Computational Methods in Applied Sciences 41

Adnan Ibrahimbegovic *Editor*

Computational Methods for Solids and Fluids

Multiscale Analysis, Probability Aspects and Model Reduction

ECCOMAS
European Community
on Computational Methods
in Applied Sciences

Springer

Computational Methods in Applied Sciences

Volume 41

Series editor

E. Oñate
CIMNE
Edificio C-1, Campus Norte UPC
Gran Capitán, s/n
08034 Barcelona, Spain
onate@cimne.upc.edu

More information about this series at http://www.springer.com/series/6899

Adnan Ibrahimbegovic
Editor

Computational Methods for Solids and Fluids

Multiscale Analysis, Probability Aspects and Model Reduction

Springer

Editor
Adnan Ibrahimbegovic
Laboratoire Mécanique Roberval
Université Technologie Compiègne
Compiègne
France

ISSN 1871-3033
Computational Methods in Applied Sciences
ISBN 978-3-319-27994-7 ISBN 978-3-319-27996-1 (eBook)
DOI 10.1007/978-3-319-27996-1

Library of Congress Control Number: 2015959590

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Preface

This book contains a selection of research works first presented at the 2nd International Conference on Multiscale Computational Methods for Solids and Fluids, which was held in Sarajevo, June 10–12, 2015. Among them are the contributions from several experts from France and Germany (A. Ibrahimbegovic, A. Ouahsine, H. Matthies, N. Limnios, and P. Villon) who jointly taught the conference short course entitled "Current Research on Solids & Fluids: Computations, FE Code Coupling, Model Reduction, Probability," on dates preceding the conference, June 8 and 9, 2015. With the complementary contributions from the short-course instructors and a number of other conference participants, we seek to provide a comprehensive review of the state of the art in this currently active research domain. The presented contributions pertain to more than one of these aspects: (i) multiscale computations for solids and fluids, (ii) probability aspects, and (iii) model reduction. Given that each contribution touched upon more than one of these aspects, it is deemed the most appropriate to present them in a random-like order of the first authors' names.

The main ideas in each chapter are briefly outlined and commented upon for the benefits of readers.

Beckers Benoit has presented his work on "Multiscale Analysis as a Central Component of Urban Physics Modeling." The original aspects of presented point of view on urban physics are in placing it in between the scales of environmental physics and building physics. Being able to simulate at this scale is principally needed for energy exchange. Namely, the urban environment with its rapid worldwide growth has become a complex dynamic interface for either top-down problem with generation of an urban microclimate or bottom-up one with influence of an urban morphology on the final consumption of the entire city, jointly used to define properly the role of cities in global warming. The main benefits pertain to energy balance, urban planning, and participation to global climate models.

Brank Boštjan et al. have presented their work on "A Path-Following Method Based on Plastic Dissipation Control." The proposed method is a multiscale solution strategy, the most suitable for dealing with softening phenomena, which allows for path-following response computations. Here, the fine-scale representation is only controlled by the total inelastic dissipation passing to coarse scale. Detailed validation is presented for elasto-plastic model. An important model ingredient concerns the discrete approximation, which relies upon the embedded discontinuity finite element method that uses rigid-plastic cohesive laws with softening to model material failure process. Illustrative developments are presented for 2D solid and frame finite elements with embedded discontinuities that describe cohesive stresses and stress resultants by rigid plasticity with softening.

Cai Shang-Gui et al. have presented their work on "Improved Implicit Immersed Boundary Method via Operator Splitting." The main challenge tackled herein is in providing an efficient approach to fluid-flow simulation over moving solid with complex shape. The efficiency of the implicit immersed boundary method is achieved via symmetry of tangent arrays and operator splitting technique. An additional moving force equation imposes the interface velocity condition exactly on the immersed surface and the rotational incremental projection method ensures that the numerical boundary layers are generated towards the velocity and pressure during the calculation. Cai Shang-Gui et al. have also presented their work on "Modelling Wave Energy Conversion of a Semi-submerged Heaving Cylinder." The proposed model, based upon full 3D viscous Navier–Stokes equations, aims at simulating the ocean wave energy conversion of a semi-submerged heaving cylinder.

Dumont Serge et al. have presented their work on "Multiscale Modeling of Imperfect Interfaces and Applications." A systematic development of multiscale approach for an interface model can be developed by using asymptotic techniques where the thickness of the interface is considered as a small parameter. At the microscale, the model can account for imperfect interface models by taking into account microcracks. The homogenization techniques of matched asymptotic for media with microcracks is used then both in 3D and 2D cases, which leads to a cracked orthotropic material. It is shown that the Kachanov-type theory leads to soft interface models, while the approach by Goidescu leads to stiff interface models. A fully nonlinear variant of the model is also proposed, derived from the St. Venant–Kirchhoff constitutive equations. Among broad set of applications, the masonry structures are chosen to illustrate the model performance.

Jehel Pierre has presented his work on "A Stochastic Multi-scale Approach for Numerical Modeling of Complex Materials—Application to Uniaxial Cyclic Response of Concrete." Accounting for inelastic behavior of heterogeneous materials, such as concrete, calls for development of multi-scale techniques that looks for sources on nonlinearity at the relevant scale, combined with stochastic methods accounting for uncertainties. The chosen model represents the mechanical response of a representative volume of concrete in uniaxial cyclic loading. The heterogeneities are represented at mesoscale and inelastic nonlinear local response is modeled in the framework of thermodynamics with internal variables. Spatial variability of the local response is represented by correlated random vector fields generated with the spectral representation method. Macroscale response is recovered through standard homogenization procedure from representative volume element and shows salient features of the uniaxial cyclic response of concrete that are not explicitly modeled at mesoscale.

Kozar Ivica has presented his work on "Relating Structure and Model," seeking to gain an additional insight into large structure modeling. The presented approach deviates from usual model building procedures that leads us to transfer of parameters between the model and the real structure. Here, the problem is addressed in a more general manner, where both modeling and discretization are formulated, subsequently seeking relationship between relevant parameters of the real structure and its model. The corresponding procedure provides how to determine the scaling matrices in parameter and measurement spaces.

Lebon Jérémy et al. have presented their work on "Fat Latin Hypercube Sampling and efficient Sparse Polynomial Chaos Expansion for Uncertainty Propagation on Finite Precision Models: Application to 2D Deep Drawing Process." The main motivation stems from uncertainty propagation in model parameters, with the variation range of random variables that may be many orders of magnitude smaller than their nominal values. This is typical of the nonlinear finite element method computations involving contact/friction and material nonlinearity. A particular attention was given to the definition of adapted design of experiment, taking into account the model sensitivity with respect to infinitesimal numerical perturbations. The samples are chosen using an adaptation of the Latin hypercube sampling, requiring them to be sufficiently spaced away to filter the discretization and the other numerical errors limiting the number of possible numerical experiments, which leave the challenge to building an acceptable polynomial chaos expansion with such sparse data.

Marenic Eduard and Adnan Ibrahimbegovic have presented their work on "Multiscale Atomistic-to-Continuum Reduced Models for Micromechanical Systems." The main focus is upon the development of multiscale-reduced models and computational strategy for micromechanical systems, with currently interesting applications to graphene. The fine scale concerns the atomistic model and is formulated and solved along with the corresponding coarse-scale model obtained by homogenization. Two mainstream multiscale methods, the quasi-continuum and bridging domain, are compared and brought to bear upon the optimal model reduction strategy. Consequently, these two methods are further advanced from their standard formulation to a unified coupling and implementation strategy. The method can also deal with a defect-damaged graphene sheet granting an excellent performance of the proposed multiscale solution strategy.

Matthies Hermann et al. have presented their work on "Inverse Problems in a Bayesian Setting." The work reveals the strong connection between the inverse problems of the parameter identification and the forward computations of uncertainty quantification with parameter uncertainty propagating through response computations. The connection of this kind is naturally placed in the Bayesian setting, where the Bayesian updates, or filters, are derived from the variational problem associated with conditional expectation. Among various constructions of filters, the most efficient seem to be the linear or nonlinear Bayesian updates based on functional or spectral approximation constructed with polynomials, which grant much higher computational efficiency in forward uncertainty quantification than the time-consuming and slowly convergent Monte Carlo sampling.

Niekamp Rainer et al. have presented their work on "Heterogeneous Materials Models, Coupled Mechanics-Probability Problems and Energetically Optimal Model Reduction." It is shown that the sound theoretical formulation of a multiscale model of damage behavior of heterogeneous materials can be cast as coupled mechanics-probability problem. In particular, such the fine-scale interpretation of damage mechanisms can provide the most meaningful probability density distribution of material parameters governing the failure phenomena, which can be described in terms of random fields. The second challenge tackled here pertains to providing an efficient solution procedure to this coupled mechanics–probability problem, formulated by the spectral stochastic finite element method. Here, the curse of dimension, with the coupled mechanics–probability problem dimension growing with a number of random fields, is handled through low-rank approach and solution space reductions. In particular, a rank-one update scheme is devised as the optimal low-rank representation with respect to the minimal energy at the given rank.

Nikolic Mijo et al. have presented the work on "Modelling of Internal Fluid Flow in Cracks with Embedded Strong Discontinuities." The proposed multiscale model can handle fluid–structure interaction problem typical of localized failure of heterogeneous rock material under internal fluid flow. Of special interest are the methods where the fine-scale mechanics failure phenomena are presented only at the coarse-scale parameter in terms of fracture energy needed to achieve the full crack creation. The crack propagation induced steep displacement gradients are accounted for by using the concept of embedded discontinuity FEM. The computational efficiency is granted by the rock mass representation by Voronoi cells, kept together by cohesive links. The latter is chosen in terms of the Timoshenko beams capable of providing the crack-induced discontinuity propagation between the rock grains both in mode I and mode II. The model can account for rock material heterogeneities with pre-existing cracks represented by weak links placed in agreement with given probability distribution.

Papamichail Chrysanthi et al. have presented their work on "Reliability Calculus on Crack Propagation Problem with a Markov Renewal Process." The fatigue crack propagation is defined in terms of a stochastic differential system that describes the evolution of a degradation mechanism. A Markov or a semi-Markov process was considered as the perturbing process of the system that models the crack evolution. With the help of Markov renewal theory, the reliability of a structure is defined in terms of analytical solution. The method reduces the complexity of the reliability calculus compared with the previous resolution method, yet delivers good agreement with experimental data set, and Monte Carlo estimations.

Prieto Juan Luis has presented his work on "Multi-scale Simulation of Newtonian and Non-Newtonian Multi-phase Flows." The special attention is given to a level-set method to capture the fluid interface along with Brownian dynamics simulations to account for the viscoelastic effects of the fluid. The solution is obtained by using the second-order semi-Lagrangian scheme and evolving the level-set function along the characteristic curves of the flow. The proposed approach can also handle the free-surface flow taking into account viscous and surface tension effects, by using a semi-Lagrangian particle level-set method and by adding the marker particles to correct the shape of the free surface. The multiscale approach is used to solve stochastic, partial differential equations by using the finitely extensible nonlinear elastic kinetic model and a variance-reduced technique on a number of ensembles of dumbbells scattered over the domain.

Ravi Srivathsan and Andreas Zilian have presented their work on "Numerical Modeling of Flow-Driven Piezoelectric Energy Harvesting Devices." The devices of this kind provide a smart replacement of batteries with low power energy harvesting of flow-induced vibrations. The theoretical formulation leads to a coupled problem involving fluid, structure, piezo-ceramics, and electric circuit. The main difficulty pertains to problem nonlinearities and the need for reliable, robust, and efficient computations, which is here achieved by a monolithic approach involving surface-coupled fluid-structure interaction, volume-coupled piezoelectric–mechanics and a control of energy harvesting circuit. A space-time finite element approximation is used for the numerical solution of the governing equations, which allows for different types of structural elements (plate, shells) with varying cross sections and material constitutions and different types of harvesting circuits.

Rosic Bojana et al. have presented their work on "Comparison of Numerical Approaches to Bayesian Updating." The main challenge concerns Bayesian process of identifying unknown probability distribution of model parameters given prior information and a set of noisy measurement data. Two approaches are possible: one that uses the classical formula for measures and probability densities, and the other that leaves the underlying measure unchanged and updates the relevant random variable. The former is numerically tackled by a Markov chain Monte Carlo procedure based on the Metropolis–Hastings algorithm, whereas the latter is implemented via the ensemble/square root ensemble Kalman filters, as well as the functional approximation approaches in the form of the polynomial chaos based linear Bayesian filter and its corresponding square root algorithm. It was shown some of the principal differences between full and linear Bayesian updates when a direct or a transformed version of measurements are taken into consideration.

Ylinen Antti et al. have presented their work on "Two Models for Hydraulic Cylinders in Flexible Multibody Simulations." In modeling hydraulic cylinders, interaction between the structural response and the hydraulic system needs to be taken into account. In this work, two approaches for modeling flexible multibody systems are presented and compared: one with truss-element-like cylinder and bending flexible cylinder models, and other with bending flexible cylinder element chosen as a super element combining the geometrically exact Reissner beam element, the C1-continuous slide-spring element needed for the telescopic movement, and the hydraulic fluid field. Both models are embedded with a friction model based on a bristle approach and can be implemented within the standard finite element environment. In time the coupled stiff differential equation system is integrated using the L-stable Rosenbrock method.

The goal of gathering these contributions in a single book from Springer ECCOMAS series is to ensure more lasting value to the results first presented at the 2nd ECCOMAS Thematic Conference on Multiscale Computations on Solids and Fluids, providing the best starting point for further exploration in this currently very active research field. I would like to thank all the authors for contributing to this goal.

Adnan Ibrahimbegovic

Contents

Multiscale Analysis as a Central Component of Urban Physics Modeling

Benoit Beckers

Abstract Urban physics is seeking its place between two better structured scales, that of environmental physics and that of building physics. The intermediate level of the city and the urban district is particularly difficult to appreciate, because it involves huge geometries that must however be precisely detailed and finely meshed. It has become very important to simulate on this scale, principally energy exchanges, because the urban environment, with its rapid worldwide growth, has become a complex dynamic interface for both top-down problems (generation of urban microclimates) and bottom-up ones (influence of urban morphology on the final consumption of the entire city), and even properly multiscale applications (involvement of cities in global warming). This paper provides a starting point (the shortwave radiative exchange), a process of cross-validation between measurements and simulations (with emphasis on the contribution of satellite imagery) and three main objectives: energy balance, urban planning and participation to global climate models.

Keywords Urban physics · Solar energy · Multi-scale · Optimization

1 Introduction

If we observe from a physical and global point of view the contemporary evolution of human society on the surface of the earth, there is essentially a very rapid urbanization, accompanied by an equally rapid urban sprawl. Between 2000 and 2030, urban occupation of the land surface of the planet should have tripled [70], which already causes serious difficulties for food self-sufficiency, exposure to natural hazards (particularly floods) and health (air pollution, Urban Heat Island) [29].

Cities already account for more than half the world's population, and soon the two-thirds; they also concentrate many industrial activities and transport. They are

B. Beckers (✉)
Urban Systems Engineering Department, Compiegne University of Technology – Sorbonne University, Compiegne, France
e-mail: benoit.beckers@utc.fr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_1

therefore necessarily of great importance in the current climate change, especially by their CO_2 production (up to 70 % of the greenhouse gases total emission [74]).

Climatologists meeting within the IPCC (Intergovernmental Panel on Climate Change) operate their digital models on the planet scale with a meshing of approximately 100 km per side. Even in the long term, it is difficult to imagine that the size of the mesh can drop below 20 km [41]. Therefore, urban structures do not appear. However, if we calculate the total heat production of urban areas on the East Coast of the United States, on the one hand, and on the East Coast of China on the other one, and if this production is injected into a climate model, especially in the jet stream in winter, we see that this contribution may explain the over-warming of permafrost, several thousands of miles away, that is, respectively, in northern Canada and Siberia [84].

In the present climate models, cities are like phantoms: we do not see them, but we can measure their effects, sometimes very far from them. In the last IPCC report, it appears that the many measures taken by thousands of cities (in the form of climate-energy policies) cannot be quantified [41]. To go further, it will be necessary to implement a multi-scale method where calculations as fine as necessary at the urban level make it possible to correctly set the climate models.

This chapter is divided into four parts. Section 2, in the form of a state of the art very concise but expanded to the main involved areas, will remind the knowledge and existing difficulties. Section 3 will indicate how, and in what order, a new framework can be built for the urban physics. Sections 4 and 5 will describe, respectively, the shortwave and the long wave models principal characteristics.

2 Urban Physics: A State of the Art

2.1 From Environmental Physics

By the late sixties, researchers began to apply the possibilities of numerical simulations to the study of energy, momentum and matter exchanges between soil and atmosphere, and by that time, T.R. Oke applied these energy balances to specific environments: the cities [59]. He identified and sought to prioritize the physical phenomena contributing to the production of the Urban Heat Island (mainly, increased net radiation and absorption of heat by urban surfaces [66]).

We can therefore speak about urban physics, as a particular field of environmental physics [7, 52].

However, at that time, the resources did not allow simulating complex geometric patterns, and measures were limited in quality, time and space. Despite substantial advances in satellite imagery, ground-based measurements are still required and are still limited to periods of a few months in some parts of the city (cost, difficulty).

Regarding the geometry, researches are oriented in two directions: the urban canyon study (simplified street), or use of regular and extruded shapes (grid plan).

At this point, despite some very encouraging intermediate results, it is not possible to accurately quantify the exchanges at the urban scale, nor to propose substantial guidelines for urban planning.

2.2 From Urban Planning

Urban planning was developed between the 1850s and 1950s, by G.E. Haussmann in Paris (bring to all people the air and the sun) and I. Cerdà in Barcelona (the inventor of the word "urbanism") to the figure of Le Corbusier (Macià plan in Barcelona, Chandigarh, Brasilia ...), first for hygienist concerns, then for urban comfort ones.

Very large buildings made possible by the invention of the elevator have created extremely dense downtowns, introducing the economic problem of the depreciation of land always in the shade (Manhattan and Chicago, then major Asian cities). Sunshine, solar right and sky rights were the main tools for planners to think in 3D [7].

We can then really speak of an intermediate scale, the neighborhood and the city, which we call "meso", between the "macro" scale of geographers and climatologists and the "micro" scale of architects.

In Barcelona, I. Cerdà turns the grid pattern of the Eixample 45° with respect to the north-south axis, to better distribute sunlight on the facades. However, in the Mediterranean climate, the delicate point is the cross ventilation, which can avoid the summer air conditioning. This problem can be treated qualitatively (introduction of light wells), but, as it should be studied at the level of each building, it is impossible to quantify globally and therefore to validate the insight of Cerdà. In addition, urban life has changed since then, and has become much noisier. If people have to close their windows at night for acoustic reasons, the picture is completely changed. The emergence of these multi-physics and multi-scale problems leads urban planners in a dead end, and Environmental Physics, which is not adapted to such fine scales, cannot support them.

2.3 From Building Physics

In the second half of the twentieth century, architects and HVAC engineers met around building energy efficiency problems and in the wider context of bioclimatic architecture [61]. Indeed, engineers introduce a scale much finer than that of the building, the equipment and devices one (windows carpentry, heat pumps ...), which we call "nano", for which they use finite-element-like computational methods [46]. However, at the micro level, these methods give way to nodal ones and geometry disappears, breaking dialogue with architects.

For decades, simulations have been limited to the study of single buildings. The scaling up to the urban block or neighborhood, made necessary by the hardening of the thermal regulations, is very difficult. Because of its too specific methods, Building Physics has as strong difficulty in getting, through the city, the macro scale, as well as Environmental Physics has to arrive at the micro scale.

2.4 From Smart City

The application of internet technology to more physical networks (smart grid), particularly urban ones (smart city), generally presupposes simplification of the geometry into topology. In doing so, many researchers in this field emphasize the "urban software" (distribution of electricity, water and transport) with respect to its "hardware" (buildings and infrastructure). This leads to privilege active systems over passive ones, thus optimizing margins rather than main topics, in the short term rather than in the long time.

However, the smart city framework is the first one that focuses on the urban scale and reaches both the nano scale (equipment) and the macro scale (for example, the national grid). It also generalizes the idea of providing permanent sensors to the city.

3 Urban Physics: A New Framework

3.1 The City as an Interface

We must first justify the nomenclature we have proposed for the scales, which is quite different from that of environmental physics. *Nano* (equipment, the meter range and below), *micro* (buildings, 10 m), *meso* (town, hundreds of meters, kilometers) and *macro* (tens and hundreds of kilometers) match with different regulatory frameworks and different actors, but also bring very different time scales. Indeed, the facilities have lifetimes of fifteen to twenty years, the buildings of the order of the century, and cities, although their development is accelerating, have often millennial layouts.

In a multi-scale analysis, the city becomes an interface between the buildings and the land. Exchange parameters are, for example, albedo, surface temperatures. An essential element is to preserve the geometry; otherwise it is not possible to understand the urban structure and to act on it or on its elements. The great recent advance is geometric: construction by procedural methods, adaptive level of detail…[20]

Zooming in the city, one sees the various buildings with their windows. The window is the primary interface between the outside and inside of the building, and the only one concerned by daylight. Windows are transparent to visible light (between 400 and 700 nm), but not for the thermal infrared: it is the greenhouse effect, which has very important consequences for the building thermal behavior. Special devices, such as glazed balconies or Trombe walls [22, 69], allow enjoying this effect at the level of an interior. Windows are generally quite complex devices, with balconies, shutters, curtains …that achieve a desired balance between solar gain, protection against excessive inputs, and reduction of thermal losses (insulation). The configuration of this interface may vary during the day (curtains) and in the year (mobile protection).

Such characteristics can be correlated to the multiscale study of composite materials [1, 44] dealing with domain decomposition methods, multi-scale and parallel simulations.

3.2 *Multiband Aspects of the Radiation Interacting with the Cities*

A black body is both perfect receiver and consequently perfect emitter. In 1900, Max Planck postulated that the electromagnetic energy of a black body is emitted not continuously (like by vibrating oscillators), but by discrete portions or quanta.

Planck's law states that the spectral radiance or radiance per unit wavelength interval L_λ, expressed in $\mathrm{W\,m^{-3}sr^{-1}}$ is given by:

$$L_{\Omega\lambda}\left(\lambda,T\right)=\frac{2hc^2}{\lambda^5}\frac{1}{e^{\frac{hc}{k\lambda T}}-1} \tag{1}$$

In this relation, T is the temperature expressed in K and λ, the wavelength expressed in m. This distribution can also be expressed in terms of frequency, but it is mandatory to express the transformation in terms of energy [72]. Let assume that the new function is $L_{\Omega\nu}$ which depends on ν and T. First, we write the equality:

$$L_{\Omega\lambda}\left(\lambda,T\right)d\lambda=L_{\Omega\nu}\left(\nu,T\right)d\nu \tag{2}$$

As $\nu=\mathrm{c}/\lambda$, $d\lambda/d\nu=-c/\nu^2$. For the next step, we can remove the negative sign which is simply clarifying that increasing wavelengths correspond to decreasing frequencies. Then

$$L_{\Omega\nu}\left(\nu,T\right)d\nu=L_{\Omega\lambda}\left(\lambda,T\right)\frac{d\lambda}{d\nu}=L_{\Omega\lambda}\left(\lambda,T\right)\frac{c}{\nu^2} \tag{3}$$

And finally:

$$L_{\Omega\nu}\left(\nu,T\right)=\frac{2h\,\nu^3}{c^2}\frac{1}{e^{\frac{h\nu}{kT}}-1} \tag{4}$$

which is obviously expressed in $\mathrm{W\,m^{-2}\,sr^{-1}s^{-1}}$

Three fundamental physical constants are present in these formula:

Planck's constant: h = 6,62606957 × 10^{-34} J s
Velocity of light: c = 299792458 $\mathrm{m\,s^{-1}}$
Boltzmann constant: k = 7,3806488 × 10^{-23} $\mathrm{J\,K^{-1}}$

Wien's law states that $\lambda_{max}T=2.898\times10^{-3}$

Fig. 1 Spectral radiance in $\mathrm{W\,m^{-3}sr^{-1}}$ as a function of the wavelength in nanometers

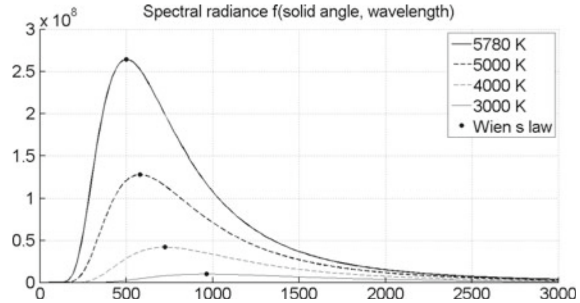
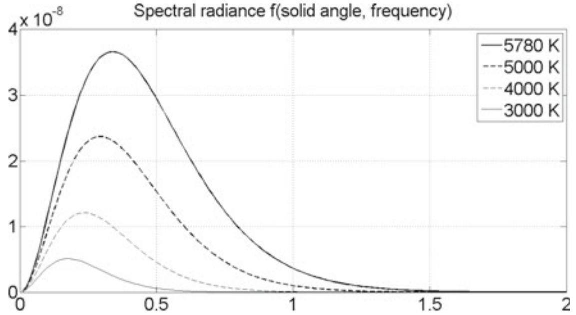


Fig. 2 Spectral radiance in $\mathrm{W\,m^{-2}sr^{-1}\,s^{-1}}$as a function of the frequency in hertz

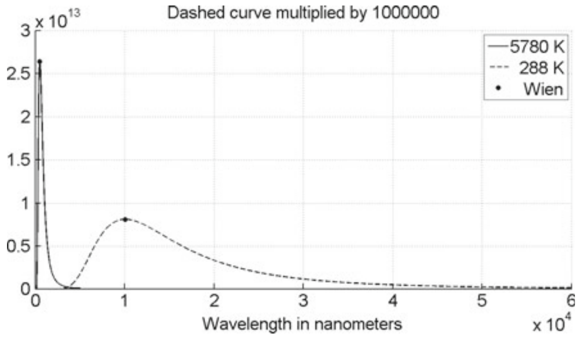


We can check it for the curves of Fig. 1, i.e. for the Sun temperature $T = 5780\,\mathrm{K}$, we obtain: $\lambda_{max} = 5.013 \times 10^{-7}\,\mathrm{m}$ ($\approx$500 nm).

