
<td>20</td>
<td>Relative thickness of the returnvane along the shroud</td>
<td>$t_{RVS_2}$</td>
<td>10</td>
<td>10.88</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>8</td>
<td>12</td>
<td>mm</td>
</tr>
</tbody>
</table>

continued on next page ...
**Random variables** $X_j$ and **design variables** $d_i$ of the initial and optimized Design; $\Phi =$ normal- $\Lambda =$ log-normal distribution

<table>
<thead>
<tr>
<th>$j$</th>
<th>$i$</th>
<th>Description</th>
<th>Symbol</th>
<th>$d_i = d_i^{IV} = F_X(x) \frac{\mu_X}{E[X]}$</th>
<th>$d_l$, $d_u$,</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>21</td>
<td>Relative thickness of the returnvane along the shroud</td>
<td>$t_{RVs_3}$</td>
<td>6</td>
<td>6</td>
<td>$\Phi$ 0.1</td>
</tr>
<tr>
<td>35</td>
<td>22</td>
<td>Describes the curvature of the vanes (hub)</td>
<td>$\beta_{RVH_2}$</td>
<td>0</td>
<td>-0.34</td>
<td>$\Phi$ -0.1</td>
</tr>
<tr>
<td>36</td>
<td>23</td>
<td>Describes the curvature of the vanes (hub)</td>
<td>$\beta_{RVH_3}$</td>
<td>0</td>
<td>-0.14</td>
<td>$\Phi$ -0.1</td>
</tr>
<tr>
<td>37</td>
<td>24</td>
<td>Relative thickness of the returnvane along the hub</td>
<td>$t_{RVH_2}$</td>
<td>6</td>
<td>6.17</td>
<td>$\Phi$ 0.1</td>
</tr>
<tr>
<td>38</td>
<td>25</td>
<td>Relative thickness of the returnvane along the hub</td>
<td>$t_{RVH_3}$</td>
<td>10</td>
<td>12</td>
<td>$\Phi$ 0.1</td>
</tr>
<tr>
<td>39</td>
<td>26</td>
<td>Axial ratio of the major axis to minor axis of the elliptical rounding of the inflow edge (hub)</td>
<td>$r_{IEH}$</td>
<td>3</td>
<td>3.12</td>
<td>$\Phi$ 0.1</td>
</tr>
</tbody>
</table>

continued on next page ...
Random variables $X_j$ and design variables $d_i$ of the initial and optimized Design; $\Phi =$normal- $\Lambda =$log-normal distribution

| $j$ | $i$ | Description                                                                 | Symbol | $d_i^0$ | $d_i^{IV}$ | $\sigma$ | $E[X_j|d_i]$ | $E[X_j^{IV}|d_i]$ | Unit |
|-----|-----|------------------------------------------------------------------------------|--------|---------|------------|----------|---------------|---------------------|------|
| 40  | 27  | Axial ratio of the major axis to minor axis of the elliptical rounding of the inflow edge (shroud) | $r_{IEis}$ | 3       | 3.14       | $\Phi$   | 0.1        | 2.4                 | 3.6  |
| 41  | 28  | Edge of the vane at the leading edge along the hub contour                   | $r_{RV_{EHIn}}$ | 1       | 0.99       | $\Phi$   | 0.1        | 0.8                 | 1.2  |
| 42  | 29  | Edge of the vane at the leading edge along the shroud contour               | $r_{RV_{ESIn}}$ | 1       | 1.03       | $\Phi$   | 0.1        | 0.8                 | 1.2  |
| 43  | 30  | Edge of the vane at the trailing edge along the hub contour                 | $r_{RV_{EHOout}}$ | 1       | 1.03       | $\Phi$   | 0.1        | 0.8                | 1.2  |
| 44  | 31  | Edge of the vane at the trailing edge along the shroud contour            | $r_{RV_{ESOut}}$ | 1       | 1.18       | $\Phi$   | 0.1        | 0.8                | 1.2  |
| 45  | 32  | Hub to shroud offset impeller                                                | $\xi_{HTSO}$ | 0.5     | 0.53       | $\Phi$   | 0.1        | 0.4                | 0.6  | %
| 46  | 33  | Point tolerance impeller                                                     | $\xi_{IPt}$ | 0.1     | 0.08       | $\Phi$   | 0.1        | 0.08               | 0.12 | %

continued on next page ...
**Random variables** $X_j$ and **design variables** $d_i$ of the initial and optimized Design; $\Phi =$normal- $\Lambda =$log-normal distribution

<table>
<thead>
<tr>
<th>$j$</th>
<th>$i$</th>
<th>Description</th>
<th>Symbol</th>
<th>$d_i^0$</th>
<th>$d_i^V$</th>
<th>$\sigma_X E[X]$</th>
<th>$d_i$</th>
<th>$d_u$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>34</td>
<td>Hub to shroud offset returnvane</td>
<td>$\xi_{RHTSO}$</td>
<td>0.5</td>
<td>0.53</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>48</td>
<td>35</td>
<td>Point tolerance returnvane</td>
<td>$\xi_{RPT}$</td>
<td>0.1</td>
<td>0.08</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>49</td>
<td>36</td>
<td>Rounding of the trailing edge</td>
<td>$r_{TE}$</td>
<td>0.4</td>
<td>0.32</td>
<td>$\Phi$</td>
<td>0.1</td>
<td>0.32</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 2: Table of all parameters

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Formula</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constraint</td>
<td>$\gamma \geq 1.5$</td>
<td>$u_1(d) = \gamma - 1.5$</td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>Minimum of $\Theta_P \geq 1.34$</td>
<td>$u_2(d) = \Theta_P - 1.34$</td>
<td></td>
</tr>
<tr>
<td>Constraint</td>
<td>$\Omega/2\pi$ needs to be $\pm 10%$ away of $f_i$</td>
<td>$u_3(d) =</td>
<td>(f_i - (\Omega/2\pi))/(\Omega/2\pi)</td>
</tr>
<tr>
<td>Constraint</td>
<td>$\Omega/\pi$ needs to be $\pm 10%$ away of $f_i$</td>
<td>$u_4(d) =</td>
<td>(f_i - (\Omega/\pi))/(\Omega/\pi)</td>
</tr>
<tr>
<td>Limit state condition</td>
<td>Condition to calculate sigma level</td>
<td>$g_1(x) = \sigma_y - \sigma_e$</td>
<td>Pa</td>
</tr>
<tr>
<td>Objective</td>
<td>Maximal efficiency (isentropic)</td>
<td>$f(d) = \eta = \frac{(p_2^{\kappa} p_1^{\kappa - 1} - 1)}{(p_2^{\kappa} p_1^{\kappa - 1} - 1)}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Table of all constraints and objectives
A FULLY COUPLED THERMO-HYDRO-MECHANICAL MODEL FOR THE ANALYSIS OF THE LINING BEHAVIOR OF UNDERGROUND CAVERNS IN AN AA-CAES SYSTEM

ANIS EL MURR *, FAOUZI HADJ-HASSEN, AHMED ROUABHI AND MICHEL TIJANI
Department of Geosciences, MINES-ParisTech, 35 rue Saint-Honoré, 77305 Fontainebleau, France
anis.el_murr@mines-paristech.fr

Key words: thermo-hydro-mechanical coupling, underground storage, AA-CAES.

Abstract. With the worldwide demand for electricity, renewable energy is attracting increasing attention. As this energy has an intermittent character, large-scale storage technologies are necessary. One of the most promising systems is the advanced adiabatic compressed air energy storage (AA-CAES) in underground lined rock caverns. The high cyclic thermal and mechanical loadings involved in the system can disturb the surrounding geological barrier and thus lead to the failure of the system. The implementation of a special lining capable of limiting the thermal losses, reducing the air leakage and ensuring the caverns stability, is therefore required.

This paper presents the governing equations for fully coupled thermo-hydro-mechanical (THM) processes in saturated deformable media filled with dry air which characterize the conditions of the storing system. The assumptions used to simplify the equations are discussed and the neglected terms are underlined. These equations take into account the dependence of thermal conductivity on temperature, convection and heat compression. The air properties are derived using Helmholtz energy. A comprehensive comparison between the proposed model and a simple THM model based on constant parameters, ideal gas and conductive flux is made in order to emphasize the phenomena that could occur and their influences. Finally, thermo-hydro-mechanical simulations of the different lining materials are carried out to analyze the advantages and the drawbacks of each solution.

1 INTRODUCTION

The use of fossil fuel reserves to produce electricity presents many drawbacks; it is a finite resource and releases CO₂ which is the major cause of global warming [1,2]. Renewable energy is becoming more and more widespread, but it has two major problems: the first one is the big distance between energy resources and energy demand, which causes additional significant costs; the second one is the stochastic nature of this source [3]. Large-scale energy storage systems can solve these problems. Two systems are known to be the best: the pumped hydroelectric storage (PHS) and the compressed air energy storage (CAES) because of their high cost effectiveness. (PHS) system has a high cost and requires a suitable topography, which is rare; so, the (CAES) is the best promising technology to store energy [1].
The first storage station using an underground compressed air reservoir was a 290MW unit built in Huntorf, Germany, in 1978. It consists of two salt caverns with a total volume of 310,000m$^3$ [4,5] and an operating pressure between 4.8MPa and 6.8MPa. The second was a 110MW unit built in MacIntosh, USA. It consists of one salt cavern with a volume of 560,000m$^3$ and an operating pressure between 4.5MPa and 7.4MPa [1]. In the USA, a big interest is given nowadays to the construction of this type of plants; the largest is the 2,700 MW Iowa Stored Energy Park [6,7,8].

This system converts the off-peak electricity taken from the grid to a compressed air and stores it in an underground cavern. Then the air is used to drive a turbine to generate electricity when needed. After the compression stage, this system uses an inter-cooler and an after-cooler in order to reduce the air temperature, which leads to the decrease of the thermal stresses on the cavern walls and the reduction of the storage volume. Fuel is combusted then inside the turbine to increase the air temperature and thus its efficiency.

The advanced adiabatic compressed air energy storage (AA-CAES) is a modification of the (CAES) system; the innovative idea is that the heat generated during compression is stored in a thermal energy storage unit (TES) and recovered in the expansion phase (Fig. 1). AA-CAES is then more efficient and less damaging to the environment (no CO$_2$ emissions).

The high cyclic thermal and mechanical loadings applied to the TES walls (temperature up to 550°C and pressure up to 15MPa) would damage the surrounding rock mass and then lead to the failure of the system. This type of loading has never been treated in a gas storage application. To solve this problem, the implementation of a special lining, which role is to help the rock mass hold the gas pressure, to ensure a sufficient thermal isolation and to reduce the air leakage to an acceptable rate, is therefore necessary.