The same data can also be expressed in terms of the frequency expressed in hertz ($\mathrm{s^{-1}}$). They are shown in Fig. 2. According to the interpretation of (2), the maximum values shown on (Fig. 1) cannot be transposed directly in Fig. 2.

If we try to get the same results for the two temperatures of Sun and Earth, we obtain two curves with very different scales, which imposes to multiply the second one by some factor (10^6 in the situation of Fig. 3).

Fig. 3 Spectral radiance in $\mathrm{W\,m^{-3}sr^{-1}}$ as a function of the wavelength in nanometers of Sun and Earth radiations



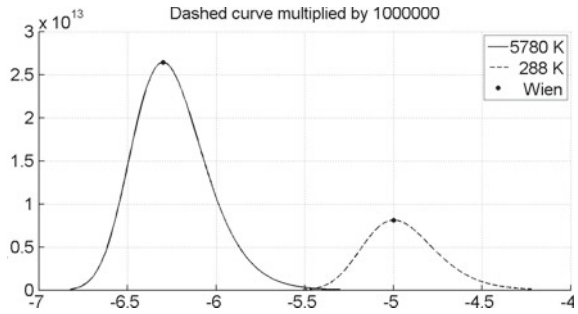


Fig. 4 Spectral radiance in $\mathrm{W\,m^{-3}\,sr^{-1}}$ as a function of the decimal logarithm of the wavelength

If we replace the decimal abscise by a decimal logarithmic one, the result is more compact and more understandable (Fig. 4).

As observed in (Figs. 3 and 4), the intersection of the continuous and the dashed curves occurs at ≈4 μm. This corresponds to the intersection point when the amplification factor of the dashed curve is equal to 30,000 ($\log_{10}(4 \times 10^{-6}) = -5.3979$).

As the black bodies spectra of ≈6000 K (Sun) and ≈300 K (Earth) are separated, it is possible to uncouple the corresponding radiations.

If one accepts the assumption of diffuse reflection essentially, it can be considered that each surface is characterized by a single parameter: the reflection coefficient. If one is interested in solar gains, this coefficient should be an average over the complete solar spectrum (between 0.32 and 4 μm), but its assessment is often restricted to the visible spectrum (from 0.4 to 0.7 μm) and weighted by the sensitivity curve of the human eye in daylight vision. For color images, the latter coefficient is split into three values (corresponding to the sensitivities of the RGB cones). The optical properties of the scene surfaces are first simplified (perfectly diffuse reflection) and then adapted to a particular receptor, the diurnal human vision. However, other receptors may be considered: plants (photosynthesis also relates to the band 0.4–0.7 μm, but with a very different sensitivity curve having its maximum at extreme red and purple, and a minimum at the center of the strip, the green color is generally reflected by the plant), photovoltaic cells (whose sensitivity extends into the near infrared, beyond micron) …

In the last decades important advances in the terrestrial and satellite measurement of solar radiation, as hyperspectral remote sensing [36, 37, 63], presage greater spectral accuracy in weather data and, therefore, greater rigor in the optical characterization of urban surfaces.

3.3 Shortwave

Because shortwaves and long waves are quite perfectly separated, the distribution of the shortwave on urban geometry does not depend on temperature. At the macro level, we only have data—meteorological ones (solar paths and clouds) and orographic ones (mountain masks—, now available all over the world. On the meso scale, the geometric model must be detailed enough (roofs slope, facades with their windows and balconies), but it has a reduced semantic of reflection coefficients. At the micro level, we go through the windows and the physical and geometric properties are maintained. On the nano scale, we can define the sensors (for example, computer monitors for the study of glare in natural light).

The simplicity of shortwave treatments makes them the necessary starting point to build a framework for multiscale analysis [5]. Shortwave analysis has the twofold advantage of a large experience acquired in the fields of entertainment and of recovering previously known principles of planners with full scale "test cases". We are presently ready to achieve shape optimization in this framework.

At the city level or at least at the district one, the first steps consist in establishing the possible objective functions, the constraints and the design parameters [10]. The main difficulty is that the general problem is typically formulated in terms of discrete variables and that the sensitivity analysis is not reachable, because the involved functions are mostly not derivable. Evolution algorithms are good candidates for this kind of optimization and have proven their effectiveness [43, 78].

Some improvements have been achieved with respect to the boundary and initial conditions in order to solve a problem closely related to the direct solar irradiation [79].

3.4 Long Waves

The next step is to go into long waves. One can imagine a city under a static atmosphere (only acting as a filter for the radiation) without inhabitants. Thermography has brought infrared in our visual experience, and it seems that we are now able to refine urban planning criteria on a better consideration of long waves (urban climate, urban comfort) [30]. The geometrical parameters calculated for the shortwave are also used for long waves (view factors), but with the consideration of temperatures, we should now at least study the radiative-conductive coupling [8]. The transition of current nodal methods to finite element ones [47, 62] has not been met yet, but the multi-scale analysis could provide an important argument in this direction (same method at all scales).

The Stefan-Boltzmann law states that the total energy (also known as irradiance or emissive power) radiated per unit surface of a black body per unit time is proportional to the fourth power of the black body thermodynamic temperature.

The total radiant energy emitted by a unit area of a blackbody radiator is obtained by integrating (1) over all wavelengths. The result is the Stefan-Boltzmann law:

$$Q = \sigma T^4 \tag{5}$$

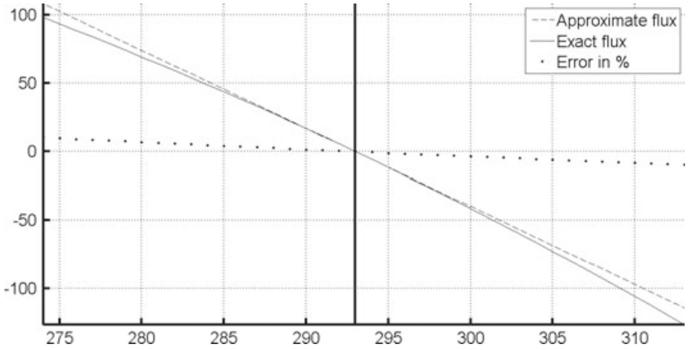


Fig. 5 Heat flux from wall at 293 K to wall at abscise temperature. Max and min error: 10.3 and −9.7 %

In this expression, Q is measured in W m^{-2}, T is the Kelvin temperature, and σ is the Stefan-Boltzmann constant (5.670373 $\times 10^{-8}$ W m^{-2} K^{-4}). It is linked to the universal constants by the relation:

$$\sigma = \frac{2\pi^5 k^4}{15h^3c^2} \tag{6}$$

Expressed as a function of the difference of temperatures $\Delta T = T_r - T_i$, with the reference temperature T_i, this law can be approximated by a linear relation [8, 51] so that, for instance, between 10 and 30 °C, the approximate solution built around 20 °C is giving a heat flux with less than 5 % error in the 20° interval around. For instance, between two infinite walls, the first one at 293 K, the flux increment for each degree more or less on the other wall is reaching $Q_{293} = 6$ W m^{-2} K^{-1} (Fig. 5). Between a source at 279 K and a receptor at 0 K, the flux is equal to 343.6 W m^{-2}. Note that the average Sun irradiance [7] at the top atmosphere (or without atmosphere on the ground) is of 342 W m^{-2}. The temperature of 279 K should be the Earth's temperature originating longwave radiations able to balance Sun's radiations in the lack of atmosphere.

4 Computational Model

The solution of radiative exchange problems is based either on ray tracing methods [39, 82] and their many variants, either on radiosity methods. The former are widely used in rendering while the latter were initially introduced in heat transfer problems [34].

Radiosity methods have the advantage of addressing the problem of radiative exchange for the entire scene. They proceed in two steps:

1. Calculation of the view factors
2. Solution of the radiosity equations

There is a clear separation between the pure geometrical step and the radiative calculations. The positive consequence is that the setting of the radiative problem is completely independent and can be modified retrospectively and inexpensively.

4.1 The Simplest Model

Given a very large urban 3D model, consisting of tens or hundreds of thousands of facets, we want to mesh most of these faces and to perform some calculations on the created meshes. The simplest calculations concern solid angles and view factors, or ultimately direct sunlight hours [9]. The most efficient solutions are using projections, mainly the stereographic one [13].

Let, for instance, calculate on the points of a virtual surface (a section of the city) the solid angle corresponding to the sky: the *SSA* (Sky Solid Angle) [25]. The sky contributes from all directions above the horizon that are not hidden by elements of the scene, and all these directions have equal weight.

If the same calculation is performed on real surfaces, and therefore opaque ones—ground of the street, facades, roofs—the geometric dimension that has a physical sense is the sky view factor (*SVF*), [60] which takes into account the fact that the grazing directions contribute less than the normal directions and only the directions whose scalar product with the normal negative contribution (an actual surface is necessarily oriented). The *SVF* is directly related to the diffuse or Lambert reflection [45].

Generally, *SSA* and *SVF* are expressed in percent, but while *SSA* is reported to the hemisphere of the sky, the *SVF* is related to the disk resulting from the orthogonal projection of the hemisphere on the plane containing the studied point. It is known as the Nusselt analogy [58]. Thus, in some configurations, *SVF* can be higher than *SSA*. For instance, the *SSA* of a spherical cap of opening α located on the top of the hemisphere is always greater than its *SVF*. Indeed, it is equal to $2\pi(1 - \cos\alpha)/2\pi = (1 - \cos\alpha)$ while its *SVF* is equal to $\pi \sin^2\alpha/\pi = \sin^2\alpha$. Their ratio is then equal to $\sin^2\alpha/(1 - \cos\alpha) = (1 - \cos^2\alpha)/(1 - \cos\alpha) = (1 + \cos\alpha)$ varying from 2 for a very small cap to 1 for the hemispherical cap. When α is small, the cap *SSA* tends to $\alpha^2/2$ while its *SVF* tends to α^2. Thus, $(SVF/SSA)_{\alpha\to 0} = 2$.

SSA and *SVF* depend only on the geometry of the scene and on the concept of horizontality (to define the sky vault). The sunshine hours add the notion of cardinal points (north direction), latitude and period of year, to set the solar paths. Most often, the sunshine hours are calculated on solstice’s extreme days and the equinox’s average days.

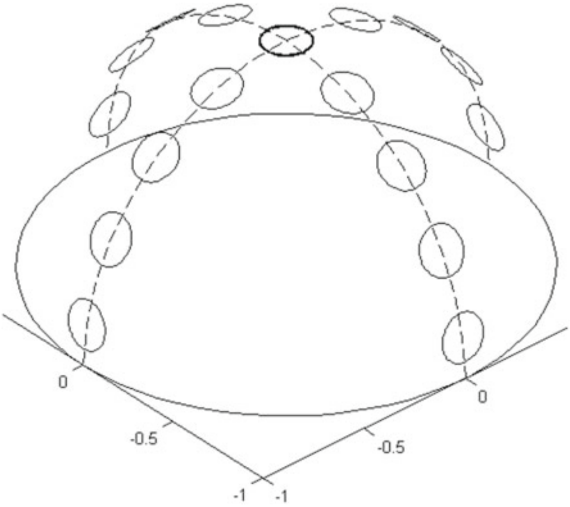


Fig. 6 Equal area disks defined on the spherical surface, so, with the same solid angles (*SSA*)

Figures 6 and 7 are showing a lot of disks located on two orthogonal meridians of a hemisphere. All the disks have the same area and thus the same *SSA*. The top disk of the hemisphere is seen in true scale in the center of Fig. 7. For the *SSA*, its cap area (very close to the disk area if it is small enough) is reported to the hemisphere area while, for the *SVF*, the corresponding disk area is reported to the base disk area validating the results presented above.

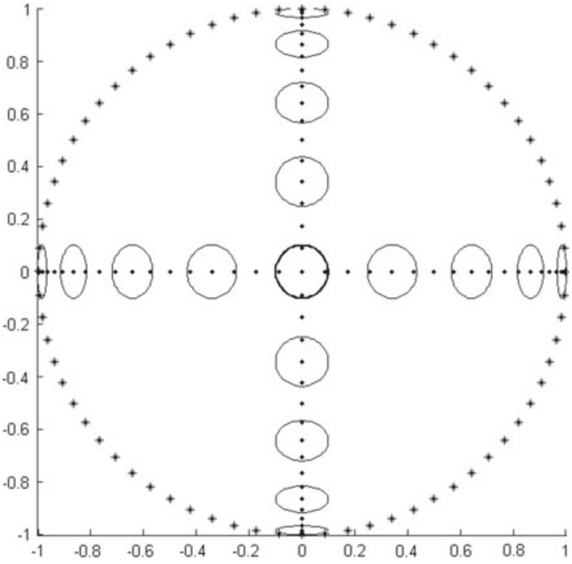


Fig. 7 *SVF* of the 17 equal area or equal solid angle spherical caps

Cities rarely have clear boundaries. More or less remote mountainous areas can greatly reduce the visibility of the sky, and thus the availability of the sun. With the exception of sea harbors or extremely flat regions, the horizon is rarely visible from the city, and only from the highest buildings.

In practice, we only manage a portion of the urban model, and the division is more or less arbitrary (e.g., according to administrative boundaries). Everything else can be projected on a cylinder centered on the studied area, in order to have a correct skyline. If the studied area is moving within the model, it is necessary to check that the encompassing cylinder remains accurate (parallax problem) [27].

Looking at the scene so organized from different points of the mesh, very different results are obtained. From the ground and the bottom of the facades, we often see very few objects, but close, while from the top of the facades, one can have a panoramic view of nearly half of the model. Numerous surfaces then appear with small view factors. From the roofs, the perceived scene is very different depending on whether the roof is tilted or horizontal. In the latter case, only are visible the upper portions of buildings higher than the considered one.

It is therefore natural to introduce technics of adaptive details in order to fully take into account these perceived scene changes, which can be very sudden, even browsing the mesh of a single flat surface, because the scene is consisting of discrete elements with well-defined edges.

The simplest city models are extrusions of urban maps. Because the radiation incident on a façade is divided into two parts (one reaches the wall, the other enters through the window), it is accounted from the glazing rate. This simplification is working well with the philosophy of the nodal methods, which tends to simplify the geometry to the maximum (e.g., a building is only represented by two nodes, one for the envelope and one for inside).

To improve the model [12], we can consider the windows as additions in front of the facades (Fig. 8). The advantage is that it avoids additional Delaunay mesh generation and that it keeps the model very simple. Thus, in (Fig. 9) all the large areas are correctly oriented (including the roofs) and the windows are present, with a final model which remains below 20,000 triangles.

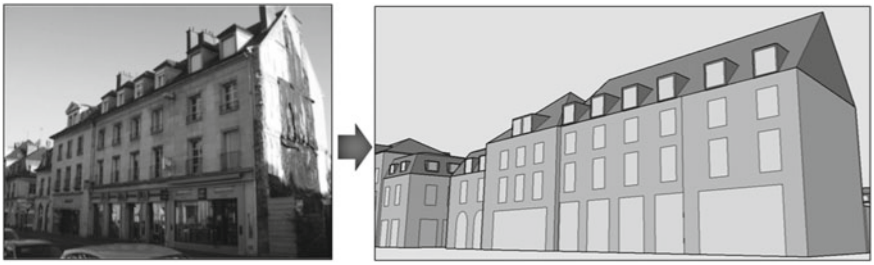


Fig. 8 Scheme of the modelling of individual buildings

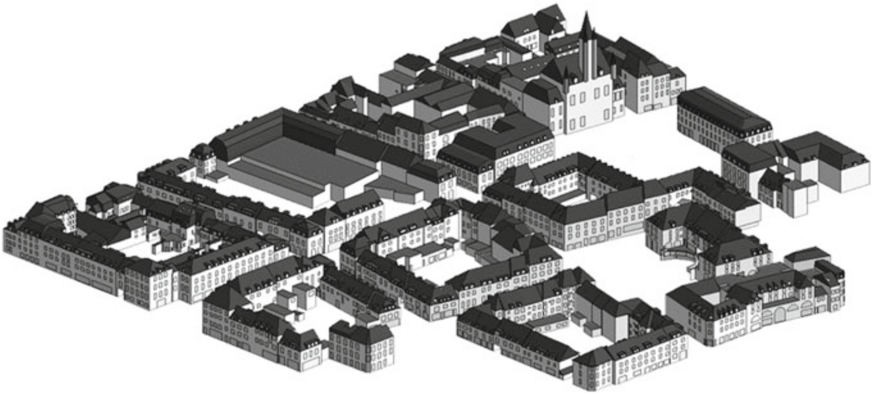


Fig. 9 Model of the Compiègne central district

It has been shown that it is impossible to take into account the thickness of the walls afterwards. To do this, we must request procedural methods and adaptive level of detail.

So, the idea of the pinhole [31] becomes very interesting, because it condenses incoming information on the window itself, and allows optimizing the shape of the window on the inner illumination criteria without restarting the computation at each step.

4.2 View Factors

In the solution of radiosity equations, the heaviest part is the computation of the coefficients of the matrix constituting the system. Indeed, the number of coefficients is potentially very high (square of the number of elements) and each one involves the treatment of the visible surface detection.

$$F_{ij} = \frac{1}{A_i} \int\limits_{A_i} \int\limits_{A_j} \frac{\cos\theta_i \cos\theta_j}{\pi r^2} V(Y_i, Y_j) dA_i dA_j \tag{7}$$

The view factor (also called form factor, angle factor or configuration factor) is the basic ingredient of radiative heat transfer studies [20, 71]. It defines the fraction of the total power leaving patch A_i that is received by patch A_j. Its definition is purely geometric. The angles θ_i and θ_j relate to the directions of the vector connecting the differential elements with the vectors normal to these elements; r is the distance between the differential elements.

Except in particular situations, it is not possible to compute the view factors explicitly [40]. An additional difficulty appears in presence of obstructions represented in the above expression by the visibility function $V\ (X_i, Y_j)$. This function is equal to 0 or 1 according to the possible presence of an obstacle that does not allow seeing an element Y_i from an element Y_j.

It is much easier to compute the differential view factor by removing the external integration that will be taken into account only in a second step to achieve the evaluation of the view factor, using, for instance, the Gauss integration rule in the concerned patch. The differential view factor in a point surrounded by the element of area *dS* is given by:

$$F_{dS-A_j} = \int\limits_{A_j} \frac{\cos\theta_i\ \cos\theta_j}{\pi r^2} V(Y_i, Y_j)\, dA_j \tag{8}$$

If the visibility function is everywhere equal to 1, the integration (4) performed on the full hemisphere is giving a view factor equal to 1. Spherical projections combined with Nusselt analogy provide an efficient solution of this problem [13].

4.3 Radiosity Equations

In order to solve efficiently the interaction problem, it is usual to set up a discrete formulation derived from the global illumination equation by making the following assumption. The environment is a collection of a finite number N of small diffusively reflecting patches each one, with uniform radiosity [15, 71]. A didactic approach of the radiosity equations solution was presented in [6], which includes the interpretation of importance, the dual of radiosity, obtained by the solution of the adjoint problem and is able to initiate new developments in discretization error analysis.

Radiosity is the radiometric quantity that is best suited for quantifying the illumination in a diffuse scene. In practice, when there is one single problem to solve, iterative solutions are used, which require the treatment of only one line of the matrix per iteration (see Sect. 4.4).

If the process is dynamic, for instance due to the movement of the Sun and the varying configurations of the sky [55], it is convenient to mesh the whole sky and to give to each element of its mesh the emittance corresponding to the concerned situation.

In this situation, it is more efficient to use the technique of combination of unitary right members [18]. It means that all the components of a column of the right members are zero except one which is equal to one. This system is solved for as many unitary right members as elements in the sky vault mesh, for instance, 145 cells in the Tregenza dome [73] and more if necessary (Fig. 10).

The creation of this kind of dome is very simple, because it is based on the two classical geographical coordinates (latitude and longitude). This choice facilitates the positioning and the navigation in the mesh [14]. Its definition is given by the sequence of numbers of elements in each ring. From these data, it is easy to compute

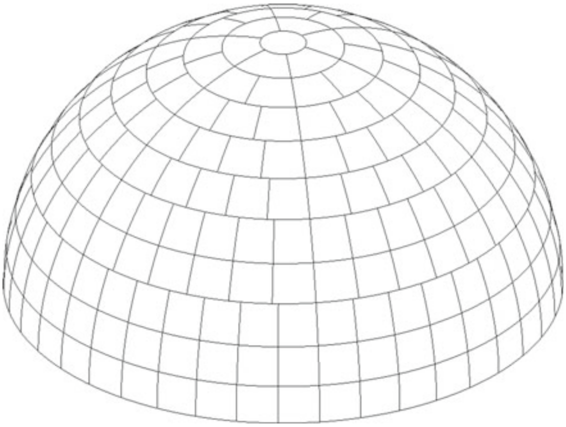


Fig. 10 Hemisphere composed of 289 equal area cells

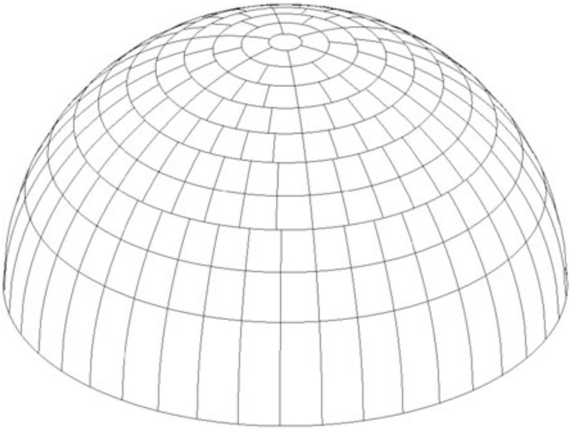


Fig. 11 Hemisphere composed of 289 equal view factor cells

the partition of a disk into equal area elements. The geometric transformation between a hemisphere and its equal area projection allows projecting the disk elements on the sphere either by using equal area projection *SSA* (Fig. 10) or equal view factor projection *SVF* (Fig. 11). The orthogonal projection of the dome of (Fig. 10) is shown in (Fig. 13) while the orthogonal projection of the dome of (Fig. 11) corresponds to Fig. 12.

After solving the radiosity equations, it is sufficient to recombine the solutions for each particular situation for which it is possible to evaluate the right member. The consequence is that the computation of the radiosities is very cheap.

Let us now define R, the diagonal matrix containing the hemispherical diffuse reflectances.

$$R_{ij} = \rho_i \delta_{ij} \tag{9}$$

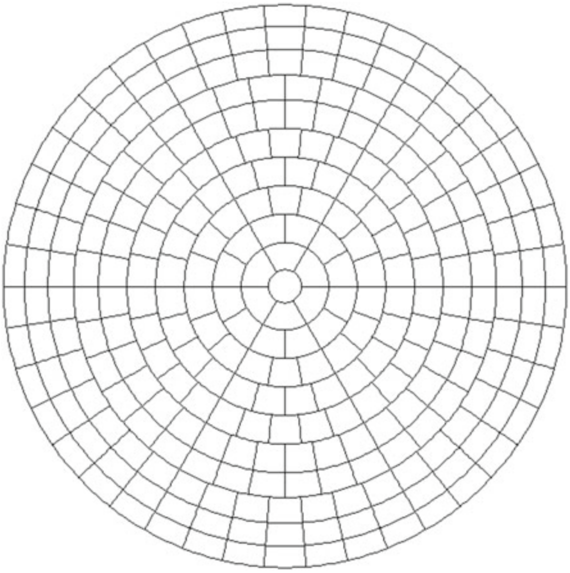


Fig. 12 Equal area cells in the base disk

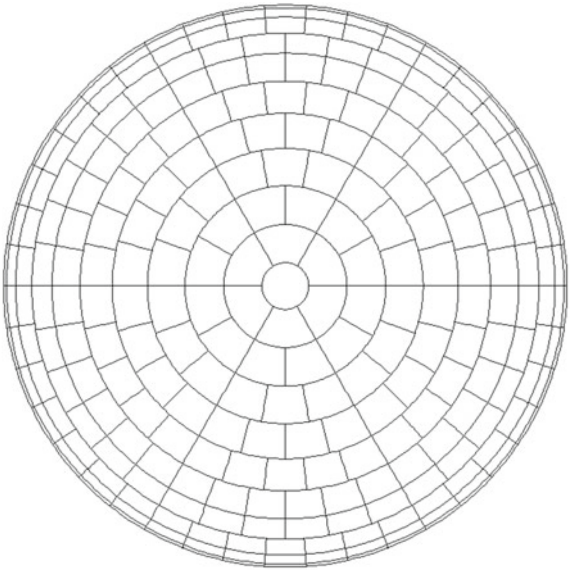


Fig. 13 Equatorial projection of the 289 equal solid angle cells

When the patches are planar polygons, the terms F_{ii} are equal to zero. These coefficients also verify the *closure property* when the environment (scene and sky), is taken into account:

$$F = \begin{pmatrix} F_{11} & F_{12} \cdots & F_{1N} \\ F_{21} & F_{22} & \vdots \\ \vdots & & \vdots \\ F_{N1} \cdots & \cdots & F_{NN} \end{pmatrix} \tag{10}$$

Let denote F the matrix of view factors coefficients between patches i and $j$as computed in (3):

$$\sum_{i=1}^{N} F_{ij} = 1; \quad i = 1, N \tag{11}$$

In the next formula, the components B_i of vector B are the radiosities, or, radiant fluxes per unit area, on patch i while the components E_i of E, are the radiant exitances. The radiosity equations can be written:

$$(I - RF)B = (I - G)\,B = MB = E \tag{12}$$

This discrete formulation leads to a linear system of equations for which many algorithms are available. The *RF* matrix, formed by the products of the view factors by the reflectances, is a non-symmetric matrix (except if all the reflectances and patch areas are equal), but the radiosity matrix M is diagonally dominant and well-conditioned.