Two major studies have been held in Sweden to study natural gas storage in lining rock caverns. In Grängesberg and in Skallen [9,10,11], the caverns were constructed in fractured granite and lined by concrete and steel. As for CAES, a study was made on a former coal mine in Hokkaido Prefecture [3], Japan. The lining consists of 0.7 m concrete filled with a synthetic rubber seal. A recent study presented in [3,8] explored the potential of CAES in lined rock caverns at shallow depth (100 m), and has shown its feasibility from a leakage and energy-efficiency perspective. The lining consists of 1 m concrete.

In the experiences stated above, the steel and the synthetic rubber have a sealing function, the concrete has the pressure transmitting function and the rock mass has the pressure absorbing function.

The Preliminary simulations for the (TES) lined with steel and concrete showed a failure of the system. The stress criteria indicate that the materials will break because of the highly compressive thermal stresses generated by the high temperature. Thus, the implementation of a cooling system in the front wall of concrete layer is quite necessary to prevent the instability of the concrete and of the rock mass. Furthermore, an insulating material is added in front of the concrete layer to ensure thermal insulation.

The understanding of the interaction of different physics in the system is necessary to design the lining of the TES. Thus, a fully Thermo-Hydro-Mechanical model in saturated deformable media filled with dry air is presented. This model takes into consideration the dependency of the thermal conductivity of the materials and dry air properties on the temperature, convection and heat compression. A comparison with a simple THM model
(constant parameters, conductive flux) is made and a geomechanical analysis of the TES lining is carried out in order to optimize the design.

2. GENERAL BALANCE EQUATIONS

In the following derivations, we consider a deformable porous media fully saturated with air, and it is considered to be in local thermal equilibrium with the solid phase. A macroscopic approach is applied, meaning that the porous medium can be treated as a continuum where volume-averaged quantities replace local ones. The balance equations are derived by considering the conservation laws of two phases (solid and fluid) and by summing the contribution of both phases.

2.1 Mass balance equation

The Eulerian forms of mass balances of both phases are given by:

\[ \partial_{t} \rho_{f} \frac{\partial \varnothing}{\partial t} + \nabla \cdot (\varnothing \rho_{f} V_{f}) = 0 \]  
\[ \partial_{t} \rho_{s} \frac{\partial (1 - \varnothing)}{\partial t} + \nabla \cdot ((1 - \varnothing) \rho_{s} V_{s}) = 0 \]

Where \( \varnothing \) is the porosity of the porous media, \( \rho_{f} \) and \( \rho_{s} \) are the mass per unit of volume for the fluid and the solid respectively. \( V_{f} \) and \( V_{s} \) are the velocities of the fluid and the solid respectively with respect to a fixed reference. Developing eq.(1) and eq.(2) leads to the total mass balance equation:

\[ \varnothing \frac{\partial \rho}{\partial t} + \rho_{f} \frac{\partial \varnothing}{\partial t} \left[ \frac{\partial \varepsilon_{v}}{\partial t} + \frac{(1 - \varnothing) D \rho_{s}}{\rho_{s}} \right] = -\nabla \cdot q_{rf} \]

where \( \varepsilon_{v} \) is the skeleton volumetric strain, \( q_{rf} \) is the flux density vector with respect to the solid particles, and \( \frac{D}{D^{c}t} \) is the co-moving time derivative for tracking the motion of a solid particle along its trajectory.

2.2 Momentum balance

In the absence of inertia forces, the law of conservation of linear momentum is given by:

\[ \nabla \cdot \sigma + \rho_{m} g = 0 \]

where \( \sigma \) is the total stress tensor, \( g \) is a vector representing the acceleration of gravity and \( \rho_{m} \) is the average density of the mixture:

\[ \rho_{m} = (1 - \varnothing) \rho_{s} + \varnothing \rho_{f} \]

2.3 Energy balance

Neglecting the dissipation of energy due to the viscosity of the air and the irreversible energy dissipation in the skeleton, and any heat transfer caused by the solid displacement, the energy balance equation for both phases is then given by:
\( (1 - \varnothing) \rho_s C_{ps} \frac{\partial T_s}{\partial t} = -\nabla \cdot q_s \tag{6} \)
\( \varnothing \rho_f C_{pf} \left( \frac{\partial T_f}{\partial t} + \rho_f C_{pf} v_{rf} \cdot \nabla T \right) = \frac{\partial p}{\partial t} - \nabla \cdot q_f \tag{7} \)

where \( q_s \) and \( q_f \) are the conductive fluxes of the solid and the fluid respectively, \( C_{ps} \) and \( C_{pf} \) are the isobaric heat capacities of the solid and the fluid respectively and \( v_{rf} \) is the velocity of the fluid particles with respect to the solid.

### 3. GENERAL CONSTITUTIVE EQUATIONS

#### 3.1 Hydraulic

Neglecting any influence of thermal gradient on the air flux (thermo-osmosis) and any change in permeability due to air slippage (Klinkenberg effect), the mass flux could be written then in the following form:

\[ q_{rf} = -\rho_f \frac{k_h}{\mu_f} (\nabla p - \rho_f g \nabla z) \tag{8} \]

where \( k_h \) is the intrinsic permeability of the media, \( \mu_f \) is the dynamic viscosity of the fluid and \( z \) is the elevation.

The thermodynamic state of the air, considered as an homogenous fluid resulting from the mixture of different gases, is described by the couple \((p, T)\). The rate of change of air density is given by:

\[ \frac{\dot{p}}{\rho_f} = \beta_f \dot{p} - 3\alpha_f \dot{T} \tag{9} \]

where \( \beta_f \) is the compressibility modulus of the fluid and \( \alpha_f \) is the linear thermal expansion coefficient (both are function of the couple \((p, T)\)). For an ideal gas, these parameters are given by:

\[ \beta_f = \frac{1}{p} \tag{10} \]
\[ \alpha_f = \frac{1}{3T} \tag{11} \]

Since we are dealing with a wide range of temperatures (up to 550°C) and pressures (up to 15MPa), the assumption that the fluid is an ideal gas may not be sufficient. Thus, using a state law for dry air proposed by Lemmon [12] would be a better approach. This law uses the Helmholtz free energy given by:

\[ f(\rho, T) = \ln(\delta) + \sum_{i=1}^{5} N_i \tau^{i-4} + N_6 \tau^{1.5} + N_7 \ln(\tau) + N_8 \ln(1 - e^{-N_1 \tau}) + N_9 \ln(1 - e^{-N_2 \tau}) + N_{10} \ln \left( \frac{\tau^2}{3} + e^{N_{13} \tau} \right) + \sum_{i=1}^{10} N_i \delta i_i \tau^{i_k} \tau^{i_k} + \sum_{k=1}^{11} N_k \delta i_k \tau^{i_k + e^{-\delta i_k}} \tag{12} \]

where \( \delta = \frac{\rho}{\rho_0}, \tau = \frac{T_0}{T} \) and \((\rho_0, T_0)\) a state of reference. The coefficients \(N_i, N_k, i_k, j_k\) and \(l_k\) are constants. Knowing this thermodynamic potential, all the thermodynamic properties of dry air can be easily computed. A comparison between the two models described above was carried out. Fig. 2 (a, b) shows \( \beta_f \) and \( C_{pf} \) as a function of pressure and for a fixed value of temperature (200°C). For both parameters, we note a relative difference going up to 10%.
3 shows $\alpha_f$ as a function of temperature and for a fixed value of pressure (5MPa). Again, a relative difference going up to 14% is remarked. For accuracy purposes, Lemmo’s model is used. Fig. 4 (a, b, c) shows $C_{pf}$, $\alpha_f$ and $\rho_f$ as a function of temperature and for different pressures according to this model.

### 3.2 Thermal

Neglecting any heat flux caused by a pressure gradient, the heat conduction flux could be written as:

$$q = -\lambda_m \nabla T$$  \hspace{1cm} (13)

where $q$ is the sum of the thermal conductive heat flux of the fluid and the thermal conductive heat flux in the solid, $\lambda_m$ is the apparent macroscopic thermal conductivity over all phases; it is expressed as:

$$\lambda_m = (1 - \phi)\lambda_s + \phi\lambda_f$$  \hspace{1cm} (14)

Experimental results showed that the thermal conductivity for the insulation material and for the refractory concrete is a linear function of temperature. Details are showed in table 1.

### 3.3 Mechanical

Adopting a small strain approach, the total strain tensor $\varepsilon$ and the volumetric strain are given by

$$\varepsilon = \frac{1}{2} (\nabla u + (\nabla u)^T)$$ \hspace{1cm} (15)

$$\varepsilon_v = \nabla \cdot u$$ \hspace{1cm} (16)

where $u$ is the displacement vector and $t$ denotes the transpose of the tensor. The effective stress law for fully saturated media is given by [13]:

$$\sigma = \sigma' - Ip$$ \hspace{1cm} (17)

where $\sigma'$ is the macroscopic effective stress tensor given by:

$$\sigma' = D : (\varepsilon - \varepsilon_T - \varepsilon_p)$$ \hspace{1cm} (18)

where $D$ is the tangential stiffness matrix, and $\varepsilon$ is the total solid skeletal strain tensor, $\varepsilon_T$ is the thermal strain caused by temperature increase; it is given by:

$$\varepsilon_T = \alpha_d I \Delta T$$ \hspace{1cm} (19)

where $\alpha_d$ is the drained linear thermal expansion coefficient and $\varepsilon_p$ is the strain caused from compression of grains by internal hydrostatic fluid pressure, it is given by:

$$\varepsilon_p = -\frac{I \Delta p}{3k_g}$$ \hspace{1cm} (20)
where \( k_g \) is the bulk modulus of solid grains. The constitutive equation required to calculate the changes in solid grains, for small changes in grain density, is given by [14]:

\[
\frac{p_s}{p_{s0}} = 1 + \frac{\rho - \rho_0}{k_g} - \alpha_g (T - T_0) - \frac{\text{trace}(\sigma' - \sigma'_0)}{(1 - \phi) 3k_g} \tag{21}
\]

where \( \alpha_g \) is the linear thermal expansion coefficient of solid grains and the subscript 0 labels a reference state.