In order to integrate the radiosity method in the environment of finite element method [35], it is suitable to work with symmetric matrices [56, 57].

The equation structure allows introducing another important property of the radiative exchanges: the *principle or reciprocity*

$$\forall(i, j) : A_i F_{ij} = A_j F_{ji} \tag{13}$$

We rewrite (12) explicitly and divide each line i by A_i/ρ_i

$$\frac{A_i}{\rho_i} B_i - A_i \sum_{k=1}^{n} B_k F_{ik} = \frac{A_i}{\rho_i} E_i \tag{14}$$

In pure diffuse reflection, this relation expresses the energy transfers between the N elements of the scene. If we use the reciprocity relation, we can transform (12) by multiplying the view factor matrix F by the diagonal matrix $S_{ij} = A_i \delta_{ij}$ of the patch areas.

We obtain then a symmetric matrix with N (N+1)/2 elements.

$$SF = \begin{bmatrix} 0 & A_1F_{12} & A_1F_{13} & \cdots \\ A_2F_{21} & 0 & A_2F_{23} & \cdots \\ A_3F_{31} & A_3F_{32} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{15}$$

Then, multiplying (12) by SR^{-1}, we can write:

$$\left(SR^{-1} - SF\right) B = SR^{-1}E \tag{16}$$

And in symmetrical form:

$$S\left(R^{-1} - F\right) B = SR^{-1}E \ \rightarrow \ B = \left(R^{-1} - F\right)^{-1} R^{-1}E \tag{17}$$

The second member $SR^{-1}E$ represents the incident power on the patch [3]. To solve this system of linear equations, a lot of very efficient methods are available. The Cholesky one [75] is very well known in the field of finite element method. We have good feedback for problems with more than one million of degrees of freedom. For thousands of degrees of freedom, it works very well on PCs.

In each line i of matrices F or SF, the nonzero terms indicate what elements are visible from element i. So, we can build an incidence matrix L composed of integers, which gives the connections between all the elements of the scene. It will help us to manage the system of equations and to identify possible ways to condense the system of equations.

Despite the fact that the heaviest part of the computation time is the evaluation of matrix F, we can also try to accelerate the step of solution by using iterative methods as explained in the next section.

4.4 Neumann Series

Because the matrix $G = RF$, defined in (11), has a norm less than one, the matrix M is invertible and the Neumann series of successive multiplications of G converges to its inverse [20, 81, 85].

$$f \ \|G\| < 1 \ then \ M^{-1} = [I - G]^{-1} = \sum_{a=0}^{\infty} G^a \tag{18}$$

This property gives indications to develop very efficient methods to solve these equations. It also gives justifications for iterative solutions. As noted by several authors [2, 26, 42], each step of the iterative process can be interpreted as the introduction of an additional reflection on all the elements of the scene.

The ability to decompose the solution of the radiosity equation in orders of reflection is very interesting, because it allows comparing this method with the ray tracing one, where the order of reflections is a usual stopping criterion. Therefore, the calculation is often stopped at the second reflection. This is true in ray trace software as Radiance [28], but it is also the case for a radiosity solver like *V-Ray* software (http://www.chaosgroup.com). In the latter, we can choose one or two reflections.

In a city, multiple reflections are possible, for instance between facades of narrow streets. Considering an average reflectance of 20 %, the energy flow is not superior to 20 % after the first reflection, 4 % after the second one and less than 1 % after the third one. However, if someone is interested in local results, the overall reasoning can be confusing, because the reflected energy may be the only available on certain surfaces, where it takes a considerable importance.

In an inner space, the radiation from the Sun and the sky through the window illuminates largely the floor and part of the walls, but it leaves the ceiling in full shade. The first reflection on the ground is the one that illuminates the ceiling. As it is generally light in color, the ceiling returns a second non-negligible reflection to the ground. This light is the first to reach parts of the ground from where the sky is not visible. Two reflections are therefore needed to get a realistic rendering of an interior space in natural light.

But what happens in an outdoor scene? In an urban scene, because we can almost always see a bit of the sky, the second reflection does not represent a substantial change in the results, and the following ones can be ignored (except in very specific configurations, as for example the entrance of a tunnel).

Modern cities all share some essential characteristics: a network of streets delineates parcels built with heights ranging from a few meters to tens of meters. However, other features are highly variable. This is the case of the coatings optical properties. Facades can be dark (brick) or light (limed walls), with a rate of glazing (and so, specular reflection) from few percent to almost 100 % (towers of glass and steel).

An important parameter of environmental physics is the albedo. This is an average reflection coefficient over a very large area. For instance, we can refer to the albedo of a planet (the Earth albedo is about 30 %, [7]). The albedo of sea ice, ocean, desert or forest is fairly easy to assess. Today, while cities cover large parts of the land area, it is necessary to know their albedo. However, the semi-regular structure of cities gives highly variable albedo. The relationship between apparently light and dark surfaces also depends on building height and density of the neighborhood.

Another characteristic of urban settings, due to the fact that cities are relatively low and very spread out, is that what we can see from a given point is very variable. From a window on a ground floor, the view can be limited to only two surfaces: the street and the facing wall. From a window at the top of a tower, we can see dozens, even hundreds of buildings. Calculating an urban geometry therefore strongly motivates to play on the buildings level of detail.

Distant buildings can be replaced by their prismatic envelopes. This kind of procedure has been used for a long time to accelerate the detection of visible surface. Several options are available; since bounding boxes [32] to prismatic envelopes and convex bounding polyhedrons.

5 Coupling Short and Long Waves in Transient Situations

The shortwave radiative exchanges limited to diffuse reflections can be calculated on the basis of the radiosity alone. More complete treatments, including the transient heat transfer in solids and the possible inclusion of the atmosphere, require more sophisticated methods. Starting from the classical equation of heat conduction in a solid:

$$div\,[k\ grad\ T] + Q = \gamma c_v \frac{\partial T}{\partial t} \tag{19}$$

Q is the heat density (W m^{-3}), T the temperature measured in degrees Kelvin, k the thermal conductivity (W m^{-1} K^{-1}), t is the time, c_v, the specific heat (J kg^{-1} K^{-1}) and γ, *the* density (kg m^{-3}). The variable conjugated with the temperature is the heat flux linked to the temperature gradient by Fourier's law.

To discretize these equations, the usual technique is the nodal method [50], used for instance in Esarad [64]. This technique also known as "Lumped Parameter Method" [65, 68, 78] offers a number of advantages among which we note that it highlights the thermal balance and heat flux. The fundamental assumption of the method is the use of isothermal nodes arranged in a network where they are connected by resistances and capacitances (electrical analogy). Its main drawback is that it requires a step of idealization from the geometrical definition of the model (CAD step) and the definition of the calculation model. For small models, it can be very useful, because it gives a summary of the exchanges. However, it offers only a coarse representation of the temperatures distribution.

An alternative is the finite elements method, originally developed in mechanical and civil engineering. In this method, the domain is covered with a congruent mesh. In each element, the field is replaced by a polynomial approximation respecting the required field continuity conditions through the interfaces (borders with neighboring elements). The border area consists of a boundary layer through which the exchanges occur with the fluid (here, the atmosphere). Through this boundary are also occurring radiative exchanges with the outside or with other elements of the scene.

In the thermal problem, the temperature field is discretized, so that the result of the simulation is a temperature map "*painted*" on the skin of the solid.

The boundary conditions consist of Dirichlet or essential conditions where the temperature T is imposed, natural or Neumann conditions where the heat flux is imposed, and Robin conditions, which are a weighted combination of Dirichlet and Neumann boundary conditions. These three zones cannot overlap and their union should be the total boundary.

The loads are of different natures:

Heat flow from the shortwave solar radiation. It is calculated separately in the "radiosity" module;

Long wave radiative fluxes are travelling towards other elements of the scene or to the atmosphere. They are proportional to the difference of fourth power of temperatures.

Convective flow proportional to the temperature difference between the surface of the solid and a reference point of the fluid in which it merges [30].

Any other heat flow that may be estimated directly or expressed in terms of temperatures, for example, evapotranspiration [53].

In brief, the solid subjected to these heat flows and where the temperature is known at least at one point will experience modifications as a result of internal heat conduction and the ability of materials to store heat. In finite element calculation, it is a classical problem, its theory was developed in the 1970s [33].

The time component is calculated by a finite difference method. At the level of discretization, we must ensure that the temporal pattern is consistent with the spatial discretization.

In this problem, the main difficulty is the calculation of the view factors of the surfaces brought into contact. It may be assumed they have been calculated in the previous step (shortwave). If they must apply for both analyses, both meshes have to coincide.

To calculate the convective exchanges, one must know the temperature of the air, which requires in principle to include modeling of the fluid.

The calculation of thermal interactions in the city encompasses three major phases:

the definition of the geometry, which must be structured and has to allow processing of very large volumes of data,

the view factors calculation, which involves the effective detection of hidden or viewed parts,

and finally, the solution of the equations of transient heat conduction in the coupled conduction-radiation problem.

The methods proposed to solve these problems are qualified, but it is still necessary to verify that the computation time is acceptable.

Treatment of massive geometric data takes advantage of advances in procedural methods and in "LOD" (methods of levels of detail) [20]. The calculation of view factors can take advantage of the progresses made in the Monte Carlo methods or in the effective treatment of hemispheric and stereographic projections.

For the solution of the coupled system, the choice of the finite element method is motivated by its ability to provide a temperature map that can be easily compared to telemetry results [76].

Today, the finite element method is widely used to solve nonlinear problems of millions of degrees of freedom and benefits from the attention of programmers who have optimized the algorithms. It may be accused of producing an enormous amount of results but the task of identifying the relevant information is reduced through visualization techniques. Use of optimization techniques and sensitivity calculation is another decisive tool to assist in the understanding and interpretation of the analyzed phenomena.

5.1 Improving the Performances of the Finite Element Solution Using Super Elements

The set of transient equations is linear with respect to conduction and convection, but not to radiation. In clear sky conditions, due to higher differences of temperatures (about −50 °C in the zenith direction, up to 50 °C or more on the ground), the heat exchanges between sky and city are highly nonlinear. However, for cloudy skies, the difference can drop drastically [80] leading to an apparent sky temperature very close to the ambient one. It is then very interesting to condense in a super element the linear part of the model and to iterate on the degrees of freedom corresponding to the elements of the city participating at high level to the heat fluxes going to the sky, i.e. by selecting only the roofs or other elements of the scene [8]. The superelement technique is well known in the fields of structural mechanics and civil engineering [20]. It is extensively used in the modeling of large structures like full aircraft or oil platforms. For this problem, procedural methods will help to organize the data [19].

5.2 Other Aspects of the City Behavior Simulation

Another challenge is to consider the convection, which works well with the nodal methods [5]. Developments have been performed, dealing with this aspect or with the interaction of heat and fluid dynamics [21, 83]. The ventilation aspects could then be addressed in this framework, with a clear objective: to be able to carry everywhere—in the squares, in the bottom of streets and inside buildings—the measurements made at specific points of the city (on some roofs, near the airport) [24].

In general the problems of fluids structures or fluids solids interactions are still a major field of research and development in the frame of multiscale and multiphysic disciplines [38, 77].

Solar radiation reaches the Earth's surface after passing through the atmospheric layer. It appears in different forms: direct, diffuse, reflected by the environment or other elements of the scene. Other phenomena also contribute, for example, the physical and chemical reactions that take place in the atmosphere and the phenomena induced by vegetation [23].

Finally we can follow the example of previous works concerned by large dimension problems [48], but wich will take benefits of the still increasing enhancements of the computers memory and processors.

6 Conclusion

A complete model of multi-scale energy exchanges should allow the following simulations: investigating major urban design options with their impact on the overall urban energy efficiency (which, among other things, allows a better evaluation of smart city type proposals); helping urban planning (optimization of urban forms based on energy conservation criteria); supplying atmospheric models at the macro scale.

Further developments should deal with the following steps:

1. A "shortwave" model for quickly achieving fast simulations on very large geometric models;
2. A "long wave" model to simulate thermography;
3. A "multiphysics" approach for finding optimal solutions adapted to the urban project;
4. A complete physical model coupling thermodynamics and aeraulics.

On the basis of an urban geometry model built on procedural principles with an adaptive level of detail, the shortwave model must be able to simulate the distribution of solar radiation from meteorological data, with time steps of a few minutes, and taking into account the reflections, in order to calculate the main thermal parameters (solar gain) and the luminous ones (Daylight Autonomy DA [67] and Useful Daylight Illuminance UDI [54]). The calculation time is an important issue for the three main anticipated applications: calculation of the urban albedo variation, urban design assistance and optimization of urban shapes (on criteria such as solar energy potential or photovoltaic solar access).

For the long wave model, calculation of surface temperatures can be achieved at a reasonable cost if we agree to simplify the convection contribution. Thermography is now present in our visual experience, and architects wish to use it soon as component of simulation in their projects. The many existing thermographic images, including at urban scale (above ground thermography performed using satellites, aircraft or drones), are altogether test cases available to calibrate the simulations. This objective requires leaving the nodal methods for Finite Element Methods.

Multiphysic studies should then focus on heat, light, acoustics and cross ventilation. To optimize the urban shapes, the last one is indeed much more available—and then efficient—than forced ventilation (wind). Therefore, these studies do not necessarily need very sophisticated simulations, but rather to follow a methodical order still to be explored.

Finally, the complete physical model should be able to take into account the heat-fluid coupling, with, as a primary objective, the transposition anywhere in the city of the data collected by weather stations. To achieve the quantification of the different contributions to urban climate, multiscale analysis and model reduction techniques will undoubtedly be necessary.

To optimize a city on criteria of comfort or energy efficiency, it is clear that prior understanding of urban climate and precise quantification of the different contributions to this climate are absolutely necessary. This is even truer in order to act on the air quality at the urban scale. Pioneering works have shown, first, that the relevant choice of a LoD (Level of Details) of the urban 3D model is essential [49] and, secondly, that any major action on the climate must be preceded by an analysis allowing to assess not only the effectiveness of the different possible actions, but also the order in which they should be carried out, otherwise unexpected—and possibly dangerous—results are obtained. Cities are systems, and a systemic approach is mandatory.

We believe that achieving the first three objectives is needed before we can work on the last one seriously. Another necessary condition is that the FEM community is interested in this subject. Raising this interest has been the main motivation of this chapter.

References

1. Allix O Gosselet P Kerfriden P Saavedra K (2012) Virtual Delamination Testing through Non-Linear Multi-Scale Computational Methods: Some Recent Progress. CMC: Computers, Materials, & Continua, 2012, 32 (2), pp. 107–132.
2. Arvo JR (1993) Linear Operators and Integral Equations in Global Illumination. In: Global Illumination, SIGGRAPH ‘93 Course Notes, vol 42, August.
3. Ashdown I (2002) Radiative Transfer Networks Revisited. Journal of the illuminating Engineering Society pages 38–51 Vol. 31, Issue 2.
4. Beckers B (2009) Geometrical interpretation of sky light in architecture projects. In: Actes de la Conférence Internationale Scientifique pour le BATiment CISBAT, September EPFL, Lausanne, Suisse.
5. Beckers B (2011a) Impact of Solar Energy on Cities Sustainability, in Architecture & Sustainable Development. In: Bodart M & Evrard A (ed) Proc. 27th Int. Conf. on Passive and Low Energy Architecture (PLEA 2011), Louvain-la-Neuve, Belgium.
6. Beckers B (2011b) Urban outlines 2D abstraction for flexible and comprehensive analysis of thermal exchanges, Conférence Internationale Scientifique pour le BATiment CISBAT 2011, EPFL, September 2011, Lausanne, Suisse.
7. Beckers B (2012) Worldwide aspects of solar radiation impact. In: Solar Energy at Urban Scale, chap. 5, Ed. B. Beckers, John Wiley and Sons Inc.
8. Beckers B (2013) Taking Advantage of Low Radiative Coupling in 3D Urban Models. In: Eurographics Workshop on Urban Data Modelling and Visualisation, Girona.
9. Beckers B Masset L (2006) Heliodon 2 Documentation and Software, www.heliodon.net.
10. Beckers B Beckers P (2008) Optimization of daylight in architectural and urban projects. In: 2th International Conference on Multidisciplinary Design Optimization and Applications, ASMDO, Gijon, Spain.
11. Beckers B Rodriguez D (2009) Helping architects to design their personal daylight. In: WSEAS Transactions on Environment and Development, Issue 7, Volume 5, pp 467–477.
12. Beckers B Rodríguez D Antaluca E Batoz JL (2010) About solar energy simulation in the urban framework: The model of Compiègne. In: 3rd International Congress Bauhaus SOLAR, 5 pages, November 10 & 11, 2010, Erfurt, Germany.
13. Beckers B Masset L Beckers P (2011) The universal projection for computing data carried on the hemisphere, Computer-Aided Design, Volume 43, Issue 2, Pages 219–226, February 2011.
14. Beckers B Beckers P (2012a) A general rule for disk and hemisphere partition into equal-area cells, Computational Geometry - Theory and Applications, Volume 45, Issue 7, Pages 275–283.
15. Beckers B Beckers P (2012b) Radiative Simulation Methods, in Solar Energy at Urban Scale, chap. 10, Ed. B. Beckers, John Wiley and Sons Inc, pp 205–236.
16. Beckers B Beckers P (2014a) Super Element Technique for Solar Energy Optimization at Urban Level. In: Proceedings of OPT-i An International Conference on Engineering and Applied Sciences Optimization M. Papadrakakis, M.G. Karlaftis, N.D. Lagaros (eds.) Kos Island, Greece, 4–6 June.

17. Beckers B Beckers P (2014b) Reconciliation of Geometry and Perception in Radiation Physics John Wiley and Sons Inc, 192 pages.
18. Beckers B Beckers P (2014c) Sky vault partition for computing daylight availability and shortwave energy budget on an urban scale. In: Lighting Research and Technology, vol. 46 n°. 6, Pages 716–728, December.
19. Besuievsky G Patow G (2011) A procedural modeling approach for automatic generation of LoD building models. In: Proceedings CISBAT 2011, Lausanne, Switzerland pp. 993–998.
20. Besuievsky G Barroso S Beckers B Patow G (2014) A Configurable LoD for Procedural Urban Models intended for Daylight Simulation. Eurographics Workshop on Urban Data Modelling and Visualization, Strasbourg, France.
21. Bouyer J Inard C Musy M (2011) Microclimatic coupling as a solution to improve building energy simulation in an urban context, Energy and Buildings Volume 43, Issue 7, July, Pages 1549–1559.
22. Boyer H Lucas F Miranville F Bastide A Morau D (2006) Simulations de dispositifs du type Mur Trombe avec CODYRUN ESIM 2006, May 2006, Toronto, Canada. pp. 81–233.
23. Bruse M Fleer H (1998) Simulating surface-plant-air interactions inside urban environments with a three dimensional numerical model. Environmental Modelling & Software 13 373–384.
24. Bueno B Roth M Norford L Li R (2014) Computationally efficient prediction of canopy level urban air temperature at the neighbourhood scale. In: Urban Climate 9, 35–53.
25. Capeluto G (2012) Dense Cities in Temperate Climates: Solar and Daylight Rights. In: Solar Energy at Urban Scale, chap. 13, Ed. B. Beckers, John Wiley and Sons Inc.
26. Cohen MF Wallace JR (1993) Radiosity and Realistic Image Synthesis, Academic Press.
27. Collin O (2008) Facteur de ciel : un paramètre d'évaluation des masques en architecture Master Thesis University of Liège Belgium.
28. Compagnon R (2004) Solar and daylight availability in the urban fabric, Energy and Buildings 36 321–328.
29. Eigenbrod VF Bell A Davies HN (2011) The impact of projected increases in urbanization on ecosystem services", Proceedings of the Royal Society B. 278:3201–3208.
30. Erell E Pearlmutter D Williamson T (2011) Urban Microclimate - Designing the Spaces Between Buildings, Routledge.
31. Fernández E Besuievsky G (2015) Inverse opening design with anisotropic lighting incidence. Comput. Graph., vol. 47, pp. 113–122.
32. Foley JD van Dam A Feiner SK Hughes JF (1990) Computer Graphics, principles and practice, Addison-Wesley publishing Company, Second Edition.
33. Fraeijs de Veubeke B Hogge M (1972) Dual analysis for heat conduction problems by finite elements. In: International Journal for Numerical Methods in Engineering, Volume 5, Issue 1, pages 65–82, September/October.
34. Gebhart B (1961) Heat Transfer McGraw-Hill Book Company.
35. Heckbert PS Winget JM (1991) Finite Element Methods for Global Illumination. In: Technical report CSD-91-643, UC Berkeley, January 8.
36. Herold M Gardner ME Noronha V Roberts DA (2003) - Spectrometry and Hyperspectral Remote Sensing of Urban Road Infrastructure, Online Journal of Space Communication Issue 3.
37. Herold M Roberts DA Gardner ME Dennison PE (2004) Spectrometry for urban area remote sensing – Development and analysis of a spectral library from 350 to 2400 nm, Remote Sensing of Environment 91 304–319.
38. Hou G Wang J Layton A (2012) Numerical Methods for Fluid-Structure Interaction - A Review Commun. Comput. Phys. Vol. 12, No. 2, pp. 337–377 August.
39. Howell JR (1997) The Monte Carlo Method in Radiative Heat Transfer Journal of Heat Transfer August 1998 Vol 120 547–560.
40. Howell JR Siegel R Menguc MP (2011) Thermal Radiation Heat Transfer", 5th ed., Taylor and Francis CRC, New York.
41. Jouzel J Debroise A (2014) Le défi climatique, Dunod.
42. Kajiya JT (1986) The Rendering Equation, ACM Siggraph Dallas August 18–22 Vol. 20, N. 4.

43. Kämpf JH Robinson D (2009) A hybrid CMA-ES and HDE optimization algorithm with application to solar energy potential, in Applied Soft Computing, vol. 9, p. 738–745.
44. Ladevèze P Loiseau O Dureisseix D (2001) A micro-macro and parallel computational strategy for highly heterogeneous structures. Int. Journal for Numerical Methods in Engineering, 52:121–138.
45. Lambert JH (1760) Photometria sive de mensura et gradibus luminis, colorum et umbrae German translation by E. Anding in Ostwald's Klassiker der Exakten Wissenschaften, Vol. 31–33, Leipzig, 1892. Cited by Peter Schröder & Pat Hanrahan, "A Closed Form Expression for the Form Factor between Two Polygons", Research Report CS-TR-404-93, January 1993.
46. Lenik K Wójcicka-Migasiuk D (2010) FEM applications to the analysis of passive solar wall elements. Journal of Achievements in Materials and Manufacturing Engineering Volume 43 Issue 1: 333–340 November.
47. Lewis RW Morgan K Thomas HR Seetharamu KN (1996) The Finite Element Method in Heat Transfer Analysis", John Wiley & Sons.
48. Mardaljevic J Janes GM (2012) Multiscale Daylight Modeling for Urban Environments. In: Solar Energy at Urban Scale, chap. 8, Ed. B. Beckers, John Wiley and Sons Inc.
49. Martilli A Roulet YA Junier M Kirchner F Rotach MW Clappier A (2003) On the impact of urban surface exchange parameterisations on air quality simulations: the Athens case Atmospheric Environment 37 4217–4231.
50. Meyer RX (1999) Elements of space technology for aerospace engineers Chap. 7, "Spacecraft thermal design", Academic Press.
51. Monteith JL Unsworth MH (2007) Principles of Environmental Physics, 3rd edition, Elsevier, 2007.
52. Moonen P Defraeye T Dorer V Blocken B Carmeliet J (2012) Urban Physics: Effect of the micro-climate on comfort, health and energy demand Frontiers of Architectural Research 1, 197–228.
53. Musy M (2012) Evapotranspiration. In: Solar Energy at Urban Scale, chap. 7, Ed. B. Beckers, John Wiley and Sons Inc.
54. Nabil A Mardaljevic J (2005) Useful daylight illuminance: a new paradigm for assessing daylight in buildings. Light. Res. Technol., vol. 37, no. 1, pp. 41–59, January.
55. Nahon R Vermeulen T Beckers B (2013) an Adaptive 3D Model for Solar Optimization at the Urban Scale. In: IET/IEEE Second International Conference on Smart and Sustainable City, August 19–20 Shanghai, China.
56. Neumann L Tobler RF (1994) New Efficient Algorithms with Positive Definite Radiosity Matrix. In: Proceedings of the Fifth Eurographics Workshop on Rendering, Darmstadt, Germany, June.
57. Nievergelt Y (1997) Making Any Radiosity Matrix Symmetric Positive Definite, Journal of the Illuminating Engineering Society, Vol. 26, No. 1, pp. 165–171, Winter.
58. Nusselt W (1928) Graphische bestimmung des winkelverhaltnisses bei der wärmestrahlung", Zeitschrift des Vereines Deutscher Ingenieure, 72(20):673.
59. Oke TR (1978) Boundary layer climates, Methuen & Co., Ltd.
60. Oke TR (1981) Canyon geometry and the nocturnal urban heat island: comparison of scale model and field observations. J Climatology 1:237–254.
61. Olgyay V Olgyay A (1963) Design With Climate: Bioclimatic Approach to Architectural Regionalism Princeton University Press.
62. Panczak T Ring S Welch M (1997) A CAD-based Tool for FDM and FEM Radiation and Conduction Modeling SAE 1998 Transactions - Journal of Aerospace - V 107–1.
63. Pirard T (2012) The Odyssey of Remote Sensing from Space: Half a Century of Satellites for Earth Observations. In: Solar Energy at Urban Scale, chap. 1, Ed. B. Beckers, John Wiley and Sons Inc.
64. Planas P Flett DC (1998) ESARAD: From R&D to Industrial Utilisation", ESA bulletin 93 - February.
65. Ramallo-González AP Eames ME Coley DA (2013) Lumped parameter models for building thermal modelling: An analytic approach to simplifying complex multi-layered constructions, Energy and Buildings 60 174–184.