4. FINAL EQUATIONS

4.1 Hydraulic

Combining the aforementioned balance and constitutive equations and neglecting the terms \( \nu_f \nabla p, \nu_f \nabla T \) and \( \nu_f \nabla \rho_f \), we obtain:

\[
\frac{1}{M} \frac{\partial p}{\partial t} = \alpha_m \frac{\partial T}{\partial t} + b \frac{\partial \phi}{\partial t} = -\frac{1}{\rho_f} \nabla (-\rho_f \frac{k_h}{\mu_f} (\nabla p - \rho_f \rho g \nabla z)) \tag{22}
\]

where:

\[
\frac{1}{M} = \phi \beta_f + \frac{b - \phi}{k_g} \tag{23}
\]

is an isothermal storage coefficient and

\[
\alpha_m = 3\phi \alpha_f + (1 - \phi) \alpha_g - (1 - b) \alpha_d \tag{24}
\]

is the linear differential thermal expansion coefficient. Note that \( b \) is the well-known Biot coefficient.

4.2 Thermal

Using the assumptions of local thermal equilibrium, neglecting \( \nabla_s \nabla T, \nabla_s \nabla p \) and \( \nu_f \nabla p \) and combining eq. (6), (7) and (11), we obtain:

\[
(\rho C_p)_m \frac{\partial T}{\partial t} + \rho_f C_p \frac{\partial \nu_f \cdot \nabla T}{\partial t} - \frac{\partial p}{\partial t} = -\nabla (-\lambda_m \nabla T) \tag{25}
\]

where:

\[
(\rho C_p)_m = (1 - \phi) \rho_s C_{ps} + \phi \rho_f C_{pf} \tag{26}
\]

4.3 Mechanical

Combining eq. (4), (15), (16) and (17) and neglecting gravity forces, we obtain the well-known equilibrium equation:

\[
G \nabla^2 u + (G + \lambda) \nabla (\varepsilon_u) - b \nabla p - 3\alpha_d k_d \nabla T = 0 \tag{27}
\]

where \( G \) and \( \lambda \) are Lamé’s elasticity constants.
5. NUMERICAL MODEL

5.1 Model geometry

The cavity is a vertical cylinder with a 20 m diameter, a 50 m height, a domed roof and rounded invert. In order to simplify the simulations, the cavity is represented by a 1D axisymmetric model neglecting the effect of gravity and the cavity extremities. The simulations were carried out on two basic configurations shown in Fig. (2, 3).

5.2 Materials model

Since no information is available on the mechanical behavior of these materials under high temperature and high pressure, an elastic behavior was assumed. The thermo-hydro-mechanical properties adopted for these materials are given in detail in table 1.

Table 1: Thermo-hydro-mechanical properties.

<table>
<thead>
<tr>
<th>Property</th>
<th>Refractory concrete</th>
<th>Insulation material</th>
<th>Concrete</th>
<th>Granite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young’s modulus (GPa)</td>
<td>5</td>
<td>1</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>Poisson ratio</td>
<td>0.2</td>
<td>0.25</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Uniaxial compressive strength (MPa)</td>
<td>40</td>
<td>2</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>Thermal conductivity (W/m/K)</td>
<td>(0.0005, 3.4)</td>
<td>(0.0001, 0.36)</td>
<td>1.75</td>
<td>2.5</td>
</tr>
<tr>
<td>$\lambda = aT + b$; (a, b), where T is in °C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat capacity (J/kg/K)</td>
<td>950</td>
<td>1100</td>
<td>880</td>
<td>970</td>
</tr>
<tr>
<td>Permeability (m²)</td>
<td>$10^{-19}$</td>
<td>$10^{-11}$</td>
<td>$10^{-18}$</td>
<td>$10^{-15}$</td>
</tr>
<tr>
<td>Drained linear thermal expansion</td>
<td>0.78×$10^{-5}$</td>
<td>0.6×$10^{-5}$</td>
<td>1×$10^{-5}$</td>
<td>0.9×$10^{-5}$</td>
</tr>
<tr>
<td>coefficient (1/K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biot’s coefficient</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

5.3 Initial and boundary conditions

The initial stress regime in the granite rock is assumed to be isotropic and equal to 3.24 MPa. The excavation of the cavern is modeled using the convergence-confinement method. The mechanical load comprises 3 main phases: an increase of the pressure during 5 days from 0 to 5.2 MPa, then the pressure is maintained constant during 46 days before performing 9 cycles of 24 hours for each of them. As shown in Fig. 7a, the pressure is cycling between 4 and 8 MPa. In the same way, the thermal load also comprises three main phases: an increase of temperature during the first 5 days from 12.5 to 370°C, then the temperature is held constant during 46 days, followed by 9 cycles oscillating between 170 and 470°C of 24 hours for each of them (Fig. 7b). The limit of the model within the granite layer was defined sufficiently far to ensure that no radial displacement will occur. The numerical code used is COMSOL MULTIPhysics. Coupling conduction and convection with poroelasticity was carried out by the earth science module.
6. RESULTS

Preliminary simulations showed that the deformation caused by the mechanical load does not significantly affect the hydrostatic pressure field, thus the predominant coupling is thermo-hydraulic. This means that once the temperature and the pressure fields are computed, they are provided to the mechanical module in order to compute the stresses.

6.1 Thermo-hydraulic coupling

As mentioned in the introduction, comparisons were made between the model presented above and a simple THM model (constant parameters, no convection, no heat compression...). For this purpose, we have chosen 24 significant dates in one cycle and showed differences in temperature and in hydrostatic pressure between both models. Fig. 8a shows that a difference from -12 to 10°C in the temperature field could occur. This difference does not significantly affect the daily thermal losses which are equal to 0.56% for the first configuration and to 0.64% for the second one. As for the pressure field, the comparison for the first configuration did not show any significant difference (daily losses: 0.18%), but in the second one, the difference could reach 1.15MPa as shown in Fig. 8b. Furthermore, the concrete layer is sufficient to maintain daily air losses below 1%. No significant difference is noted between both models from an air leakage point of view.

6.2 Stability analysis

The stability of the cavity is analyzed using the Mohr-Coulomb’s failure criterion given by:

\[ f(\sigma_1', \sigma_2', \sigma_3') = k\sigma_1' - \sigma_3' \leq R_c \]  \hspace{1cm} (28)

where \( \sigma_1', \sigma_2', \) and \( \sigma_3' \) are the effective principal stresses, \( k \) is the passive pressure coefficient \( (k = 3 \) and corresponds to a friction angle of 30°) and \( R_c \) is the uniaxial compressive strength. The materials are considered to support no tensile stress.

\[ \sigma_1' \leq 0 \]  \hspace{1cm} (29)

The Mohr-Coulomb’s criterion shows that the proposed model is more critical than the classic one, and that there is no failure in shear for both configurations \( S = 1 \) for the first configuration and \( S = 2 \) for the second configuration, where \( S \) is the safety factor, Fig. 9 (a, b)). A tensile stress might occur in the refractory concrete due to the term \( \alpha_m \frac{\partial T}{\partial t} \) in the final hydraulic equation. The tensile stress is not only dependent on the linear thermal expansion coefficient of air, but also on the Biot’s coefficient, the more this coefficient is close to 1, the more the stresses are higher. This result requires a deep understanding of the poroelastic behavior of the refractory concrete and an accurate measurement of the Biot’s coefficient.

7. CONCLUSION

A fully coupled THM model has been developed to study the lining behavior in a TES cavern of an AA-CAES system. Comprehensive comparisons were made with a classic THM
model in order to point out the major differences. The daily heat losses and daily air losses are below 1% for both lining configurations, which is acceptable. It has also been shown that there is no failure by excess of shear; in order to increase the safety factor especially in the first configuration \((S = 1)\), the use of thermal joints (inserted between the blocks of the insulation material) that compensate the thermal expansion and thus reduce thermo-mechanical stresses is therefore necessary. Simulation also showed that a tensile stress could occur. A detailed study of the poroelastic properties of the refractory concrete is needed.

**ACKNOWLEDGEMENT**

This work was performed under the project SEARCH. The authors express their thanks to the French National Research Agency (ANR) for sponsoring this program as well as to the partners of this project GDF SUEZ, St Gobain and CEA/Liten.

**REFERENCES**


Figure 1: The concept of an AA-CAES system

Figure 2 (a, b): $\beta_f$ and $C_{pf}$ as a function of pressure and for a fixed value of temperature (200°C).
**Figure 3:** $\alpha_f$ as a function of temperature and for a fixed value of pressure (5MPa).

**Figure 4 (a, b, c):** $C_{p_f}$, $\alpha_f$ and $\rho_f$ as a function of temperature and for different pressures according to Lemmon’s model[12].

**Figure 5:** Configuration 1: insulation material + cooling system (in red) + concrete+ granite.
**Figure 6:** Configuration 2: refractory concrete + insulation material + cooling system + concrete + granite.

**Figure 7:** Mechanical (a) and thermal (b) loadings for one cycle (24 hours).

**Figure 8:** Comparison with a classic THM model, differences for temperature (a) and pressure (b) between both models for the second configuration.