66. Rasheed A Robinson D Clappier A Naranayan C Lakehal D (2011) Representing complex urban geometries in mesoscale modeling. In: International Journal of Climatology, 31: 289–301.
67. Reinhart CF Wienold J (2011) The Daylighting Dashboard - A Simulation-Based Design analysis for Daylit Spaces. In: Building and Environment Volume 46 Issue 2 pages 386–396.
68. Robinson D Haldi F Kämpf JH Leroux P Perez D Rasheed A Wilke U (2009) CitySim: Comprehensive Micro-simulation of Resource Flows for Sustainable Urban Planning. In Eleventh International IBPSA Conference Glasgow, Scotland, July 27–30.
69. Saadatian O Sopian K Lim CH Asim N Sulaiman MY (2012) Trombe walls: A review of opportunities and challenges in research and development Renewable and Sustainable Energy Reviews 16 (2012) 6340–6351.
70. Seto KC Güneralp B Hutyra LR (2012) Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. In: Proceedings of the National Academy of Sciences of USA, 109(40): 16083–16088.
71. Sillion FX Puech C (1994) Radiosity and Global Illumination, Morgan Kaufmann Publishers Inc.
72. Soffer BH Lynch DK (1999) Some paradoxes, errors, and resolutions concerning the spectral optimization of human vision Am. J. Phys. 67 (11), November.
73. Tregenza PR (1987) Subdivision of the sky hemisphere for luminance measurements, Lighting Research and Technology 19 (1), pp. 13–14.
74. UN-Habitat (2011) Cities and Climate Change: Global Report on Human Settlements.
75. van de Geijn R (2011) Notes on Cholesky Factorization. In: Report TX 78712 University of Texas at Austin.
76. van Eekelen T (2012) Radiation Modeling Using the Finite Element Method. In Solar Energy at Urban Scale, chapter 11, Ed. B. Beckers, John Wiley and Sons Inc.
77. van Loon R Anderson PD van de Vosse FN Sherwin SJ (2007) Comparison of various fluid-structure interaction methods for deformable bodies Computers & Structures Volume 85, Issues 11–14, June-July Pages 833–843.
78. Vermeulen T Kämpf JH Beckers B (2013) Urban Form Optimization for the Energy Performance of Buildings Using Citysim. In CISBAT 2013, Sept. 4–6, Lausanne, Switzerland pp. 915–920.
79. Vermeulen T Knopf-Lenoir C Villon P Beckers B (2015) Urban layout optimization framework to maximize direct solar irradiation. In: Computers, Environment and Urban Systems, 51, p. 1–12.
80. Vollmer M Möllmann KP (2010) Infrared Thermal Imaging: Fundamentals, Research and Applications. Wiley-VCH, 612 pages.
81. Wang X Cen S Li C (2013) Generalized Neumann Expansion and Its Application in Stochastic Finite Element Methods, Mathematical Problems in Engineering Volume 2013, Article ID 325025, 13 pages Hindawi Publishing Corporation.
82. Ward GJ (1994) The RADIANCE Lighting Simulation and Rendering System", SIGGRAPH '94 Proceedings of the 21st annual conference on Computer graphics and interactive techniques.
83. Zhai ZJ Chen QY (2006) Sensitivity analysis and application guides for integrated building energy and CFD simulation. Energy and Buildings 38 1060–1068.
84. Zhang GJ Cai M Hu A (2013) Energy consumption and the unexplained winter warming over northern Asia and North America. In: Nature Climate Change, 3, 466–470.
85. Zhu D Li B Liang P (2015) On the Matrix Inversion Approximation Based on Neumann Series in Massive MIMO Systems. In: arXiv:1503.05241v1 [cs.IT] 17 Mar 2015 accepted to conference; Proc. IEEE ICC 201.

A Path-Following Method Based on Plastic Dissipation Control

Boštjan Brank, Andjelka Stanić and Adnan Ibrahimbegovic

Abstract A path-following method that is based on controlling incremental plastic dissipation is presented. It can be applied for analysis of an elasto-plastic solid or structure. It can be also applied for complete failure computation of a solid or structure that is performed by using a material failure model. In this work, we applied it for computations with the embedded-discontinuity finite elements that use rigid-plastic cohesive laws with softening to model material failure process. The most important part of the path-following method is the constraint function. Several constraint functions are derived and proposed for geometrically nonlinear small strain elasto-plasticity with linear isotropic hardening. The constraint functions are also derived for the embedded-discontinuity finite elements. In particular, they are derived for 2-d solid (and frame) embedded-discontinuity finite elements that describe cohesive stresses (or forces and moments) in the discontinuity curve (or point) by rigid-plasticity with softening. Numerical examples are presented in order to illustrate performance of the discussed path-following method.

1 Introduction

The most used path-following method in the nonlinear finite element analysis of solids and structures is probably the Crisfield's cylindrical arc-length method, see e.g. [1]. It can be successfully used for solving geometrically linear problems as well as many types of geometrically and materially nonlinear problems. However,

B. Brank (✉) · A. Stanić
Faculty of Civil and Geodetic Engineering, University of Ljubljana,
Jamova cesta 2, 1000 Ljubljana, Slovenia
e-mail: bbrank@fgg.uni-lj.si

A. Ibrahimbegovic
Laboratory Roberval, University of Technology of Compiègne,
Compiègne, France

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_2

it might fail when computing solid or structural failure due to material failures. For this kind of problems, several modified arc-length methods were proposed, see e.g. [7, 15] and references therein. The problem of those modifications is that they are very much problem dependent. A general and robust path-following method for complete failure computation of solids and structures due to material failures is still to be designed.

The most important part of the path-following method is the constraint equation. Recently, [17] presented a path-following method based on constraint equation that is controlling dissipation of inelastic material. This kind of path-following method can be used when solid or structural material is modelled by an inelastic material model, e.g. by an elasto-plastic or damage model. In [17], several constraint functions were presented. In particular, they derived constraint functions for geometrically linear and geometrically nonlinear damage, and geometrically linear elasto-plasticity.

In this work we extent the ideas of [17] to geometrically nonlinear small strain elasto-plasticity and to embedded-discontinuity formulations. In particular, we derive explicit and implicit constraint functions that control incremental dissipation for small strain elasto-plasticity with isotropic hardening. Moreover, we derive explicit constraint functions, based on plastic dissipation control, for solid and beam embedded-strong-discontinuity-in-displacements (or rotation) finite elements. These kind of elements have become during the last years an interesting tool for modelling and simulation of solid or structural failure due to material failure, see e.g. [8, 12, 13]. Two of the authors of this work have been involved in derivation of the embedded-strong-discontinuity finite elements for analysis of different structures and solids. We refer to [2, 9, 14] for planar Euler-Bernoulli beams (Fig. 1), to [10] for Timoshenko beams (Fig. 2), and to [3, 4, 6] for 2d-solids (Fig. 3). In several cases, convergence problems were observed when computing failure analysis with those complex elements and standard cylindrical arc-length method. This problem is addressed in this work. The aim of this work is therefore a derivation of a novel path-following method, based on dissipation control, which should be more robust for analysis of solid and structural failure problems by the embedded-discontinuity finite elements.

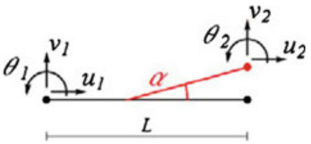


Fig. 1 Euler-Bernoulli beam finite element with embedded-strong-discontinuity in rotation. Cross-section softening in rotation is described by rigid-plasticity with softening

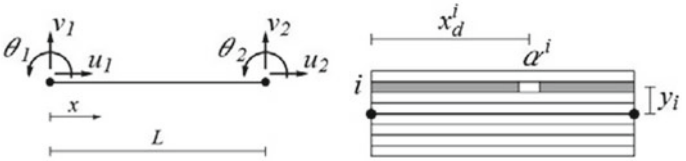


Fig. 2 Multi-layered Timoshenko beam finite element with layer-wise embedded-strong discontinuity in axial displacement. Material failure at each layer is described by rigid-plasticity with softening

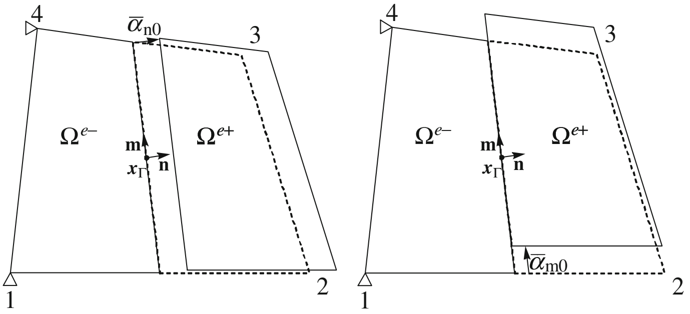


Fig. 3 2-d solid embedded-discontinuity quadrilateral: constant separations along the discontinuity line for mode I (*left*) and mode II (*right*). Cohesive stresses in discontinuity are described by rigid-plasticity with softening

2 Path-Following Method Framework

In the nonlinear finite element method for solids and structures, one has to solve a system of nonlinear equations related to the equilibrium of the nodes of the finite element mesh

$$\boldsymbol{R}\left(\boldsymbol{u}\left(t\right),\lambda\left(t\right)\right)=\boldsymbol{R}^{int}\left(\boldsymbol{u}\left(t\right)\right)-\boldsymbol{f}^{ext}\left(\lambda\left(t\right)\right)=\boldsymbol{0} \qquad (1)$$

where $\boldsymbol{R}^{int}$ and $\boldsymbol{f}^{ext}$ are vectors of internal and external (equivalent) nodal forces (and moments, if they are present in the formulation), respectively, $\boldsymbol{u}$ is vector of unknown nodal displacements (and rotations, if they are present in the formulation), λ is the load convergence problems when factor, and $t \geq 0$ is a monotonically increasing parameter called the pseudo-time or the arc-length. In many practical cases, the system of Eq. (1) is possible to solve only by introducing an additional constraint equation

$$g\left(\boldsymbol{u}\left(t\right)-\boldsymbol{u}\left(t-\Delta t\right),\lambda\left(t\right)-\lambda\left(t-\Delta t\right)\right)=0 \qquad (2)$$

where Δ is a (small) incremental change. Solving (1) and (2) simultaneously is called the path-following method or the arc-length method, see e.g. [1]. The solution of (1) and (2) is searched for at discrete pseudo-time points $0 = t_0, t_1, \ldots, t_n, t_{n+1}, \ldots, t_{final}$. Assume that configuration of solid or structure at t_n is known; it is defined by the pair $\{\boldsymbol{u}(t_n), \lambda(t_n)\} = \{\boldsymbol{u}_n, \lambda_n\}$. At searching for the next configuration at $t_{n+1} = t_n + \Delta t_n$, we decompose $\boldsymbol{u}_{n+1}$ and λ_{n+1} as

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + \Delta\boldsymbol{u}_n, \quad \lambda_{n+1} = \lambda_n + \Delta\lambda_n \tag{3}$$

where $\Delta\boldsymbol{u}_n$ and $\Delta\lambda_n$ are the increments of the displacement vector and the load vector, respectively. With (3), Eqs. (1) and (2) can be rewritten for t_{n+1} as

$$\begin{aligned} &\boldsymbol{R}_{n+1}(\boldsymbol{u}_n, \lambda_n; \Delta\boldsymbol{u}_n, \Delta\lambda_n) = \boldsymbol{0} \\ &g_{n+1}(\Delta\boldsymbol{u}_n, \Delta\lambda_n) = 0 \end{aligned} \tag{4}$$

where $\Delta\boldsymbol{u}_n$ and $\Delta\lambda_n$ are the unknowns. The solution of (4) is searched for iteratively by the Newton-Raphson method. At iteration i, the following linear system has to be solved

$$\begin{bmatrix} \boldsymbol{K}^i_{n+1} & \boldsymbol{R}^i_{n+1,\lambda} \\ \left[g^i_{n+1,\boldsymbol{u}}\right]^T & g^i_{n+1,\lambda} \end{bmatrix} \begin{Bmatrix} \Delta\tilde{\boldsymbol{u}}^i_n \\ \Delta\tilde{\lambda}^i_n \end{Bmatrix} = - \begin{Bmatrix} \boldsymbol{R}^i_{n+1} \\ g^i_{n+1} \end{Bmatrix} \tag{5}$$

for the pair $\left\{\Delta\tilde{\boldsymbol{u}}^i_n, \Delta\tilde{\lambda}^i_n\right\}$, where $(\circ)_{,\lambda}$ and $(\circ)_{,\boldsymbol{u}}$ denote the derivatives of $(\circ)$ with respect to $\Delta\lambda_n$ and $\Delta\boldsymbol{u}_n$, respectively, and $\boldsymbol{K}^i_{n+1} = \boldsymbol{R}^i_{n+1,\boldsymbol{u}}$ is the tangent stiffness matrix. New iterative guess is obtained as $\Delta\boldsymbol{u}^{i+1}_n = \Delta\boldsymbol{u}^i_n + \Delta\tilde{\boldsymbol{u}}^i_n$ and $\Delta\lambda^{i+1}_n = \Delta\lambda^i_n + \Delta\tilde{\lambda}^i_n$. System of Eq. (5) can be effectively solved by the bordering algorithm, see e.g. [18] for details. When the iteration loop ends due to fulfilment of a convergence criterion, the configuration $\{\boldsymbol{u}_{n+1}, \lambda_{n+1}\}$ at t_{n+1} is obtained and search for the solution at the next pseudo-time point can start.

The above presentation is valid for any kind of the constraint function g_{n+1} in (4). However, the robustness and efficiency of the path-following method depend crucially on the specific form of this function. In what follows, we will elaborate for the case when g_{n+1} controls the structural plastic dissipation, which can be computed when elasto-plastic and/or rigid-plastic material models are used.

3 Dissipation Constraint for Geometrically Nonlinear Small Strain Elasto-plasticity

In this section, we will present several formulations for defining constraint equation $g_{n+1} = 0$ in (4) that controls structural plastic dissipation.

3.1 Explicit Formulation—Version 1

The rate of plastic dissipation in an elasto-plastic solid or structure is defined as (see e.g. [8])

$$\dot{D} = \dot{P} - \dot{\Psi} \tag{6}$$

where $\dot{P}$ is the pseudo-time rate of the total energy the solid/structure is receiving, and $\dot{\Psi}$ is the rate of the thermodynamic (i.e. the free energy or the stored energy) potential for plasticity. For the discretized solid/structure in the framework of the geometrically nonlinear and inelastic finite element method, $\dot{P}$ can be written as

$$\dot{P} = \int\limits_V \boldsymbol{S}^T \dot{\boldsymbol{E}} dV = \boldsymbol{f}^{ext,T} \dot{\boldsymbol{u}} = \lambda \hat{\boldsymbol{f}}^{ext,T} \dot{\boldsymbol{u}} \tag{7}$$

where $\boldsymbol{S}$ and $\boldsymbol{E}$ are vectors comprising 2nd Piola-Kirchhoff stresses and Green-Lagrange strains, respectively, and V is initial volume. Moreover, it was assumed in (1) and (7) that the external forces are conservative and can be described as $\boldsymbol{f}^{ext} = \lambda \hat{\boldsymbol{f}}^{ext}$, where $\hat{\boldsymbol{f}}^{ext}$ is a fixed pattern of nodal forces. The free energy potential (i.e. the stored energy) of a solid/structure, based on the St. Venant-Kirchhoff elasticity and plasticity with linear isotropic hardening, is

$$\Psi = U + H \tag{8}$$

where the stored energy due to elastic deformations is

$$U = \int\limits_V \frac{1}{2} \boldsymbol{E}^{e,T} \boldsymbol{D} \boldsymbol{E}^e dV = \int\limits_V \frac{1}{2} \boldsymbol{S}^T \boldsymbol{D}^{-1} \boldsymbol{S} dV \tag{9}$$

and the stored energy due to material hardening is

$$H = \int\limits_V \frac{1}{2} K_h \xi_h^2 dV \tag{10}$$

Here, $\boldsymbol{E}^e = \boldsymbol{E} - \boldsymbol{E}^p$ are elastic strains, $\boldsymbol{E}^p$ are plastic strains, $\boldsymbol{D}$ is symmetric constitutive matrix, $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{E}^e$, K_h is hardening modulus, and ξ_h is strain-like variable that controls isotropic hardening. For any other type of isotropic and/or kinematic hardening, H in (10) has to be changed accordingly. Derivation of U with respect to the pseudo-time gives

$$\dot{U} = \int\limits_V \dot{\boldsymbol{E}}^T \boldsymbol{C}^{ep} \boldsymbol{D}^{-1} \boldsymbol{S} dV = \dot{\boldsymbol{u}}^T \int\limits_V \boldsymbol{B}^T \boldsymbol{C}^{ep} \boldsymbol{D}^{-1} \boldsymbol{S} dV \tag{11}$$

where $\boldsymbol{C}^{ep}$ and $\boldsymbol{B}$ denote the consistent symmetric elasto-plastic tangent modulus and the strain-displacement matrix, respectively. The following relations were used in (11): $\dot{\boldsymbol{S}} = \boldsymbol{C}^{ep}\dot{\boldsymbol{E}}$, $\dot{\boldsymbol{E}} = \boldsymbol{B}\dot{\boldsymbol{u}}$. Derivation of H with respect to the pseudo-time yields

$$\dot{H} = \int\limits_V K_h \xi_h \dot{\xi}_h dV = \int\limits_V K_h \xi_h \left(\frac{\partial \xi_h}{\partial \boldsymbol{u}}\right)^T \dot{\boldsymbol{u}} \tag{12}$$

Let us use the forward Euler pseudo-time step to express dissipation at pseudo-time point t_{n+1} by using known dissipation at t_n

$$D_{n+1} = D_n + \dot{D}_n \Delta t_n, \quad \Delta t_n = t_{n+1} - t_n \tag{13}$$

Let us further define the following constraint equation

$$g_{n+1} = D_{n+1} - D_n - \tau_n = 0 \tag{14}$$

where τ_n is a predefined (required) value of dissipation at pseudo-time step $\left[t_n, t_{n+1}\right]$. It follows from (13), (6)–(8), (11) and (12) that (14) can be rewritten as

$$g_{n+1} = \dot{D}_n \Delta t_n - \tau_n = \Delta \boldsymbol{u}_n^T \left(\lambda_n \hat{\boldsymbol{f}}^{ext} - \boldsymbol{f}_n^*\right) - \tau_n = 0 \tag{15}$$

where $\Delta \boldsymbol{u}_n = \dot{\boldsymbol{u}}_n \Delta t_n$ was defined, and

$$\boldsymbol{f}_n^* = \int\limits_V \boldsymbol{B}_n^T \boldsymbol{C}_n^{ep} \boldsymbol{D}^{-1} \boldsymbol{S}_n dV + \int\limits_V K_h \xi_{h,n} \left(\frac{\partial \xi_h}{\partial \boldsymbol{u}}\right)_n dV \tag{16}$$

It follows from (15) that the derivatives needed in (5) are simply

$$g_{n+1,\lambda} = 0\ , \quad g_{n+1,\boldsymbol{u}} = \lambda_n \hat{\boldsymbol{f}}^{ext} - \boldsymbol{f}_n^* \tag{17}$$

Most of the terms in Eq. (16) are computed during the elasto-plastic analysis and can be readily used to compute (15) and (17). An exception is $(\partial \xi_h / \partial \boldsymbol{u})_n$. In practice, one should only compute $\boldsymbol{f}_n^*$ for configuration at t_n and use it in the path-following method when iterating to find configuration at t_{n+1}.

The second integral on the right hand side of (16) should be usually smaller than the first integral, which might be a justification for neglecting the former integral when computing (16), i.e.

$$\boldsymbol{f}_n^* \rightarrow \boldsymbol{f}_n^{*,approx} = \int\limits_V \boldsymbol{B}_n^T \boldsymbol{C}_n^{ep} \boldsymbol{D}^{-1} \boldsymbol{S}_n dV \tag{18}$$

In such a case, the corresponding approximation $\left(\dot{D}_n \Delta t_n\right)^{approx}$ would be bigger than $\dot{D}_n \Delta t_n$ in (15).

3.2 Explicit Formulation—Version 2

An alternative for the expression (6) for the rate of plastic dissipation in an elasto-plastic solid or structure is (see e.g. [8])

$$\dot{D} = \int\limits_V \left(\dot{\boldsymbol{E}}^{p,T} \boldsymbol{S} + \dot{\xi}_h q\right) dV \tag{19}$$

where $q = -K_h \xi_h$. Since $\dot{\boldsymbol{E}}^p = \dot{\boldsymbol{E}} - \dot{\boldsymbol{E}}^e = \left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}^{ep}\right)\dot{\boldsymbol{E}}$, one can rewrite (19) as

$$\dot{D} = \int\limits_V \dot{\boldsymbol{u}}^T \boldsymbol{B}^T \left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}^{ep}\right)^T \boldsymbol{S} dV - \int\limits_V \dot{\xi}_h K_h \xi_h dV \tag{20}$$

If (20) is used in (15), the constraint Eq. (15) transforms to

$$g_{n+1} = \dot{D}_n \Delta t_n - \tau_n = \Delta \boldsymbol{u}_n^T \bar{\boldsymbol{f}}_n - \tau_n = \boldsymbol{0} \tag{21}$$

where

$$\bar{\boldsymbol{f}}_n = \int\limits_V \boldsymbol{B}_n^T \left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}_n^{ep}\right)^T \boldsymbol{S}_n dV - \int\limits_V K_h \xi_{h,n} \left(\frac{\partial \xi_h}{\partial \boldsymbol{u}}\right)_n dV \tag{22}$$

Comparison of (15) and (16) with (21) and (22) yields (note that $\left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}_n^{ep}\right)^T = \boldsymbol{I} - \boldsymbol{C}_n^{ep}\boldsymbol{D}^{-1}$)

$$\lambda_n \hat{\boldsymbol{f}}^{ext} = \int\limits_V \boldsymbol{B}_n^T \boldsymbol{S}_n dV \tag{23}$$

which states that the equivalent external nodal forces are in equilibrium with the internal nodal forces at t_n, see e.g. [8], a condition already accomplished at the start of the current pseudo-time step $\left[t_n, t_{n+1}\right]$. This leads us to a conclusion that constraints (15) and (21) are completely equivalent but computed differently. The derivatives of (21) are

$$g_{n+1,\lambda} = 0\,, \quad g_{n+1,\boldsymbol{u}} = \bar{\boldsymbol{f}}_n \tag{24}$$

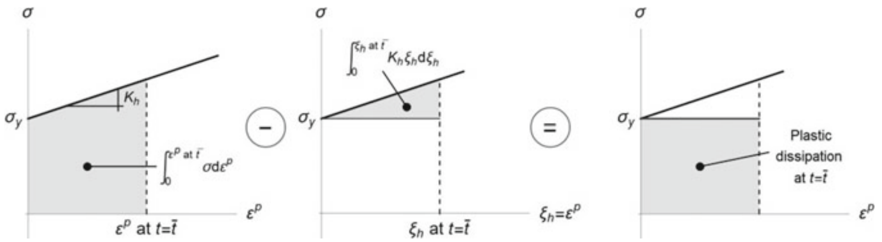

σ
K_h
σ_y
$\int_0^{\varepsilon^p \text{ at } \bar{t}} \sigma d\varepsilon^p$
ε^p
ε^p at $t=\bar{t}$
−
σ
$\int_0^{\xi_h \text{ at } \bar{t}} K_h \xi_h d\xi_h$
σ_y
ξ_h at $t=\bar{t}$
$\xi_h=\varepsilon^p$
=
σ
σ_y
Plastic
dissipation
at $t=\bar{t}$
ε^p

Fig. 4 Plastic dissipation at bulk point of elasto-plastic material with hardening for pseudo-time increment $[t_y, \bar{t}]$ and 1d case

It can be seen from (19) and (22) that approximation (18) corresponds to approximation

$$\dot{D} \rightarrow \dot{D}^{approx} = \int_V \dot{\boldsymbol{E}}^{p,T} \boldsymbol{S} dV \tag{25}$$

For a 1d case, i.e. stretching/compressing of a bar, integration of (25) relates to Fig. 4 (left), and integration of (19) and (20) relates to Fig. 4 (right). In Fig. 4, plastic yielding at a material point of such a bar is presented, where σ, ε^p, and σ_y denote stress, plastic strain and yield stress, respectively. Let us look at the plastic dissipation at the end of the pseudo-time increment $[t_y, \bar{t}]$, where t_y is the pseudo-time point at the beginning of plastic yielding. Since the stress monotonically increases during this increment, $\varepsilon^p = \xi_h$, see [8]. Plastic dissipation at a material point at $t = \bar{t}$ is the grey area on Fig. 4 (right). When energy storage in the material point due to material hardening is neglected, an approximation of plastic dissipation at $t = \bar{t}$ is obtained, which is bigger than the plastic dissipation and corresponds to the grey area on Fig. 4 (left).