**Figure 9 (a, b):** Mohr-Coulomb’s stress criterion for both configurations.
<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abascal, R.</td>
<td>906</td>
</tr>
<tr>
<td>Acevedo, H.</td>
<td>1033</td>
</tr>
<tr>
<td>Agelet de Saracibar, C.</td>
<td>159</td>
</tr>
<tr>
<td>Al-Qadad, A.</td>
<td>786</td>
</tr>
<tr>
<td>Albers, B.</td>
<td>58</td>
</tr>
<tr>
<td>Alcaraz, M.</td>
<td>1269</td>
</tr>
<tr>
<td>Alhama, F.</td>
<td>1407</td>
</tr>
<tr>
<td>Alhama, I.</td>
<td>1269, 1407</td>
</tr>
<tr>
<td>Alifanov, O.M.</td>
<td>840</td>
</tr>
<tr>
<td>Alonso, A.A.</td>
<td>1395</td>
</tr>
<tr>
<td>Améndola, F.</td>
<td>1418</td>
</tr>
<tr>
<td>Aminbaghai, M.</td>
<td>1383</td>
</tr>
<tr>
<td>Andrà, H.</td>
<td>1201</td>
</tr>
<tr>
<td>Antelo, L.T.</td>
<td>1395</td>
</tr>
<tr>
<td>António, J.</td>
<td>1322</td>
</tr>
<tr>
<td>Antunes, A.R.E.</td>
<td>562</td>
</tr>
<tr>
<td>Ardia, O.J.</td>
<td>491</td>
</tr>
<tr>
<td>Arias-Méndez, A.</td>
<td>1395</td>
</tr>
<tr>
<td>Ariza, M.P.</td>
<td>1213</td>
</tr>
<tr>
<td>Arranz Carreno, A.</td>
<td>550</td>
</tr>
<tr>
<td>Asao, S.</td>
<td>217</td>
</tr>
<tr>
<td>Asfour, S.</td>
<td>1142</td>
</tr>
<tr>
<td>Athanassoulis, G.A.</td>
<td>597</td>
</tr>
<tr>
<td>Bahai, H.</td>
<td>585</td>
</tr>
<tr>
<td>Balsa-Canto, E.</td>
<td>1395</td>
</tr>
<tr>
<td>Barrera, P.</td>
<td>1310</td>
</tr>
<tr>
<td>Barretta, R.</td>
<td>253, 287</td>
</tr>
<tr>
<td>Bastos, S.M.A.</td>
<td>562</td>
</tr>
<tr>
<td>Bazilevs, Y.</td>
<td>229</td>
</tr>
<tr>
<td>Behnke, R.</td>
<td>1082</td>
</tr>
<tr>
<td>Bekele, Y.W.</td>
<td>773</td>
</tr>
<tr>
<td>Belibassakis, K.A.</td>
<td>597</td>
</tr>
<tr>
<td>Benasciutti, D.</td>
<td>609</td>
</tr>
<tr>
<td>Bendani, K.</td>
<td>818</td>
</tr>
<tr>
<td>Benedikt, M.</td>
<td>989, 1009, 1021</td>
</tr>
<tr>
<td>Bergallo, M.B.</td>
<td>966</td>
</tr>
<tr>
<td>Bermúdez, A.</td>
<td>538</td>
</tr>
<tr>
<td>Biasi, V.</td>
<td>883</td>
</tr>
<tr>
<td>Bijl, H.</td>
<td>937</td>
</tr>
<tr>
<td>Bikard, J.</td>
<td>151</td>
</tr>
<tr>
<td>Billon, N.</td>
<td>151</td>
</tr>
<tr>
<td>Birken, P.</td>
<td>937</td>
</tr>
<tr>
<td>Blom, D.S.</td>
<td>937</td>
</tr>
<tr>
<td>Boccone, V.</td>
<td>1302</td>
</tr>
<tr>
<td>Boffi, D.</td>
<td>323</td>
</tr>
<tr>
<td>Bogdevičius, M.</td>
<td>661</td>
</tr>
<tr>
<td>Bols, J.</td>
<td>1000</td>
</tr>
<tr>
<td>Boman, R.</td>
<td>159</td>
</tr>
<tr>
<td>Bonet, J.</td>
<td>503, 550</td>
</tr>
<tr>
<td>Boria, S.</td>
<td>1051</td>
</tr>
<tr>
<td>Brandt, L.</td>
<td>685</td>
</tr>
<tr>
<td>Brant, S.R.C.</td>
<td>1257</td>
</tr>
<tr>
<td>Brulliard, V.</td>
<td>1130</td>
</tr>
<tr>
<td>Buchschmid, M.</td>
<td>720</td>
</tr>
<tr>
<td>Bungartz, H-J.</td>
<td>241</td>
</tr>
<tr>
<td>Burczynski, T.</td>
<td>387</td>
</tr>
<tr>
<td>Bussetta, P.</td>
<td>159</td>
</tr>
<tr>
<td>Bustamante, C.A.</td>
<td>1280, 1439</td>
</tr>
<tr>
<td>Calvo, N.A.</td>
<td>966</td>
</tr>
<tr>
<td>Camiola, V.D.</td>
<td>515</td>
</tr>
<tr>
<td>Carrazedo, R.</td>
<td>631</td>
</tr>
<tr>
<td>Cavallini, N.</td>
<td>323</td>
</tr>
<tr>
<td>Cerracchio, P.</td>
<td>311</td>
</tr>
<tr>
<td>Cerrato, A.</td>
<td>906</td>
</tr>
<tr>
<td>Cerutti, F.</td>
<td>1302</td>
</tr>
<tr>
<td>Cervera, M.</td>
<td>159</td>
</tr>
<tr>
<td>Chantrail, T.</td>
<td>1373</td>
</tr>
<tr>
<td>Chekired, M.</td>
<td>1292</td>
</tr>
<tr>
<td>Chemin, S.</td>
<td>918</td>
</tr>
<tr>
<td>Chetverushkin, B.</td>
<td>978</td>
</tr>
<tr>
<td>Chiumenti, M.</td>
<td>159</td>
</tr>
<tr>
<td>Cheuffart, Q.</td>
<td>171</td>
</tr>
<tr>
<td>Cliffe, A.</td>
<td>104</td>
</tr>
<tr>
<td>Cocks, A.C.F.</td>
<td>275</td>
</tr>
<tr>
<td>Coda, H.B.</td>
<td>631</td>
</tr>
<tr>
<td>Cordão Neto, M.P.</td>
<td>1257</td>
</tr>
<tr>
<td>Coroneo, M.</td>
<td>37</td>
</tr>
<tr>
<td>Cremanns, K.</td>
<td>1475</td>
</tr>
<tr>
<td>Cremonesi, M.</td>
<td>756</td>
</tr>
<tr>
<td>Daru, V.</td>
<td>574</td>
</tr>
<tr>
<td>Datcheva, M.</td>
<td>67</td>
</tr>
<tr>
<td>de Araujo, R.</td>
<td>479</td>
</tr>
<tr>
<td>De Ridder, J.</td>
<td>668</td>
</tr>
<tr>
<td>De Schutter, G.</td>
<td>806</td>
</tr>
<tr>
<td>Degroote, J.</td>
<td>668, 1000</td>
</tr>
<tr>
<td>Delfs, J-O.</td>
<td>949</td>
</tr>
<tr>
<td>Deng, X.</td>
<td>229</td>
</tr>
<tr>
<td>di Prisco, C.</td>
<td>756</td>
</tr>
<tr>
<td>Di Sciucua, M.</td>
<td>311, 355</td>
</tr>
<tr>
<td>Di Stefano, V.</td>
<td>1109</td>
</tr>
<tr>
<td>Dialami, N.</td>
<td>159</td>
</tr>
<tr>
<td>Divo, E.</td>
<td>127, 491</td>
</tr>
<tr>
<td>Dlugosz, A.</td>
<td>387</td>
</tr>
<tr>
<td>Dominguez-Mota, F.J.</td>
<td>1310</td>
</tr>
<tr>
<td>Drakšas, M.</td>
<td>661</td>
</tr>
<tr>
<td>Duan, B.Y.</td>
<td>527</td>
</tr>
</tbody>
</table>
Duan, W.................................................. 1045
Dubini, G.................................................. 685
Dubujet, Ph.................................................. 1168
Duchaine, F.................................................. 1154
Dulkiravich, G.S.................................................. 367
Dunne, F.P.E.................................................. 275
Durany, J.................................................. 1062
D’Ascenzo, N.................................................. 978
Eglite, I.................................................. 1249
El Murr, A.................................................. 1502
Erinosho, T.O.................................................. 275
Ern, A.................................................. 574
Errera, M-P.................................................. 918
Fahmy, M.A.................................................. 862
Farias Filho, U.P.................................................. 562
Faveargeon, J.................................................. 851
Favre, L.................................................. 1418
Ferraioli, F.................................................. 379