Equation that corresponds to (18) and (25), i.e. (15) without the "hardening term" (more precisely, (15) without the second integral on the right hand side of Eq. (16)), was used in [17] for geometrically linear elasto-plastic problems. However, it is clear from the above derivations that the constraints (15) and (21) can be both used for geometrically linear as well as for geometrically nonlinear small-strain elasto-plastic problems.

3.3 Implicit Formulations

In this section, we present implicit counterparts of version-1 and version-2 explicit formulations presented above. The backward Euler pseudo-time step can be used in (13), i.e.

$$D_{n+1} = D_n + \dot{D}_{n+1}\Delta t_n, \quad \Delta t_n = t_{n+1} - t_n \tag{26}$$

which leads to the following constraint equation (compare with (15))

$$g_{n+1} = \dot{D}_{n+1}\Delta t_n - \tau_n = \Delta \boldsymbol{u}_n^T \left(\lambda_{n+1}\hat{\boldsymbol{f}}^{ext} - \boldsymbol{f}_{n+1}^*\right) - \tau_n = 0 \tag{27}$$

where

$$\boldsymbol{f}_{n+1}^* = \int_V \boldsymbol{B}_{n+1}^T \boldsymbol{C}_{n+1}^{ep} \boldsymbol{D}^{-1} \boldsymbol{S}_{n+1} dV + \int_V K_h \xi_{h,n+1} \left(\frac{\partial \xi_h}{\partial \boldsymbol{u}}\right)_{n+1} dV \tag{28}$$

Note that $\Delta \boldsymbol{u}_n$ in (27) is defined as $\Delta \boldsymbol{u}_n = \dot{\boldsymbol{u}}_{n+1}\Delta t_n$. Expressions for the derivatives of g_{n+1} are much more complex in comparison with (17) and can be written as

$$g_{n+1,\lambda} = \Delta \boldsymbol{u}_n^T \hat{\boldsymbol{f}}^{ext}, \quad g_{n+1,\boldsymbol{u}} = \lambda_{n+1}\hat{\boldsymbol{f}}^{ext} - \boldsymbol{f}_{n+1}^* - \left(\boldsymbol{f}_{n+1,\boldsymbol{u}}^*\right)^T \Delta \boldsymbol{u}_n \tag{29}$$

The problem is to derive $\boldsymbol{f}_{n+1,\boldsymbol{u}}^*$, which demands, among other derivatives, the derivative of elasto-plastic tangent modulus $\boldsymbol{C}_{n+1,\boldsymbol{u}}^{ep}$.

Alternatively to (27), the constraint equation can be expressed as (compare with (21))

$$g_{n+1} = \dot{D}_{n+1}\Delta t_n - \tau_n = \Delta \boldsymbol{u}_n^T \bar{\boldsymbol{f}}_{n+1} - \tau_n = 0 \tag{30}$$

where

$$\bar{\boldsymbol{f}}_{n+1} = \int_V \boldsymbol{B}_{n+1}^T \left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}_{n+1}^{ep}\right)^T \boldsymbol{S}_{n+1} dV - \int_V K_h \xi_{h,n+1} \left(\frac{\partial \xi_h}{\partial \boldsymbol{u}}\right)_{n+1} dV \tag{31}$$

and

$$g_{n+1,\lambda} = 0, \quad g_{n+1,\boldsymbol{u}} = \bar{\boldsymbol{f}}_{n+1} - \left(\bar{\boldsymbol{f}}_{n+1,\boldsymbol{u}}\right)^T \Delta \boldsymbol{u}_n \tag{32}$$

where, again, $\bar{\boldsymbol{f}}_{n+1,\boldsymbol{u}}$ calls for derivative of elasto-plastic tangent modulus $\boldsymbol{C}_{n+1,\boldsymbol{u}}^{ep}$.

It is obvious that $\boldsymbol{f}_{n+1,\boldsymbol{u}}^*$ and $\bar{\boldsymbol{f}}_{n+1,\boldsymbol{u}}$, needed in (29) and (32), respectively, are not easy to derive and compute, which renders explicit formulations more attractive for implementation than implicit. On the other hand, the explicit formulations might turn to be less robust than the implicit ones.

4 Dissipation Constraint for Embedded Discontinuity Finite Elements

Let us derive the plastic dissipation constraint for a situation when the material failure in solid is modelled by the embedded-displacement-discontinuity finite element formulation and inelastic softening cohesive traction-separation law is used at the discontinuity. In what follows, we will restrict to 2-d solids with a single fracture curve (i.e. with a single discontinuity) and to frames with softening plastic hinges. Let the bulk of the 2-d solid or frame be modelled as elastic and let the cohesive stresses at the discontinuity be modelled by rigid-plasticity with linear softening.

The free energy potential (i.e. the stored energy) of the solid can be written as

$$\Psi = (U - S) + S_s \tag{33}$$

where the stored energy due to elastic deformations of the bulk

$$U = \int_V \frac{1}{2}\boldsymbol{E}^T\boldsymbol{D}\boldsymbol{E}dV = \int_V \frac{1}{2}\boldsymbol{S}^T\boldsymbol{D}^{-1}\boldsymbol{S}dV \tag{34}$$

is diminished for S due to localized plastic deformations at the failure curve. Due to softening rigid-plasticity, those plastic deformations equal to kinematic variables $\boldsymbol{\alpha}$ that describe material separation along the discontinuity. The S in (33) is defined as

$$S = \int_\Gamma \boldsymbol{\alpha}^T\boldsymbol{t}d\Gamma \tag{35}$$

where $\boldsymbol{t}$ is vector of cohesive stresses in discontinuity, and Γ is length of the discontinuity curve. The S_s in (33) is due to the linear softening and takes the form

$$S_s = \int_\Gamma \frac{1}{2}K_s\xi_s^2 d\Gamma \tag{36}$$

In (36), $K_s < 0$ is softening modulus, and ξ_s is displacement-like variable that controls softening. The pseudo-time derivatives of (34), (35) and (36) are

$$\dot{U} = \int_V \dot{\boldsymbol{E}}^T\boldsymbol{S}dV, \quad \dot{S} = \int_\Gamma \dot{\boldsymbol{\alpha}}^T\boldsymbol{t}d\Gamma, \quad \dot{S}_s = \int_\Gamma K_s\xi_s\dot{\xi}_s d\Gamma \tag{37}$$

The derivatives in (37) can be expressed by $\dot{\boldsymbol{u}}$ using $\dot{\boldsymbol{E}} = \boldsymbol{B}\dot{\boldsymbol{u}}$ and the chain rule

$$\dot{\boldsymbol{\alpha}} = \frac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{u}}\dot{\boldsymbol{u}}, \quad \dot{\xi}_s = \left(\frac{\partial\xi_s}{\partial\boldsymbol{\alpha}}\right)^T\frac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{u}}\dot{\boldsymbol{u}} \tag{38}$$

The constraint equation can be defined for forward Euler pseudo-time step (according to the version-1 of above presented explicit formulation, see (15)) as

$$g_{n+1} = \dot{D}_n\Delta t_n - \tau_n = \Delta\boldsymbol{u}_n^T\left(\lambda_n\hat{\boldsymbol{f}}^{ext} - \boldsymbol{f}_n^*\right) - \tau_n = 0 \tag{39}$$

where $\boldsymbol{f}_n^*$ in (39) is now defined as

$$\boldsymbol{f}_n^* = \int\limits_V \boldsymbol{B}_n^T \boldsymbol{S}_n dV + \underbrace{\int\limits_\Gamma \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{u}}\right)_n^T \boldsymbol{t}_n d\Gamma + \int\limits_\Gamma K_s \xi_{s,n} \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{u}}\right)_n^T \left(\frac{\partial \xi_s}{\partial \boldsymbol{\alpha}}\right)_n d\Gamma}_{\bar{\boldsymbol{f}}_n^*} \tag{40}$$

The equilibrium of solid or structure at t_n demands equality of external and internal nodal forces, i.e. $\lambda_n \hat{\boldsymbol{f}}^{ext} = \int\limits_V \boldsymbol{B}_n^T \boldsymbol{S}_n dV$. Thus, inserting (40) in (39) yields

$$g_{n+1} = \dot{D}_n \Delta t_n - \tau_n = \Delta \boldsymbol{u}_n^T \bar{\boldsymbol{f}}_n^* - \tau_n = 0 \tag{41}$$

where $\bar{\boldsymbol{f}}_n^*$ is indicated in (40). The derivatives of g_{n+1} are the expressions from (17) with $\boldsymbol{f}_n^*$ from (40). In the implementation of embedded-discontinuity finite elements, kinematic variables $\boldsymbol{\alpha}$ are condensed on the element level. This enables to compute $(\partial \boldsymbol{\alpha} / \partial \boldsymbol{u})_n$ in (40) as assembly of element contributions. Since the condensation on the element level (e) yields

$$\Delta \boldsymbol{\alpha}_n^{(e)} = \left(\boldsymbol{K}^{\boldsymbol{\alpha\alpha},(e)}\right)_n^{-1} \boldsymbol{K}_n^{\boldsymbol{\alpha u},(e)} \Delta \boldsymbol{u}_n^{(e)}, \quad \boldsymbol{K}_n^{(e)} = \begin{bmatrix} \boldsymbol{K}^{\boldsymbol{uu}} & \boldsymbol{K}^{\boldsymbol{u\alpha}} \\ \boldsymbol{K}^{\boldsymbol{\alpha u}} & \boldsymbol{K}^{\boldsymbol{\alpha\alpha}} \end{bmatrix}_n^{(e)} \tag{42}$$

one has

$$\left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{u}}\right)_n^{(e)} = \left(\boldsymbol{K}^{\boldsymbol{\alpha\alpha},(e)}\right)_n^{-1} \boldsymbol{K}_n^{\boldsymbol{\alpha u},(e)} \tag{43}$$

where $\boldsymbol{K}_n^{(e)}$ is the element stiffness matrix at t_n, which can be decomposed as shown in (42). How $(\partial \xi_s / \partial \boldsymbol{u})_n$ in (40) is computed will not be further elaborated.

The third integral on the right hand side of (40) might be neglected, i.e.

$$\bar{\boldsymbol{f}}_n^* \rightarrow \bar{\boldsymbol{f}}_n^{*,approx} = \int\limits_\Gamma \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{u}}\right)_n^T \boldsymbol{t}_n d\Gamma \tag{44}$$

The corresponding approximation $\left(\dot{D}_n \Delta t_n\right)^{approx}$ is smaller than $\dot{D}_n \Delta t_n$ in (41) since $K_s < 0$. This is illustrated for 1d case, i.e. stretching/compressing of a bar, in Fig. 5, where plastic dissipation for a point at the discontinuity is presented. For a 1d case integration of (41) relates to Fig. 5 (right), and integration of (41) by using (44) relates to Fig. 5 (left). In Fig. 5, f is cohesive stress, f_f is material failure stress at which

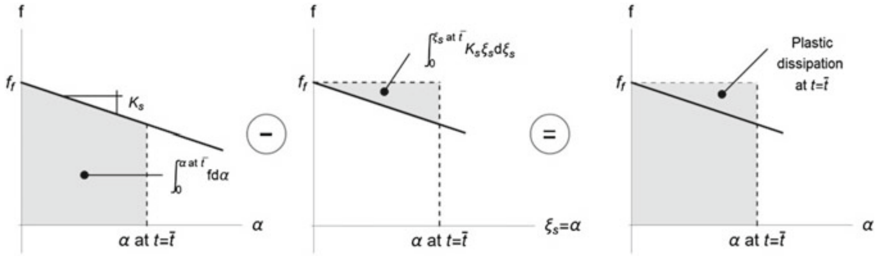

f
f_f
K_s
$\int_0^{\alpha \text{ at } \bar{t}} f d\alpha$
α
α at $t=\bar{t}$
−
f
f_f
$\int_0^{\xi_s \text{ at } \bar{t}} K_s \xi_s d\xi_s$
α at $t=\bar{t}$
$\xi_s=\alpha$
=
f
f_f
Plastic dissipation at $t=\bar{t}$
α at $t=\bar{t}$
α

Fig. 5 Plastic dissipation at discontinuity point for rigid-plastic material with linear softening for pseudo-time increment $[t_y, \bar{t}]$ and 1d case

softening begins, and α is separation. The plastic dissipation at the end of the pseudo-time increment $[t_f, \bar{t}]$ is shown, where t_f is pseudo-time point at material failure. Since the cohesive stress monotonically decreases during this increment, $\alpha = \xi_s$. Plastic dissipation at a material point at $t = \bar{t}$ is the grey area on Fig. 5 (right). When material softening is neglected, an approximation of plastic dissipation at $t = \bar{t}$ is obtained, which is smaller than the plastic dissipation and corresponds to the grey area on Fig. 5 (left).

The above derivation is the embedded-discontinuity softening-rigid-plasticity counterpart of the concepts introduced above in section “Explicit formulation—version 1”. If one wants to exploit the concepts from “Explicit formulation—version 2”, the rate of plastic dissipation has to be considered, which is defined as (since the plastic dissipation takes place only at the discontinuity curve)

$$\dot{D} = \int_{\Gamma} \left(\dot{\boldsymbol{\alpha}}^T \boldsymbol{t} + \dot{\xi}_s q_s \right) d\Gamma \tag{45}$$

where $q_s = -K_s \xi_s$. It is straightforward to show that the constraint equations related to (45) is (41). The corresponding implicit formulation will not be considered here.

When dealing with frames with elastic bulk and softening plastic hinges, the integral over the discontinuity curve in the expressions above is replaced with the sum over discontinuity points (i.e. the sum over softening plastic hinges). For example, in such a case, (35) transforms to

$$S = \sum_{i=1}^{n_p} \boldsymbol{\alpha}_i^T \boldsymbol{t}_i \tag{46}$$

where n_p is number of discontinuity points in the frame, and $\boldsymbol{\alpha}_i$ and $\boldsymbol{t}_i$ are vectors comprising jumps in displacements and rotations and cohesive forces and moments, respectively at i-th softening plastic hinge.

5 Numerical Examples

In this section, we present two examples. The following finite elements are used: (i) 4-node assumed natural strain (ANS) shell element with stress-resultant Ilyushin-Shapiro elasto-plasticity (which is a shell stress-resultant counterpart of shell J_2 plasticity with hardening and von Mises yield criterion), see [5], (ii) stress-resultant planar Euler-Bernoulli beam element with embedded-discontinuity in rotation representing softening plastic hinge. The used material models are elasto-plasticity with hardening (for the bulk) and rigid-plasticity with softening (for the discontinuity point), see Fig. 1 and [2, 9]. Those elements and plastic dissipation based path-following method have been implemented into the computer code AceFEM, see [11]

The path-following method that is used for the analysis of the first example is based on Eq. (15) with $\boldsymbol{f}_n^*$ computed with (18). For the initial, i.e. elastic, part of the response, standard cylindrical Crisfield arc-length method was used, see e.g. [1], with

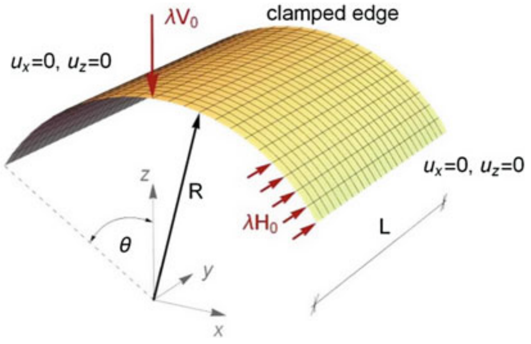
$$g_{n+1} = \Delta \boldsymbol{u}_n^T \Delta \boldsymbol{u}_n - \tau_n \tag{47}$$

see also [16]. The path-following method that is used for the analysis of the second example is based on the sum of two above presented equations: (i) equation (15) that takes into account plastic dissipation due to hardening plasticity ($\boldsymbol{f}_n^*$ in (15) was computed with (18)), and (ii) Eq. (41) that takes into account plastic dissipation due to softening ($\bar{\boldsymbol{f}}_n^*$ in (41) was computed with (44) and (46)).

5.1 Geometrically Nonlinear Elasto-plastic Shell Analysis

We consider cylindrical panel from Fig. 6, which is subjected to a set of horizontal axial forces λH_0 (where $H_0 = 1000\,\text{N}$), applied at each node of the mesh at curved edge at $y = 0$, and to a vertical point load λV_0 (where $V_0 = 10\,\text{N}$). Geometry and boundary conditions of the panel with thickness h are presented in Fig. 6. The panel is made of isotropic elasto-plastic material (steel) with elastic modulus $E = 210{,}000\,\text{N/mm}^2$, Poisson's coefficient $\boldsymbol{\nu} = 0.3$, yield stress $\boldsymbol{\sigma}_y = 235\,\text{N/mm}^2$ and hardening modulus $K_h = 0$. Finite element mesh consists of 24 × 24 elements. Table 1 presents the input data for used path-following methods. The analysis started with standard arc-length, which was later replaced with the dissipation controlled path-following method.

Figure 7 shows initial and deformed finite element meshes. Load factor λ versus vertical displacement curves for the two nodes, marked on Fig. 7 (left), are presented in Fig. 8. At point A on Fig. 8, the Crisfield cylindrical arc-length, see (47) and Table 1, failed to converge. If the solution method was switched to the dissipation



$R =$ 1270 mm
$L =$ 2032 mm
$\theta = 57.30°$
h = 5 mm

Fig. 6 Shell panel data

Table 1 Elasto-plastic shell: data for used path-following methods

	τ_0	$\tau_{n,max}$	Desired number of iterations	Convergence tolerance
Arc-length	0.5	1	8	10^{-8}
Dissipation based	10^2 Nmm	10^5 Nmm	8	10^{-8}

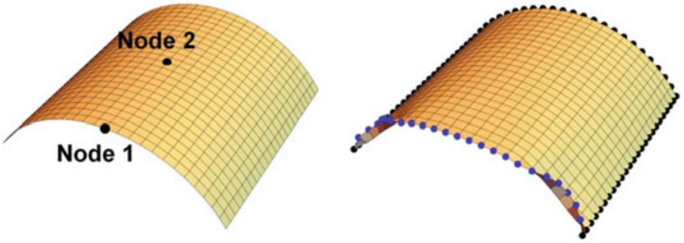


Fig. 7 Initial mesh and deformed mesh (at the end of the analysis, see Fig. 8)

based path-following method after the plasticity had started (which was before point A), the solution path could be traced much beyond point A. This example illustrates that the dissipation based path-following method may be superior to the cylindrical arc-length for elasto-plastic problems.

5.2 *Failure of Steel Frame*

A planar steel frame from Fig. 9 is analysed with the stress-resultant elasto-plastic geometrically linear beam Euler-Bernoulli finite element with the embedded strong

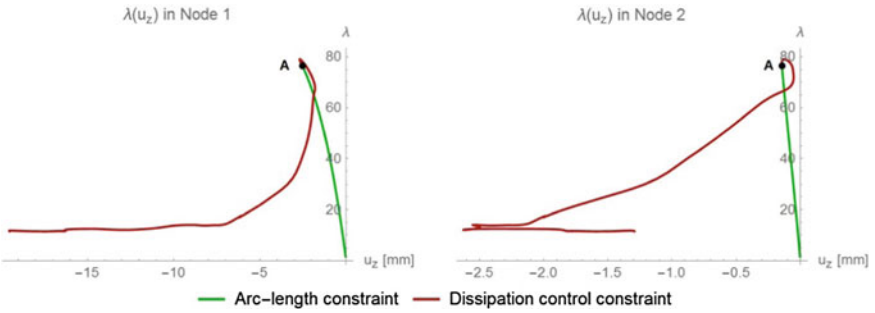


Fig. 8 Load factor versus vertical displacement u_z for nodes 1 and 2

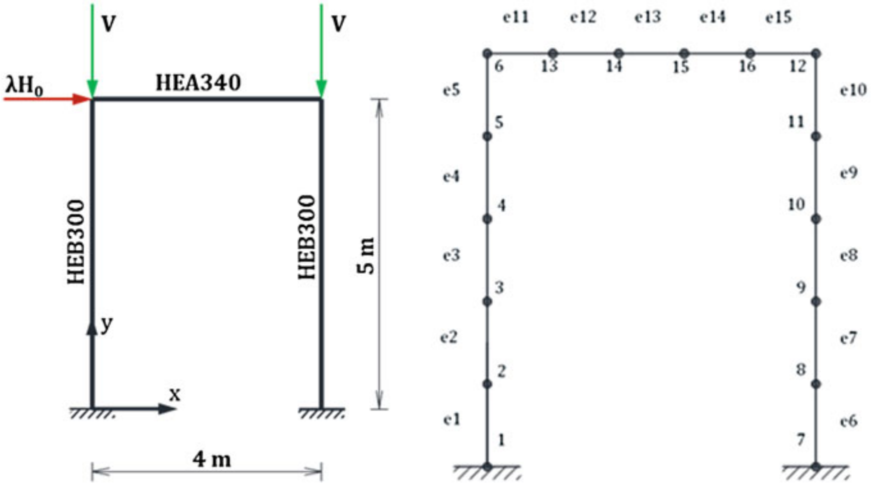


Fig. 9 Planar steel frame and the finite element mesh

discontinuity in rotation, see e.g. [2] and Fig. 1. The columns and the beam are discretized as shown on Fig. 9 (right). The material Young's modulus is $E = 210{,}000\,\text{N/mm}^2$ and the yield stress is $\boldsymbol{\sigma}_y = 235\,\text{N/mm}^2$. It is assumed that the yield moment of a cross-section depends on the axial force N as

$$M_y(N) = W\left(\sigma_y - \frac{|N|}{A}\right) \qquad (48)$$

where W is the bending resistance cross-section modulus, and A is the cross-section area. Moreover, it is assumed that the ultimate moment M_u is also a function of the axial force N as

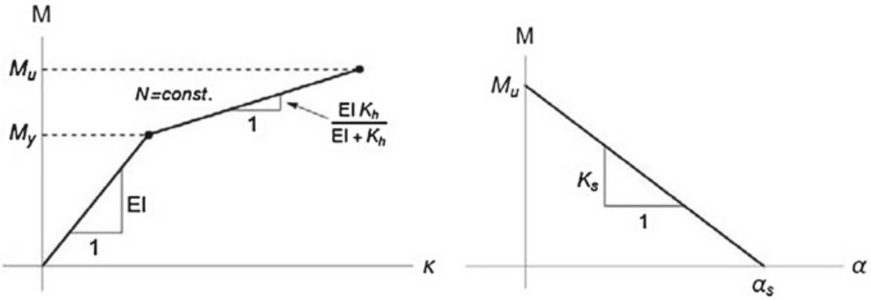


Fig. 10 Moment-curvature relation for the integration point. Moment versus jump in rotation line for the plastic hinge

$$M_u\left(N\right)=\begin{cases} M_u^{ref,0}\left(1.03+0.85\frac{N}{N_y}\right) & if \quad N<-0.035N_y \\ M_u^{ref,0} & if \quad N\geq -0.035N_y \end{cases} \tag{49}$$

where $N_y = A\sigma_y$ and $M_u^{ref,0} = W_{pl}\sigma_y$, where W_{pl} is the plastic modulus. The stress-resultant elasto-plasticity for the bulk is presented in Fig. 10 (left). The rigid-plasticity with softening, used in rotational softening plastic hinge, is presented in Fig. 10 (right). It is assumed that softening rotational hinge response is governed by linear softening modulus K_s.

The data for the HEA340 are: $A = 12721.5\,\text{mm}^2$, modulus of inertia $I = 264{,}213{,}316\,\text{mm}^4$, $W_{pl} = 1{,}761{,}321\,\text{mm}^3$, linear hardening modulus $K_h = 5.3{\cdot}10^{11}\,\text{Nmm}^2$, $K_s = -2.0{\cdot}10^9\,\text{Nmm}$. The data for the HEB300 are: $A = 14{,}282\,\text{mm}^2$, $I = 241{,}867{,}801\,\text{mm}^4$, $W_{pl} = 1{,}780{,}471\,\text{mm}^3$, $K_h = 6.3{\cdot}10^{11}\,\text{Nmm}^2$ and $K_s = -2.0{\cdot}10^9\,\text{Nmm}$. The load consists of the horizontal force λH_0, where $H_0 = 35\,\text{kN}$ and two vertical forces $V = 2800\,\text{kN}$ that remain constant throughout the analysis.

The standard arc-length method (see Table 2) failed to converge at point A on Fig. 11. If we replaced it by the path-following method with dissipation control (see Table 2) after the activation of the first plastic hinge, complete failure was computed. Figure 12 (left) shows deformed configuration at point B marked on Fig. 11. The value of plastic rotation at softening plastic hinges at that configuration are presented in Fig. 12 (right).

Table 2 Planar steel frame: data for used path-following methods

	τ_0	$\tau_{n,max}$	Desired number of iterations	Convergence tolerance
Arc-length	500	500	5	10^{-10}
Dissipation based	1 Nmm	10^5 Nmm	5	10^{-10}

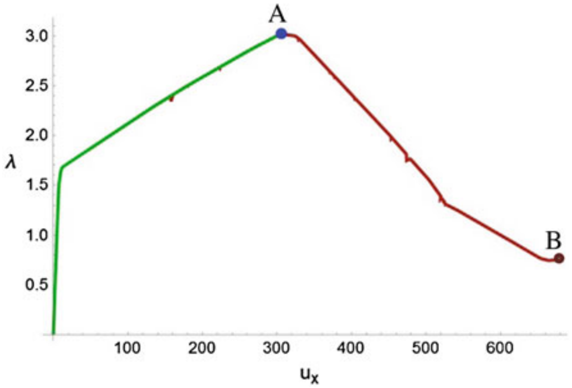


Fig. 11 Load factor versus horizontal displacement of the upper left corner of the frame

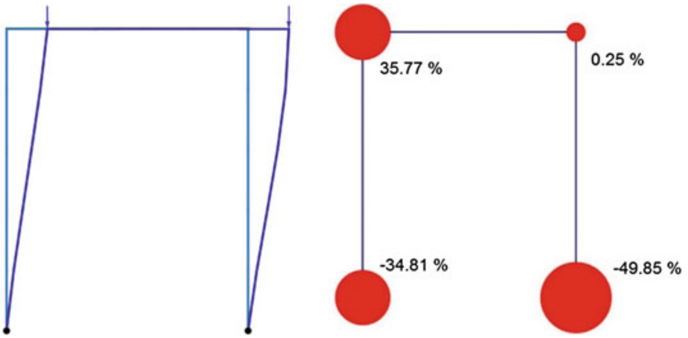


Fig. 12 Deformed configuration. Plastic hinges and corresponding ratio of plastic rotation α/α_s (in percentage)

6 Conclusions

In the first part of this work, we studied in detail different constraint functions for controlling incremental plastic dissipation in geometrically nonlinear elasto-plastic solid and structural problems. The derived constraint functions can be used to govern a dissipation-based path-following method for elasto-plasticity with isotropic hardening, which is an extension of the work presented in [17]. It turned out (see 1st example in "Numerical examples") that the resulting path-following method can be superior to the standard cylindrical Crisfiled's arc-length method [1]. Moreover, one should have in mind that the latter method sometimes allows for unrealistic, spurious elastic unloading of a complete structure [15]. This cannot happen with the dissipation-based path-following method, since elastic unloading of complete structure is not possible when using this method.

In the second part of the work, the dissipation-based path-following method was extended to embedded-discontinuity finite element formulations. Those formulations are used to model material failure in solids and structures. The constraint functions were derived for 2d-solid and plane-frame embedded-discontinuity finite elements that represent cohesive stresses (or forces and moments) in the discontinuity by rigid-plasticity with softening. It turned out that the dissipation-based path-following method is very suitable for computation of complete failure of solids and structures by using embedded-discontinuity finite elements (see 2nd example in "Numerical examples"). It should be also very robust for any other finite element formulation involving material softening.

References

1. Crisfield, M.A. 1991. Non-linear Finite Element Analysis of Solids and Structures, Vol. 1: Essentials, Chichester. John Wiley & Sons: 345 p.
2. Dujc, J., Brank, B., Ibrahimbegovic, A. 2010. Multi-scale computational model for failure analysis of metal frames that includes softening and local buckling. The fractional Fourier transform and applications. Computer Methods in Applied Mechanics and Engineering, 199: 1371–1385.
3. Dujc, J., Brank, B., Ibrahimbegovic, A., Brancherie, D. 2010. An embedded crack model for failure analysis of concrete solids. Computers and Concrete, 7 (4): 331–346.
4. Dujc, J., Brank, B., Ibrahimbegovic, A. 2010. Quadrilateral finite element with embedded strong discontinuity for failure analysis of solids. CMES - Computer Modeling in Engineering and Sciences, 69 (3): 223–258.
5. Dujc, J., Brank, B. 2012. Stress resultant plasticity for shells revisited. Computer Methods in Applied Mechanics and Engineering, 247–248: 146–165.
6. Dujc, J., Brank, B., Ibrahimbegovic, A. 2013. Stress-hybrid quadrilateral finite element with embedded strong discontinuity for failure analysis of plane stress solids. International Journal for Numerical Methods in Engineering, 94 (12): 1075–1098.
7. Geers, M.G.D. 1999. Enhanced solution control for physically and geometrically non-linear problems. Part I - The subplane control approach. International Journal for Numerical Methods in Engineering, 46 (2): 177–204.
8. A. Ibrahimbegovic. 2009. Nonlinear solid mechanics. Theoretical formulations and finite element solution methods. Springer: 594 p.
9. Jukić, M., Brank, B., Ibrahimbegović, A. 2013. Embedded discontinuity finite element formulation for failure analysis of planar reinforced concrete beams and frames. Engineering Structures, 50: 115–125.
10. Jukić, M., Brank, B., Ibrahimbegovic, A. 2014. Failure analysis of reinforced concrete frames by beam finite element that combines damage, plasticity and embedded discontinuity. Engineering Structures, 75: 507–527.
11. Korelc, J. 2015. AceGen and AceFEM. Available at http://www.fgg.uni-lj.si/Symech
12. Linder, C., Armero, F. 2007. Finite elements with embedded strong discontinuities for the modeling of failure in solids. International Journal for Numerical Methods in Engineering, 72 (12), pp. 1391–1433.
13. Oliver, J., Huespe, A.E. 2004. Theoretical and computational issues in modelling material failure in strong discontinuity scenarios. Computer Methods in Applied Mechanics and Engineering, 193 (27–29): 2987–3014.
14. Piculin, S., Brank, B. 2015. Weak coupling of shell and beam computational models for failure analysis of steel frames. Finite Elements in Analysis and Design, 97: 20–42.
15. Pohl, T., Ramm, E., Bischoff, M. 2014. Adaptive path following schemes for problems with softening. Finite Elements in Analysis and Design, 86: 12–22.

16. Stanić, A., Brank, B., Korelc, J. 2015. On consistently linearized path-following method and its application to structural failure problems. Submitted for publication.
17. Verhoosel, C.V., Remmers, J.J.C., Gutiérrez, M.A. 2009. A dissipation-based arc-length method for robust simulation of brittle and ductile failure. International Journal for Numerical Methods in Engineering, 77 (9): 1290–1321.
18. Wriggers, P. 2008. Nonlinear Finite Element Methods. Springer.

Improved Implicit Immersed Boundary Method via Operator Splitting

Shang-Gui Cai, Abdellatif Ouahsine, Julien Favier and Yannick Hoarau

Abstract We present an implicit immersed boundary method via operator splitting technique for simulating fluid flow over moving solid with complex shape. An additional moving force equation is derived in order to impose the interface velocity condition exactly on the immersed surface. The moving force matrix is formulated to be symmetric and positive definite, thus its calculation is computational inexpensive by using the conjugate gradient method. Moreover, the proposed immersed boundary method is incorporated into the rotational incremental projection method as a plug-in. No numerical boundary layers will be generated towards the velocity and pressure during the calculation. The method is validated through various benchmark tests.

1 Introduction

The immersed boundary method has gained popularity in recent years for its simplicity and efficiency in simulating flows with complex moving geometries. Traditional arbitrary Lagrangian-Eulerian (ALE) formulation [16] builds the mesh that conforms

S.-G. Cai · A. Ouahsine (✉)
Sorbonne Universités, Université de Technologie de Compiègne, CNRS,
UMR 7337 Roberval, Centre de Recherche Royallieu,
60319, 60203 Compiègne Cedex, France
e-mail: shanggui.cai@utc.fr

A. Ouahsine
e-mail: ouahsine@utc.fr

J. Favier
Aix-Marseille Université, CNRS, Centrale Marseille,
M2P2 UMR 7340,
13451 Marseille, France

Y. Hoarau
ICUBE, Strasbourg University,
UMR 7357, 2 Rue Boussingault,
67000 Strasbourg, France

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_3

the solid body. For moving geometries, the mesh is deformed or re-established at each time step. It is always important, but usually difficult, to maintain the quality of the mesh in ALE formulation. Meshless method, for example the natural element method [4], removes the issues of excessive mesh distortion. But it also has problems of their own. The immersed boundary method circumvents the mesh problems by employing a non body conforming mesh and adopting a boundary force to represent the immersed solid.

The immersed boundary method is first introduced by Peskin [12] for studying blood flow through a elastic beating heart. Lai and Peskin [10] extended the method to apply to rigid boundary problems by using a stiff spring. Goldstein et al. [7] and Saiki and Biringen [15] used the feedback forcing strategy. However, the time step is severely limited. Later Mohd-Yosuf [11] and Fadlun et al. [6] proposed the direct forcing immersed boundary method to enlarge the time step, by modifying the discretized momentum equation to set the desired interface velocity value on the immersed surface. To circumvent the oscillation towards the boundary force in case of moving boundaries, Uhlmann [20] proposed to calculate the boundary force on the Lagrangian positions and then spread the force to the surrounding fluid cells using a regularized delta function.

The crucial aspects of the immersed boundary method are the correct imposition of the interface velocity and the accurate evaluation of the boundary force. However, the boundary force depends on the fluid velocity, which is not known a priori.

Therefore, aforementioned methods apply an explicit scheme to evaluate the boundary force, which results in an inexact force between the fluid and solid and deteriorates the no-slip wall condition. Kempe and Fröhlich [9] improved the accuracy of the method of Uhlmann [20] by adding an additional forcing loop between the viscous and projection step. However, the error will not vanish within a small number of iteration. To achieve a desired tolerance, many more iterations are needed, which is in general time-consuming.

To reduce the error and achieve a fully implicit result, iterative schemes can be employed [1, 2], where the pressure Poisson equation is solved with only one iteration during the sub-iterations to spare the computational time. Considering the boundary force as a Lagrange multiplier for satisfying the interface velocity condition, Taira and Colonius [17] proposed the immersed boundary projection method, where the fluid equations together with the constraints are formulated into one algebraic system. The overall system is solved by an inexact factorization. The boundary force and the pressure are combined into a modified Poisson equation and then solved in the projection step. The divergence free condition and the interface velocity condition are satisfied simultaneously in the immersed boundary projection method but at the cost of increasing system dimension. Especially in case of moving boundaries, the matrix of the modified Poisson equation has to be updated and the pre-conditioner needs to be recalculated at each time level, which makes the method less efficient.

In the present paper, we propose an efficient implicit immersed boundary method by using operator splitting. An additional moving force equation is derived for imposing the interface velocity condition. The moving force matrix is formulated to be symmetric and positive definite, thus conjugate gradient method can be applied directly.

The overall scheme is performed in a very efficient fractional manner, i.e., the viscous prediction step, the immersed boundary forcing step and the projection step. Hence the fluid and solid are separated both from the mesh and the matrix. Furthermore, the present immersed boundary solver can be plugged into any existing fluid code easily. To illustrate this, we integrate the proposed immersed boundary method into the rotational incremental pressure-correction projection method, which does not generate numerical boundary layers on the velocity and pressure during the calculation [8].

The organization of the paper is as follows. The proposed immersed boundary method is presented in details in Sect. 2. Various numerical simulations are performed in Sect. 3 in order to validate the proposed method. We draw the conclusions in the final section.

2 Methodology

The dimensionalised incompressible fluid Navier-Stokes equations read

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla \cdot (\boldsymbol{u} \otimes \boldsymbol{u}) = -\nabla p + \frac{1}{Re}\nabla^2 \boldsymbol{u} + \boldsymbol{f}, \tag{1a}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{1b}$$

$$\boldsymbol{u} = \boldsymbol{U}_s \quad \text{on } \partial\Omega_s, \tag{1c}$$

where $\boldsymbol{u}$ is the fluid velocity vector, p the pressure, Re the Reynolds number and $\boldsymbol{U}_s$ the solid velocity vector. The Reynolds number is defined as $Re = UL/\nu$, where U, L and ν are the reference length, reference velocity and kinematic viscosity, respectively. Here we assume appropriate initial and boundary conditions are assigned to the Navier-Stokes equations. The immersed boundary force $\boldsymbol{f}$ here is used to satisfy the interface velocity condition (1c). We henceforth designate upper case letters and lower case letters for the quantities on the Lagrangian and Eulerian locations, respectively.

The fluid equations are discretized in time with the explicit second order Adams-Bashforth scheme for the non-linear terms and the implicit Crank-Nicolson scheme for the linear terms, yielding

$$\frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t} + \left[\frac{3}{2}\mathscr{N}(\boldsymbol{u}^n) - \frac{1}{2}\mathscr{N}(\boldsymbol{u}^{n-1})\right] = -\mathscr{G}p^{n+1} + \frac{1}{2Re}\mathscr{L}(\boldsymbol{u}^{n+1} + \boldsymbol{u}^n) + \boldsymbol{f}^{n+1}, \tag{2a}$$

$$\mathscr{D}\boldsymbol{u}^{n+1} = 0, \tag{2b}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{U}_s^{n+1} \quad \text{on } \partial\Omega_s^{n+1}, \tag{2c}$$

where $\mathscr{L}$, $\mathscr{N}$, $\mathscr{G}$, $\mathscr{D}$, are the discrete linear, non-linear, gradient and divergence operators, respectively. The pressure and boundary force are treated implicitly as a matter of fact that they are the Lagrange multipliers for satisfying the divergence free and the interface velocity condition at each time step. The semi-discrete equations (discrete in time but continuous in space) can be solved by the finite difference method, finite volume method, finite element method, or other spatial discretization methods.

The boundary force has a physical meaning that it represents the interaction between the fluid and solid. It is an unknown that needs to be solved along with the velocity field. How it is evaluated differs one immersed boundary method from another and determines the accuracy of the overall scheme. Next we review some popular immersed boundary methods for rigid boundaries in the literature and propose the novel method.

2.1 Overview of the Immersed Boundary Methods

The immersed boundary method of Peskin [12] was originally developed for elastic membranes. For rigid bodies, Lai and Peskin [10] used a spring with a large stiffness value κ to fix the boundary point $\boldsymbol{X}(s,t)$ on the equilibrium position $\boldsymbol{X}^e(s,t)$. The boundary force expression is given by

$$\boldsymbol{F}(s,t) = \kappa(\boldsymbol{X}^e(s,t) - \boldsymbol{X}(s,t)). \tag{3}$$

where s is the Lagrangian coordinate of the immersed boundary, irrespective to the underlying fluid Eulerian grids. Generally the fluid and solid grids are not coincident, especially when the staggered grid is used. The boundary force $\boldsymbol{F}(s,t)$ is evaluated on the Lagrangian locations and distributed to the fluid Eulerian grids using a regularized delta function

$$\boldsymbol{f}(\boldsymbol{x},t) = \int_s \boldsymbol{F}(s,t)\delta_h(\boldsymbol{x} - \boldsymbol{X}(s,t))ds \tag{4}$$

where the one-dimensional function has the form of

$$\delta_h(r) = \begin{cases} \dfrac{1}{8h}(3 - 2|r|/h + \sqrt{1 + 4|r|/h - 4(r/h)^2}), & |r|/h \leqslant 1, \\ \dfrac{1}{8h}(5 - 2|r|/h - \sqrt{-7 + 12|r|/h - 4(r/h)^2}), & 1 \leqslant |r|/h \leqslant 2, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where h is the cell width of the staggered grid in the r-direction. The one-dimensional function is plotted in Fig. 1 with a unit cell width. The time step is kept small in order to ensure the maximum displacement of any boundary point is negligible. Goldstein et al. [7] and Saiki and Biringen [15] used an alternative feedback forcing strategy

$$\boldsymbol{F}(s,t) = -\alpha \int_0^t \left(\boldsymbol{U}(s,t') - \boldsymbol{U}_s(s,t')\right) dt' - \beta \left(\boldsymbol{U}(s,t) - \boldsymbol{U}_s(s,t')\right), \tag{6}$$

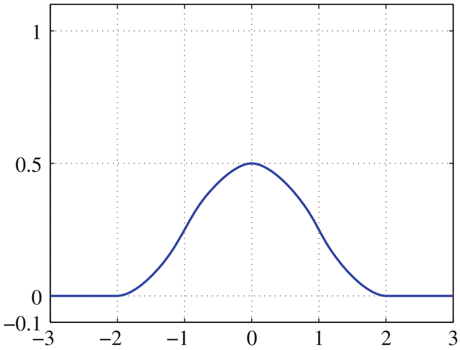


Fig. 1 The one-dimensional function of the regularized delta function of Peskin [13]

where $\alpha \gg 1$ and $\beta \gg 1$ are some large artificial constants. This approach behaves as a system of springs and dampers to correct $\boldsymbol{U}(s,t)$ to $\boldsymbol{U}_s(s,t)$ in a feedback manner. The major shortcomings of the feedback forcing approach are that big values of α and β can cause a stiff equation and an over small time step is allowed [6, 7].

To avoid those artificial parameters, Mohd-Yosuf [11] and Fadlun et al. [6] proposed the direct forcing immersed boundary method by replacing $\boldsymbol{u}^{n+1}$ with $\boldsymbol{U}_s$ in the momentum equation. In case of moving boundaries, Uhlmann [20] suggested to evaluate the boundary force on the Lagrangian locations and spread it to the Eulerian cells with the regularized delta function, in order to avoid large oscillation towards the force. The evaluation of the boundary force follows

$$\boldsymbol{F}(s,t) = \frac{\boldsymbol{U}_s - \boldsymbol{U}(s,t)}{\Delta t} \quad \text{on } \partial\Omega_s, \tag{7}$$

where

$$\boldsymbol{U}(s,t) = \int_{\boldsymbol{x}} \boldsymbol{u}(\boldsymbol{x},t)\delta_h(\boldsymbol{x} - \boldsymbol{X}(s,t))d\boldsymbol{x}. \tag{8}$$

As we can see that the boundary force is a function of the fluid velocity in the direct forcing immersed boundary method, which unfortunately is not known a priori. A simple way to handle this problem is to use the past or predicted velocity fields. The explicit method is quick but leaves the boundary force inexact. As a consequence, the final velocity will not satisfy the interface velocity condition. Kempe and Fröhlich [9] added a small number of forcing loop to improve this accuracy, but the error is not substantially changed.

To achieve a fully implicit scheme, Taira and Colonius [17] proposed the immersed boundary projection method by combining the two Lagrange multipliers, namely the boundary force and the pressure, into one algebraic system

$$\begin{bmatrix} \mathscr{A} & \mathscr{Q} \\ \mathscr{Q}^{\mathrm{T}} & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{u}^{n+1} \\ \lambda \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \tag{9}$$

where $\lambda = [p \quad \boldsymbol{F}^{n+1}]^{\mathrm{T}}$, $\mathscr{Q} = [\mathscr{G}, \mathscr{S}^{\mathrm{T}}]$, $\mathscr{A} = \frac{1}{\Delta t}[I - \frac{\Delta t}{2Re}\mathscr{L}]$ and r_1, r_2 include the boundary conditions, explicit terms and the interface velocity value. This system is solved by an approximate block LU decomposition

$$\begin{bmatrix} \mathscr{A} & 0 \\ \mathscr{Q}^{\mathrm{T}} & -\mathscr{Q}^{\mathrm{T}}\mathscr{B}^N\mathscr{Q} \end{bmatrix} \begin{bmatrix} I & \mathscr{B}^N\mathscr{Q} \\ 0 & I \end{bmatrix} \begin{pmatrix} \boldsymbol{u}^{n+1} \\ \lambda \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} + \begin{pmatrix} -\frac{\Delta t^N}{2^N}\mathscr{L}^N\mathscr{Q}\lambda \\ 0 \end{pmatrix}, \tag{10}$$

where $\mathscr{B}^N$ is the N-th order Taylor series expansion of $\mathscr{A}^{-1}$

$$\mathscr{B}^N = \Delta t I + \frac{\Delta t^2}{2Re}\mathscr{L} + \frac{\Delta t^3}{(2Re)^2}\mathscr{L}^2 + \cdots + \frac{\Delta t^N}{(2Re)^{N-1}}\mathscr{L}^{N-1} = \sum_{j=1}^{N} \frac{\Delta t^j}{(2Re)^{j-1}}\mathscr{L}^{j-1}. \tag{11}$$

In practice, this algebraic equation is solved as follows

$$\mathscr{A}\boldsymbol{u}^* = r_1, \tag{12a}$$

$$\mathscr{Q}^{\mathrm{T}}\mathscr{B}^N\mathscr{Q}\lambda = \mathscr{Q}^T\boldsymbol{u}^* - r_2, \tag{12b}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^* - \mathscr{B}^N\mathscr{Q}\lambda. \tag{12c}$$

And $N = 3$ is suggested in [17] for achieving positive-definiteness of the modified Poisson equation (12b).

The immersed boundary projection method has successfully satisfied the divergence free condition and the interface velocity condition simultaneously. However, the fluid and solid are still coupled in terms of matrix at the projection procedure, which could make the method less efficient. The matrix of the modified Poisson equation has to be updated and its pre-conditioner needs to be re-computed when the boundary moves. It is known that the most time-consuming part in the projection method is solving the Poisson equation. Additionally, the dimension of the modified Poisson equation is increased, which requires more storage and iterations for a given tolerance.

2.2 Moving Immersed Boundary Method

In view of the advantages and disadvantages of the immersed boundary methods discussed previously, we propose the moving immersed boundary method. For the sake of simplicity, we rewrite the fluid equations as

$$\frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t} = \mathscr{H}(\boldsymbol{u}) + \mathscr{P}(\boldsymbol{u}) + \mathscr{F}(\boldsymbol{u}), \tag{13a}$$

$$\mathscr{D}\boldsymbol{u}^{n+1} = 0, \tag{13b}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{U}_s^{n+1} \quad \text{on } \partial\Omega_s^{n+1}, \tag{13c}$$

where

$$\mathscr{H}(\boldsymbol{u}) = -\left[\frac{3}{2}\mathscr{N}(\boldsymbol{u}^n) - \frac{1}{2}\mathscr{N}(\boldsymbol{u}^{n-1})\right] + \frac{1}{2Re}\mathscr{L}(\boldsymbol{u}^{n+1} + \boldsymbol{u}^n) - \mathscr{G}p^n, \tag{14a}$$

$$\mathscr{P}(\boldsymbol{u}) = -\mathscr{G}\phi^{n+1}, \tag{14b}$$

$$\mathscr{F}(\boldsymbol{u}) = \boldsymbol{f}^{n+1}, \tag{14c}$$

where ϕ is the pseudo pressure.

It is worth noting that the momentum Eq. (13a) is constrained by the divergence free condition (13b) and the interface velocity condition (13c). To decouple this, we perform the following operator splitting algorithm:

(I) Viscous prediction step, ignoring the immersed objects.

$$\frac{\hat{\boldsymbol{u}} - \boldsymbol{u}^n}{\Delta t} = \mathscr{H}(\boldsymbol{u}). \tag{15}$$

(II) Immersed boundary forcing step for satisfying the interface velocity condition.

$$\frac{\tilde{\boldsymbol{u}} - \hat{\boldsymbol{u}}}{\Delta t} = \mathscr{F}(\boldsymbol{u}), \tag{16a}$$

$$\tilde{\boldsymbol{u}} = \boldsymbol{U}_s^{n+1} \quad \text{on } \partial\Omega_s. \tag{16b}$$

Using the direct forcing concept and following Uhlmann [20], we evaluate the boundary force on the Lagrangian locations and spread it to the fluid Eulerian cells, namely

$$\mathscr{I}\mathscr{S}\boldsymbol{F}^{n+1} = \frac{\boldsymbol{U}_s - \mathscr{I}\hat{\boldsymbol{u}}}{\Delta t}, \tag{17a}$$

$$\tilde{\boldsymbol{u}} = \hat{\boldsymbol{u}} + \Delta t\mathscr{S}\boldsymbol{F}^{n+1}, \tag{17b}$$

where $\mathscr{I}$ is the interpolation operator matrix for transferring the quantities from the Eulerian cells to Lagrangian markers, $\mathscr{S}$ the spreading operator matrix in the opposite direction. Donate $\mathscr{M} = \mathscr{I}\mathscr{S}$ the moving force matrix, then we obtain a concise form of the moving force equation

$$\mathscr{M}\boldsymbol{F}^{n+1} = \frac{\boldsymbol{U}_s - \mathscr{I}\hat{\boldsymbol{u}}}{\Delta t}. \tag{18}$$

The moving force matrix $\mathscr{M}$ is found to be symmetric and positive-definite. More importantly, its dimension only depends on the number of Lagrangian markers on the immersed surface, which in general is much smaller than the dimension of fluid matrix. Therefore, compared to the modified Poisson equation in the immersed boundary projection method [17], the moving force equation is much easier to work with. Even though the interface velocity condition is enforced before the projection step, we have found that the velocity on the immersed boundary is essentially unchanged after the projection step. The same observation has also been made by Kempe and Fröhlich [9] and Fadlun et al. [6].