Ferreira, P.................................................. 1322
Feyel, F.................................................. 883
Fichera, C.................................................. 708
Filopoulos, S.................................................. 597
Fischlischweiger, M.................................................. 1343
Fleischhauer, R.................................................. 1082
Flitz, F.................................................. 957
Florez, W.F.................................................. 1280, 1439
Flumian, F.................................................. 609
Foca, M.................................................. 1354
Fontoura, S.A.B.................................................. 794
Formisano, A.................................................. 379
Franssen, J-M.................................................. 895
Gardini, F.................................................. 323
Garzon, M.................................................. 193
Gastaldi, L.................................................. 323
Gatzhammer, B.................................................. 241
Genes, M.C.................................................. 205
Gherlone, M.................................................. 311, 355
Gicquel, L.................................................. 1154
Gil, A.J.................................................. 503, 550
Gimenez, J.M.................................................. 966
Gleim, T.................................................. 423
Gomes, I.F.................................................. 1257
Gómez, D.................................................. 538
González, G.................................................. 1310
González, J.A.................................................. 25, 906
Granados, J.M.................................................. 1439
Gravouil, A.................................................. 1373
Gray, L.J.................................................. 193
Guarracino, F.................................................. 264
Guimañáes, T.A.................................................. 479
Guo, Z.................................................. 696
Görke, U-J.................................................. 80
Hadj-Hassen, F.................................................. 1502
Hage Chehade, F.................................................. 786
Halama, R.................................................. 1430
Hanoglu, U.................................................. 116
Harapin, A.................................................. 732
Haupt, M.C.................................................. 467, 651
Hayakawa, H.................................................. 1191
Hayhurst, D.R.................................................. 432
Hernández, A.................................................. 458
Hijjaj, M.................................................. 828
Hofer, A.................................................. 1009, 1021
Horst, P.................................................. 467, 651
Hrabovský, J.................................................. 1179, 1383
Hsu, M-C.................................................. 229
Hudobivnik, B.................................................. 1331
Idelsohn, S.R.................................................. 966
Inclan, E.J.................................................. 367
Inoue, N.................................................. 794
Iurlaro, L.................................................. 355
Jędrusik, S.................................................. 1237
Janutêniënë, J.................................................. 661
Jaques, D.................................................. 806
Jasper, T.................................................. 1475
Jaspers, S.J.................................................. 345
Jaurè, S.................................................. 1154
Kabel, M.................................................. 1201
Kacem, M.................................................. 1168
Kalbacher, T.................................................. 949
Kaliske, M.................................................. 1082
Kassab, A.J.................................................. 127, 491
Kawai, M.................................................. 1191
Kebrâiaei, R.................................................. 139
Kikuchi, K.................................................. 643
Kim, H.J.................................................. 1365
Kim, H.Y.................................................. 1365
Kim, T.H.................................................. 1365
Kiselev, A.................................................. 1101
Kolditz, O.................................................. 80, 873, 949
Kolyshkin, A.A.................................................. 1249
Korelec, J.................................................. 1331
Korobenko, A.................................................. 229
Kosenko, I.I.................................................. 92
Kowollik, D.S.C.................................................. 467
Krabbenhoft, K.................................................. 744
Krumphals, A.................................................. 1237
Kubota, Y.................................................. 643
Kuhl, D.................................................. 423
Kunieda, M.................................................. 335
Kuramae, H.................................................. 1226
Kutíš, V.................................................. 1179, 1383
Laredj, N.................................................. 818
Lautenschläger, C.E.R.................................................. 794
Le Corre, S.................................................. 1130
Lecomte-Beckers, J.................................................. 183