(III) Projection step for obtaining a divergence free velocity $\boldsymbol{u}^{n+1}$.

$$\frac{\boldsymbol{u}^{n+1} - \tilde{\boldsymbol{u}}}{\Delta t} = \mathscr{P}(\boldsymbol{u}), \tag{19a}$$

$$\nabla \cdot \boldsymbol{u}^{n+1} = 0. \tag{19b}$$

Applying the divergence operator to (19a) and using the divergence free condition (19b), we have

$$\mathscr{L}\phi^{n+1} = \frac{1}{\Delta t}\mathscr{D}\tilde{\boldsymbol{u}}, \tag{20a}$$

$$\boldsymbol{u}^{n+1} = \tilde{\boldsymbol{u}} - \Delta t \mathscr{G}\phi^{n+1}. \tag{20b}$$

The time splitting error due to the implicit treatment of viscous terms is found to be $\frac{1}{2Re}\mathscr{D}\hat{\boldsymbol{u}}$, by adding up those sub-steps and comparing to (2a). This error can introduce a numerical boundary layer to the pressure and velocity. The strength of this numerical boundary layer decreases when increasing the Reynolds number. But for low Reynolds number flows, the effects of the numerical boundary layer become quite severe. However it is advantageous to employ implicit schemes for the viscous terms for stability when the Reynolds number is small. Guermond et al. [8] indicated that the numerical boundary layer can be avoided by absorbing this error into the pressure, viz,

$$p^{n+1} = p^n + \phi^{n+1} - \frac{1}{2Re}\mathscr{D}\hat{\boldsymbol{u}}. \tag{21}$$

The consistent pressure boundary condition is also obtained by considering the identity $\nabla^2 \boldsymbol{u} = -\nabla \times \nabla \times \boldsymbol{u} + \nabla(\nabla \cdot \boldsymbol{u})$. This type of fractional step method for solving the Navier-Stokes equations is termed as the rotational incremental pressure-correction projection method in [8].

3 Numerical Results

3.1 Flow over a Stationary Circular Cylinder

First we consider the uniform flow over a stationary circular cylinder to validate our proposed method, since abundant experimental and numerical results are available. The flow pattern exhibits differently according to the Reynolds number $Re = u_\infty D/\nu$ based on the free stream velocity u_∞ and the cylinder diameter D.

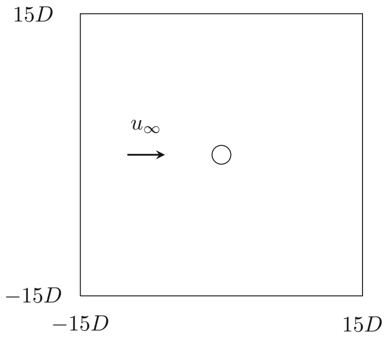


Fig. 2 The computational domain for the flow over a stationary cylinder problem (the diameter of the cylinder is scaled to make it more visible)

Both the steady case $Re = 20$ and the unsteady case $Re = 185$ have been studied here with our proposed method.

The schematic view of the fluid domain is depicted in Fig. 2. The cylinder with a unit diameter $D = 1$ is centered at the origin of the computational domain of $[-15D, 15D] \times [-15D, 15D]$. The mesh is distributed uniformly with a resolution of around $0.03D$ for both steady and unsteady cases.

A uniform free-stream velocity $u_\infty = 1$ is prescribed at the inlet. Free slip boundary conditions are assigned to lateral boundaries. The convective outflow boundary condition $\partial \boldsymbol{u}/\partial t + u_\infty \partial \boldsymbol{u}/\partial x = 0$ is applied to the outlet. The time step is selected based on the CFL condition $\text{CFL} = u_{\max} \Delta t/h \leqslant 1$.

To compare with the results in the literature, we monitor the drag and lift coefficients in both cases which are computed by

$$C_D = \frac{F'_x}{\frac{1}{2}\rho u_\infty^2 D}, \tag{22}$$

$$C_L = \frac{F'_y}{\frac{1}{2}\rho u_\infty^2 D}, \tag{23}$$

where $\rho u_\infty^2 D = 1$ and the force on the cylinder, $\boldsymbol{F}'$, can be evaluated directly on the immersed surfaces by integrating the forces

$$\boldsymbol{F}' = \begin{pmatrix} F'_x \\ F'_y \end{pmatrix} = -\int_s \boldsymbol{F}(\boldsymbol{X}) ds. \tag{24}$$

At $Re = 20$, a steady recirculation zone is developed behind the cylinder. The characteristic wake dimensions are illustrated in Fig. 3, including the recirculation zone length l, the horizontal distance from the vortex center to the cylinder a, the vertical distance between the two vortex centers b and the separation angle θ. The

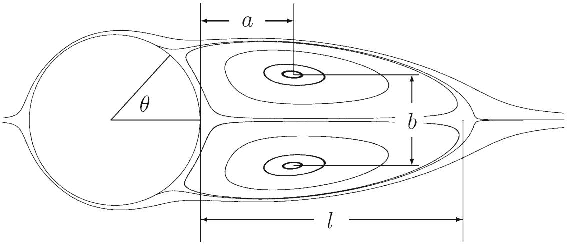


Fig. 3 The definition of the characteristic wake dimensions for the steady-state flow around a circular cylinder

Table 1 Comparison of wake dimensions and drag coefficient for the steady-state flow past a circular cylinder at $Re = 20$

	l/D	a/D	b/D	$\theta°$	C_D
Present	0.94	0.39	0.41	42.9	2.12
Taira and Colonius [17]	0.94	0.37	0.43	43.3	2.06
Coutanceau and Bouard [3]	0.93	0.33	0.46	45.0	–
Tritton [19]	–	–	–	–	2.09

wake dimensions together with the coefficients of drag and lift are presented in Table 1 and compared to the numerical results of the immersed boundary projection method of Taira and Colonius [17] and the experimental results from Tritton [19]

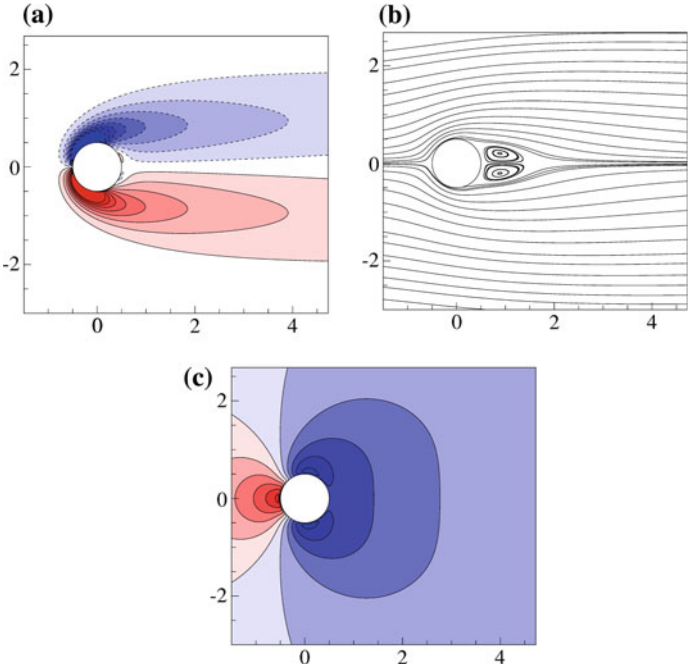


Fig. 4 Fields of the steady-state flow around a cylinder at $Re = 20$. **a** Vorticity contours, where *dashed lines* represent negative values; **b** streamlines; **c** pressure contours

and Coutanceau and Bouard [3]. The corresponding flow profiles are shown in Fig. 4. Good agreements have been found.

The flow becomes unsteady when increasing the Reynolds number to $Re = 185$. Periodic vortex shedding is observed from Fig. 7. The time evolution of the drag and lift coefficients are shown in Fig. 5. The Strouhal number is defined as $St = Df/u_\infty$, where f is the shedding frequency (see Fig. 6). Table 2 compares the Strouhal number, the mean drag coefficient and the r.m.s. lift coefficient against results from other immersed boundary methods [14, 18, 21] and the experimental data of Williamson [22]. Excellent agreements are obtained, especially for the Strouhal number.

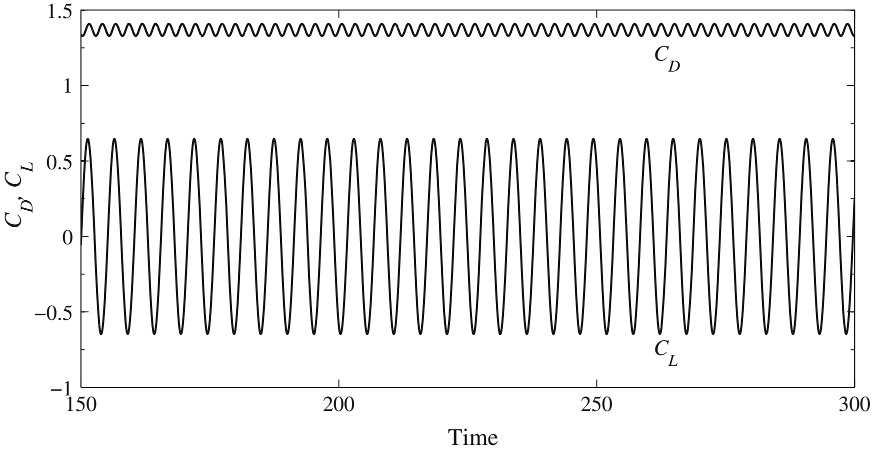
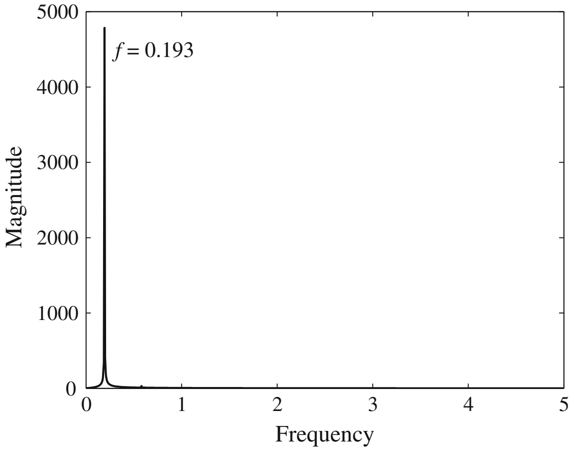


Fig. 5 Time history of the drag and lift coefficients for the flow over a cylinder at $Re = 185$

Fig. 6 Shedding frequency for the flow over a cylinder at $Re = 185$



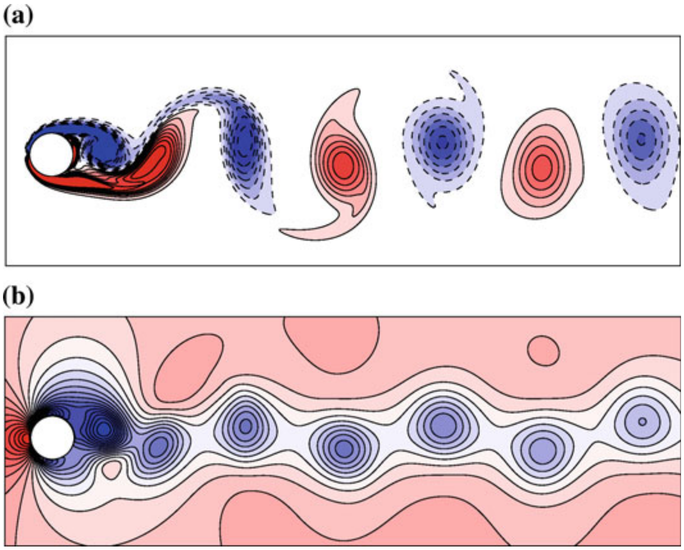


Fig. 7 Profiles for the unsteady-state flow around a cylinder at $Re = 185$. **a** Vorticity contours, where *dashed lines* represent negative values; **b** pressure contours

Table 2 Comparison of drag coefficient, lift coefficient and Strouhal number for the unsteady-state flow over a cylinder at $Re = 185$

	St	$\overline{C}_D$	C_L^{rms}
Present	0.193	1.369	0.456
Pinelli et al. [14]	0.196	1.430	0.423
Toja-Silva et al. [18]	0.195	1.31	–
Vanella and Balaras [21]	–	1.377	0.461
Williamson [22]	0.193	–	–

3.2 Flow over an Oscillating Circular Cylinder

To illustrate our proposed method for simulating fluid flow over moving boundary problems, we consider the flow induced by an oscillating circular cylinder in a fluid at rest, which is studied numerically and experimentally by Dütsch et al. [5].

The motion of the cylinder is given by a harmonic oscillation $x_c(t) = -A \sin (2\pi f t)$, where A is the oscillation amplitude and f the frequency. The moving force matrix is updated each time the boundary moves to a new position. The flow pattern is charactered by two main parameters, namely the Reynolds number $Re = U_{\max} D/\nu$ and the Keulegan-Carpenter number $KC = U_{\max}/f D$, where $U_{\max}$ designates the maximum velocity of the cylinder. The two main parameters are set to $Re = 100$ and $KC = 5$ according to Dütsch et al. [5].

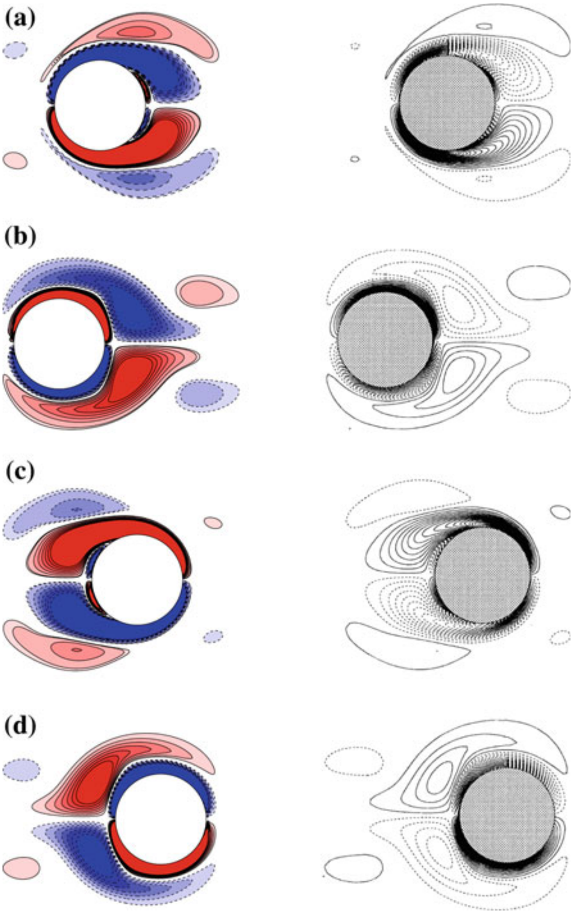


Fig. 8 Comparison of vorticity fields for the flow over an oscillating cylinder problem at $Re = 100$ and $KC = 5$ at four different positions: **a** 0° , **b** 96°, **c** 192°, **d** 288°. The present results are plotted in the *left* column and compared to the results of Dütsch et al. [5] in the *right* column

The left column of Fig. 8 shows the vorticity contours with present method from –3 to 3 with an incremental of 0.4 at four different phases (0°, 96°, 192°, 288°). The same structures are observed by comparing to the results of Dütsch et al. [5] in the right column. Figures 9, 10 and 11 plot the velocity profiles at three different phases 180°, 210°, 330°, respectively. At each phase, the two velocity components (u, v) are displayed at four different locations $(x/D = -6, 0, 6, 12)$ against the numerical and experimental results of Dütsch et al. [5]. The results with our proposed immersed boundary method are very close to the numerical data of Dütsch et al. [5] with the body conforming mesh method. Since our method treats all the domain as the fluid, the velocities inside the solid are not zero. But we only consider the exterior flow field. The present method gives a rather good accuracy, as the two curves from numerical solutions overlap outside the cylinder.

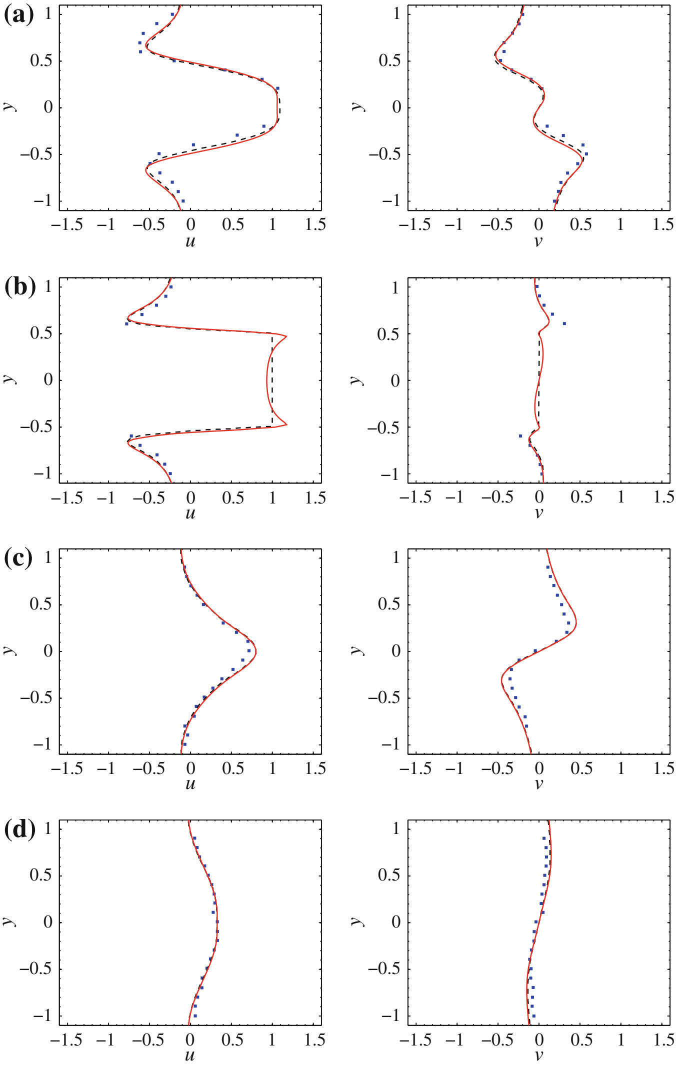


Fig. 9 Comparison of velocity components at the phase 180° at four cross-sections: **a** $x/D = -6$, **b** $x/D = 0$, **c** $x/D = 6$, **d** $x/D = 12$. *Solid lines* represent the results with present method. The *dashed lines* are the numerical solution of Dütsch et al. [5]. The *squares* are the experimental results of Dütsch et al. [5]

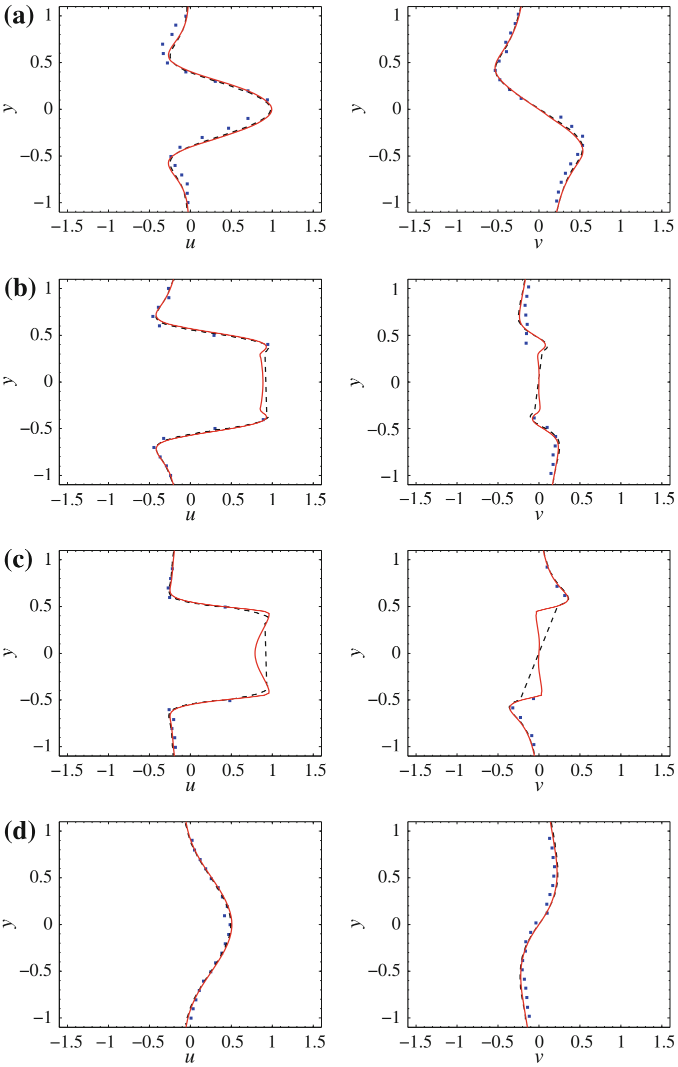


Fig. 10 Comparison of velocity components at the phase 210° at four cross-sections: **a** $x/D = -6$, **b** $x/D = 0$, **c** $x/D = 6$, **d** $x/D = 12$. *Solid lines* represent the results with present method. The *dashed lines* are the numerical solution of Dütsch et al. [5]. The *squares* are the experimental results of Dütsch et al. [5]

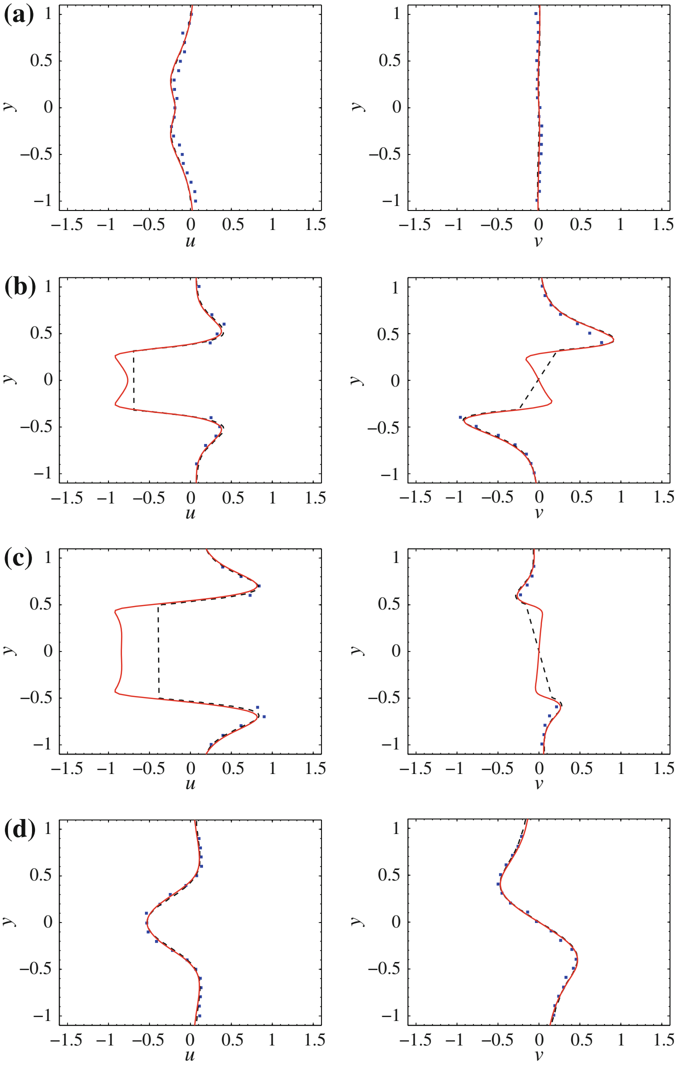


Fig. 11 Comparison of velocity components at the phase 330° at four cross-sections: **a** $x/D = -6$, **b** $x/D = 0$, **c** $x/D = 6$, **d** $x/D = 12$. *Solid lines* represent the results with present method. The *dashed lines* are the numerical solution of Dütsch et al. [5]. The *squares* are the experimental results of Dütsch et al. [5]

4 Conclusions

In this paper, a novel implicit formulation of the immersed boundary method for simulating incompressible fluid flow over complex stationary or moving boundaries was presented. The proposed immersed boundary method is based on the operator splitting technique that allows to separate the fluid from the solid with respect to the mesh and the matrix. The overall scheme is solved sequentially following the viscous prediction step, the immersed boundary forcing step and the projection step. In the forcing step, a moving force equation is derived to determine the boundary force implicitly. The moving force system is formulated to be symmetric and positive-definite and quickly solved with the conjugate gradient method. The proposed immersed boundary solver can be easily inserted into any fluid solver as a plug-in. The results have shown the efficiency and accuracy of the present method.

Acknowledgments The first author acknowledges the financial support of the China Scholarship Council.

References

1. Cai, S.-G., Ouahsine, A., Smaoui, H., Favier, J., Hoarau, Y.: An efficient implicit direct forcing immersed boundary method for incompressible flows. Journal of Physics: Conference Series. **574**, 012165 (2014)
2. Cai, S.-G., Ouahsine, A., Favier, J., Hoarau, Y., Smaoui, H.: Immersed boundary method for the simulation of lid-driven cavity flow with an embedded cylinder. COUPLED PROBLEMS 2015-Proceedings of the 6th International Conference on Coupled Problems in Science and Engineering. Venice, Italy, 1130–1137 (2015)
3. Coutanceau, M., Bouard, R.: Experimental determination of the main features of the viscous flow in the wake of a circular cylinder in uniform translation. Part 1. Steady flow. Journal of Fluid Mechanics. **79**, 231–256 (1977)
4. Darbani, M., Ouahsine, A., Villon, P., Naceur, H., Smaoui, H.: Meshless method for shallow water equations with free surface flow. Applied Mathematics and Computation. **217**, 5113–5124 (2011)
5. Dütsch, H., Durst, F., Becker, S., Lienhart, H.: Low-Reynolds-number flow around an oscillating circular cylinder at low Keulegan-Carpenter numbers. Journal of Fluid Mechanics. **360**, 249–271 (1998)
6. Fadlun, E.A., Verzicco, R., Orlandi, P., Mohd-Yusof, J.: Combined immersed-boundary finite-difference methods for three-dimensional complex flow simulations. Journal of Computational Physics. **161**, 35–60 (2000)
7. Goldstein, D., Handler, R., Sirovich, L.: Modeling a no-slip flow boundary with an external force field. Journal of Computational Physics. **105**, 354–366 (1993)
8. Guermond, J.L., Minev, P., Shen, J.: An overview of projection methods for incompressible flows. Computer Methods in Applied Mechanics and Engineering. **195**, 6011–6045 (2006)
9. Kempe, T., Fröhlich, J.: An improved immersed boundary method with direct forcing for the simulation of particle laden flows. Journal of Computational Physics. **231**, 3663–3684 (2012)
10. Lai, M.-C., Peskin, C.S.: An immersed boundary method with formal second-order accuracy and reduced numerical viscosity. Journal of Computational Physics. **160**, 705–719 (2000)
11. Mohd-Yosuf, J.: Combined immersed boundary/b-spline methods for simulation of flow in complex geometries. Annual Research Briefs, Center for Turbulence Research, 317–327 (1997)

12. Peskin, C.S.: Flow patterns around heart valves: A numerical method. Journal of Computational Physics, **10**, 252–271 (1972)
13. Peskin, C.S.: The immersed boundary method. Acta Numerica, **11**, 479–517 (2002)
14. Pinelli, A., Naqavi, I.Z., Piomelli, U., Favier, J.: Immersed-boundary methods for general finite-difference and finite-volume Navier-Stokes solvers. Journal of Computational Physics, **229**, 9073–9091 (2010)
15. Saiki, E.M., Biringen, S.: Numerical simulation of a cylinder in uniform flow: Application of a virtual boundary method. Journal of Computational Physics, **123**, 450–465 (1996)
16. Souli, M., Ouahsine, A., Lewin, L.: ALE formulation for fluid-structure interaction problems. Computer Methods in Applied Mechanics and Engineering, **190**, 659–675 (2000)
17. Taira, K., Colonius, T.: The immersed boundary method: A projection approach. Journal of Computational Physics, **225**, 2118–2137 (2007)
18. Toja-Silva F., Favier, J., Pinelli, A.: Radial basis function (RBF)-based interpolation and spreading for the immersed boundary method. Computers & Fluids, **105**, 66–75 (2014)
19. Tritton, D.J.: Experiments on the flow past a circular cylinder at low Reynolds numbers. Journal of Fluid Mechanics, **6**, 547–567 (1959)
20. Uhlmann, M.: An immersed boundary method with direct forcing for the simulation of particulate flows. Journal of Computational Physics, **209**, 448–476 (2005)
21. Vanella, M., Balaras, E.: A moving-least-squares reconstruction for embedded-boundary formulations. Journal of Computational Physics, **228**, 6617–6628 (2009)
22. Williamson, C.H.K.: Defining a universal and continuous Strouhal-Reynolds number relationship for the laminar vortex shedding of a circular cylinder. Physics of Fluids, **31**, 2742–2744 (1988)

Modelling Wave Energy Conversion of a Semi-submerged Heaving Cylinder

Shang-Gui Cai, Abdellatif Ouahsine and Philippe Sergent

Abstract In the current paper, a numerical model for simulating the ocean wave energy conversion of a semi-submerged heaving cylinder is presented. Contrary to the convectional potential flow theory, our solution is based on the full three-dimensional viscous Navier-Stokes equations. An efficient numerical wave tank is established to generate waves according to the wave theory. The coupling between the fluid equations with the rigid body dynamics are also taken into consideration in the present study.

1 Introduction

The ocean wave energy, as the most conspicuous form of ocean energy, is of considerable interest towards its conversion. Various types of wave energy converters (WEC) have been designed and tested in recent years, ranging from the simplest floating body oscillating against a fixed frame of reference, often termed as the point absorber, to more complicated systems such as the wave carpet [1], SEAREV [2], Pelamis, and self-rectifying air turbines [3], etc.

Numerous efforts have been devoted to WEC performance analysis with the potential flow theory, assuming the fluid to be inviscid, which is less true in reality. Nowadays numerical approaches with modern Computational Fluid Dynamics (CFD) techniques provide a powerful tool for solving the full viscous Navier-Stokes

S.-G. Cai · A. Ouahsine (✉)
Centre de Recherche Royallieu, Sorbonne Universités,
Université de Technologie de Compiègne, CNRS, UMR 7337, CS 60319,
60203 Roberval, Compiègne Cedex, France
e-mail: shanggui.cai@utc.fr

A. Ouahsine
e-mail: ouahsine@utc.fr

P. Sergent
Compiègne, France

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_4

equations, thus the viscous effects of boundary layer separation, turbulence, wave breaking and over-topping can be predicted.

Basically, to establish a numerical wave tank for studying the wave energy conversion, a free surface model and a wave generation technique are required. Instead of working with the shallow water equations [4], we focus on the full Navier-Stokes equations. In general, two types of free surface capturing approaches for Navier-Stokes equations can be utilized, i.e., the surface tracking method and the interface capturing method. The former treats the free surface as a sharp boundary evolving with time. However, this method is quite computational expensive. The interface capturing methods, such as the Marker-and-Cell (MAC) method, Volume of Fluid (VOF) method and level set approach, are very efficient since they are based on an Eulerian approach which do not adapt the mesh for representing the free surface. The VOF method is employed in the present study, since it is the most widely used model for describing the free surface, for example in the ship hydrodynamics [5]. The numerical wave can be generated either from an oscillating piston or appropriate boundary conditions. The oscillating piston is intentionally avoided in the present study, as it requires deforming boundary meshes each time. Instead, various wave theories can be employed for this purpose, including the linear Airy wave theory (Stokes' first order theory), Stokes' second order theory, Stokes' fifth order theory, etc.

Among a variety of wave energy converters, the semi-submerged heaving cylinder with one degree of freedom is considered in this study. At each time level, the cylinder oscillates vertically under the impact of incoming waves and affects the distribution of the surrounding fluids, resulting in a typical fluid-structure interaction problem. A fluid-structure interaction algorithm is then proposed in order to simulate the two-way interactions.

This paper is organized as follows. The mathematical descriptions of the physical problems are first presented both for the fluid and solid parts. Next we discuss the wave generation techniques and display the wave propagation according to the wave theory. In the following, the pressure-velocity decoupling method and the fluid-structure interaction algorithm are illustrated in details. With the proposed model, numerical results with respect to the energy extraction are then shown. The conclusions are drawn finally.

2 Problem Formulation

2.1 Fluid Equations

The viscous fluid flow is governed by the Navier-Stokes equations

$$\frac{\partial(\rho \boldsymbol{u})}{\partial t} + \nabla \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u}) = \nabla \cdot \boldsymbol{\sigma}_f + \rho \boldsymbol{g}, \tag{1}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0, \tag{2}$$

where ρ is the fluid density, $\boldsymbol{u}$ the fluid velocity vector and $\boldsymbol{g}$ the gravity vector. Under the Stokes assumption for a Newtonian fluid, the fluid stress $\boldsymbol{\sigma}_f$ is then given by

$$\boldsymbol{\sigma}_f = -p\boldsymbol{I} + \boldsymbol{\tau} = -(p + \lambda \nabla \cdot \boldsymbol{u})I + 2\mu \boldsymbol{\varepsilon}, \tag{3}$$

where p is the pressure, μ the dynamic viscosity, λ the bulk viscosity and $\boldsymbol{\varepsilon}$ is the strain tensor given by $(1/2)(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T)$. For an incompressible fluid (ρ = const), the governing equations become

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla \cdot (\boldsymbol{u} \otimes \boldsymbol{u}) = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \boldsymbol{u} + \boldsymbol{g}, \tag{4}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{5}$$

where $\nu = \mu / \rho$ is the kinematic viscosity.

To capture the free surface between the water and air, the volume of fluid (VOF) method [6] is applied by introducing a scalar field, namely the liquid volume fraction α, which is convected by

$$\frac{\partial \alpha}{\partial t} + \nabla \cdot (\boldsymbol{u} \alpha) = 0. \tag{6}$$

The multiphase flow interface can be found when $\alpha \in (0, 1)$. Generally $\alpha = 0$ represents the air and $\alpha = 1$ the water. For each cell, the fluid density is calculated by

$$\rho = (1 - \alpha)\rho_{air} + \alpha \rho_{water}. \tag{7}$$

The viscosity is determined in the same manner. Therefore, one single momentum equation is solved throughout the domain with the velocity fields shared by the two phases.

2.2 Rigid Body Equations of Motion

The wave energy converter considered in the present study is a semi-submerged cylinder with one-dimensional freedom [7, 8], usually termed as a point-absorber (see Fig. 1). The cylinder is constrained by a power take-off (PTO) device, whose motion is given by

$$m\ddot{Z} = m\dot{U} = F_{\text{wave}} + F_{\text{gravity}} + F_{\text{PTO}}, \tag{8}$$

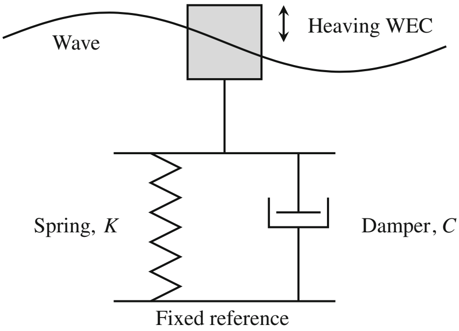


Fig. 1 The PTO device of a heaving *wave* energy converter

where m, Z, U are the mass, vertical displacement and vertical velocity of the cylinder, respectively. In the present study, the diameter and height of the cylinder are set to $D = 2m, L = 2m$. F_{wave} is the wave force on the cylinder, which can be obtained by integrating the pressure and viscous stresses on the solid surface. $F_{gravity}$ is the weight of cylinder and F_{PTO} is the force from the PTO device, which can be represented as a mass-spring-damping system

$$F_{\text{PTO}} = CU + KZ, \tag{9}$$

where C is the damping coefficient and K is the spring stiffness. The damping coefficient can be also expressed as

$$C = 2\xi\sqrt{Km}, \tag{10}$$

where ξ is the damping ratio. Therefore, the generated power from PTO is

$$P_{\text{PTO}} = CU^2. \tag{11}$$

2.3 Boundary Conditions and Wave Generation

Figure 2 displays the computational domain for the wave energy conversion problem. The symmetry boundary condition is applied to the lateral boundaries to simulate the open sea situation. And no-slip wall boundary condition is specified at the cylinder surface. The computational domain with the cylinder are discretized as shown in Fig. 3. In order to capture the free surface, 20 nodes have been used near the still water level.

The waves are generated at the inlet side and propagated towards the cylinder. The wave profile is illustrated in Fig. 4. We employ the Airy wave theory, assuming the wave to be linear with small amplitude. Other wave theories and their range of

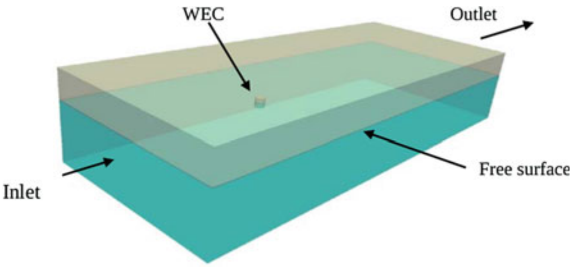


Fig. 2 Schematic representation of the computational domain

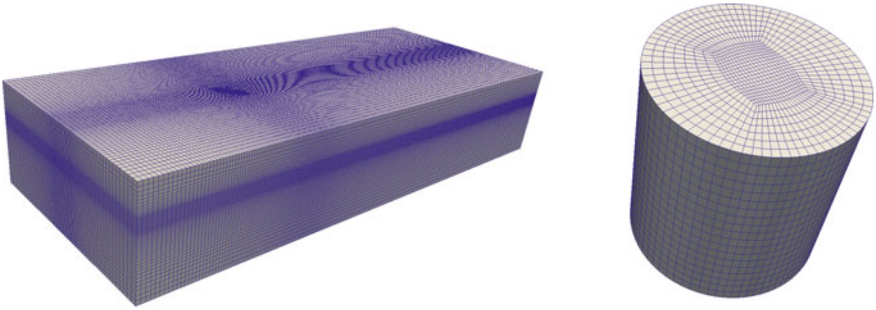
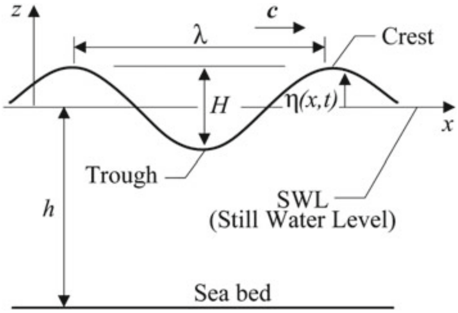


Fig. 3 Computational domain discretization and the surface mesh of the heaving cylinder

Fig. 4 Definition of progressive surface wave parameters

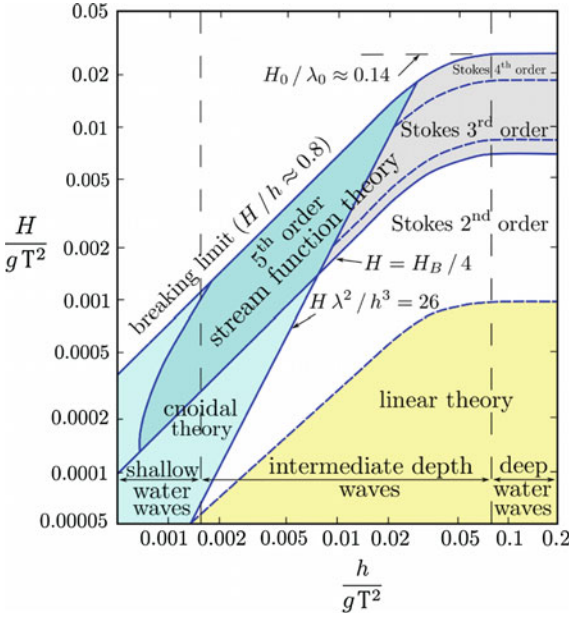


validity are shown in Fig. 5 according to [9]. Therefore, the free surface profile in linear wave theory can be described as

$$\eta = \frac{H}{2}\cos(kx - \omega t), \tag{12}$$

where H, k, ω are the wave height, wave number and wave angular frequency, respectively. As indicated in [9], the velocity profiles in deep water ($h/\lambda \geq 1/2$) are given by

Fig. 5 Range of validity of various wave theories



$$u = \frac{\pi H}{T} e^{kz} \cos(kx - \omega t), \tag{13a}$$

$$w = \frac{\pi H}{T} e^{kz} \sin(kx - \omega t), \tag{13b}$$

and in intermediate water $(1/20 \leq h/\lambda \leq 1/2)$

$$u = \frac{\pi H}{T} \frac{\cosh(k(z+h))}{\sinh(kh)} \cos(kx - \omega t), \tag{14a}$$

$$w = \frac{\pi H}{T} \frac{\sinh(k(z+h))}{\sinh(kh)} \sin(kx - \omega t), \tag{14b}$$

and in shallow water $(h/\lambda \leq 1/20)$

$$u = \frac{H}{2} \sqrt{\frac{g}{h}} \cos(kx - \omega t), \tag{15a}$$

$$w = \frac{\pi H}{T} \frac{z+h}{h} \sin(kx - \omega t), \tag{15b}$$

where T represents the wave period. The wave parameters are listed in Table 1. Therefore, our simulation is in the deep water condition. With these given values, the wave is generated and compared to the analytical solution in Fig. 6.

Table 1 Wave parameters for the wave energy conversion problem.

H (m)	h (m)	λ (m)	k (rad/m)	T (s)	ω (rad/s)
1	12	20	0.314	3.581	1.755

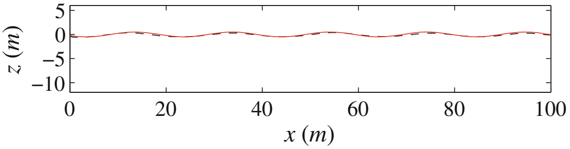


Fig. 6 Simulated wave propagation (*black dashed*) corresponding to the Airy wave theory (*red*)

2.4 Pressure-Velocity Decoupling Algorithm

The difficulty of the numerical solution of fluid Navier-Stokes equations lies in the fact that the velocity and pressure are coupled through the incompressibility constraint. The fluid governing Eqs. (4) and (5) when discretized with the standard finite volume method in a control volume (see Fig. 7) can be written as

$$a_P \boldsymbol{u}_P^{n+1} = \boldsymbol{H}(\boldsymbol{u}^{n+1}) - \nabla p^{n+1}, \tag{16a}$$

$$\nabla \cdot \boldsymbol{u}^{n+1} = \sum_f \boldsymbol{u}_f^{n+1} \cdot \boldsymbol{S}_f = 0, \tag{16b}$$

where $\boldsymbol{H}(\boldsymbol{u}^{n+1})$ contains the convective, diffusive, temporal terms and all the source terms, namely

$$\boldsymbol{H}(\boldsymbol{u}^{n+1}) = -\sum_n a_N \boldsymbol{u}_N^{n+1} + \frac{\boldsymbol{u}^n}{\Delta t} + \boldsymbol{S}_{\boldsymbol{u}}. \tag{17}$$

Apparently the pressure term does not explicitly appear in the continuity equation. To derive the equation for the pressure, we insert the discrete momentum equation into the discrete continuity equation, which gives

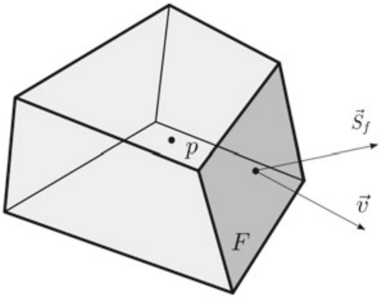
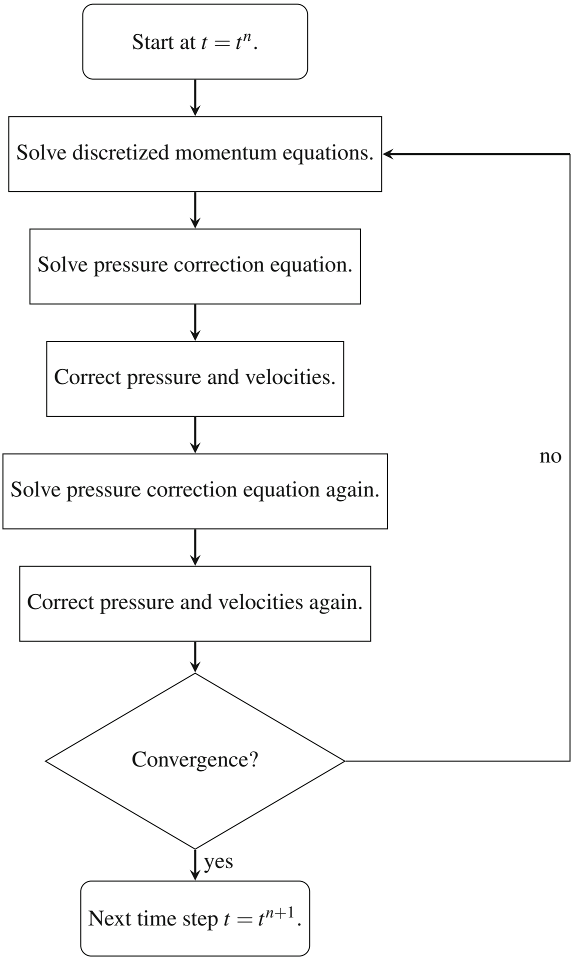


Fig. 7 A control volume of the fluid domain

$$\nabla \cdot \left(\frac{1}{a_P} \nabla p^{n+1} \right) = \nabla \cdot \left(\frac{\boldsymbol{H}(\boldsymbol{u}^{n+1})}{a_P} \right). \tag{18}$$

To decouple the velocity and pressure, we adopt the PISO (Pressure implicit with splitting of operator) algorithm originally proposed by Issa [10]. Other types of pressure-velocity decoupling algorithms and their comparisons can be found in [11]. PISO algorithm features a predictor-corrector splitting scheme, which predicts the velocity fields with an initial guess of pressure and corrects the pressure two times with the predicted velocity, as shown in Fig. 8.

Fig. 8 PISO algorithm for decoupling pressure and velocity


Start at $t = t^n$.
Solve discretized momentum equations.
Solve pressure correction equation.
Correct pressure and velocities.
Solve pressure correction equation again.
Correct pressure and velocities again.
Convergence?
no
yes
Next time step $t = t^{n+1}$.

2.5 Fluid-Structure Interaction

At each time step, the cylinder oscillates vertically under the impact of incoming waves and redistributes the surrounding fluids, which results in a typical fluid-structure interaction problem.

In the present study, we solve the fluid and solid equations with a segregated scheme for the sake of efficiency. We first start with the fluid equations with previous solid positions and velocities. Once the fluid fields are calculated, we can obtain the total force exerted on the solid and use it to solve the solid equations. A new fluid mesh is then established with the new position of solid. The overall scheme is shown in Fig. 9.

Note that the fluid and solid are strongly coupled by nature, thus a relative small time step is used in our explicit implementations. Otherwise, iterations between the fluid and solid parts are required at each time step, which is in general time-consuming. Highly efficient iterative schemes are always demanding. Besides, only small amplitude of oscillation is allowed, so that the mesh around the moving body will not be largely distorted. However this can be circumvented by the non-body conforming method, such as the immersed boundary method [12, 13]. Large displacements of solid and even arbitrary motions can be tackled easily. This will be considered in the future work.

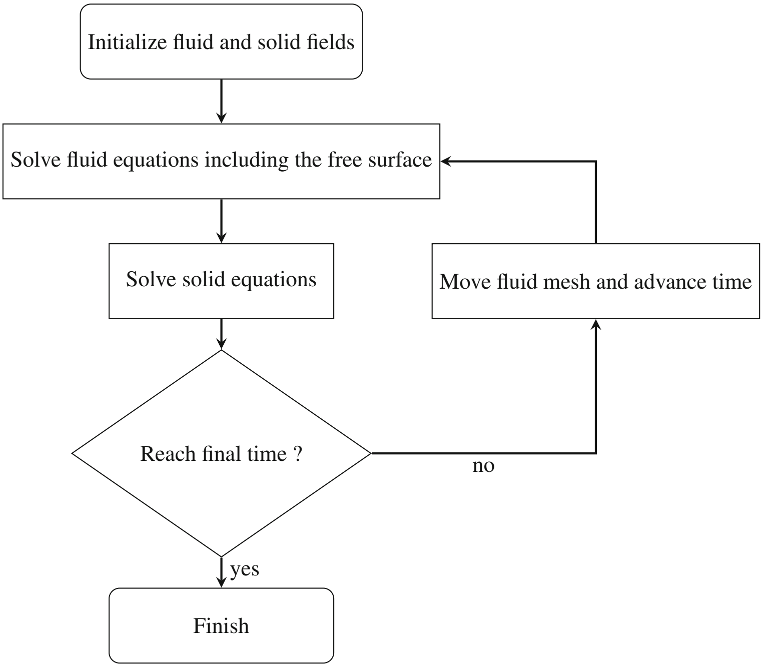

Initialize fluid and solid fields
Solve fluid equations including the free surface
Solve solid equations
Move fluid mesh and advance time
Reach final time ?
no
yes
Finish

Fig. 9 Solution procedure for the fluid-structure interaction

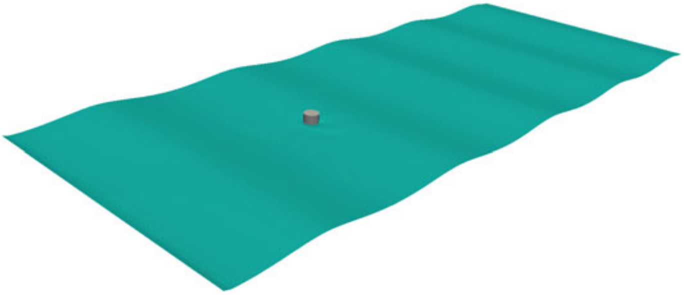


Fig. 10 Instantaneous free surface elevation

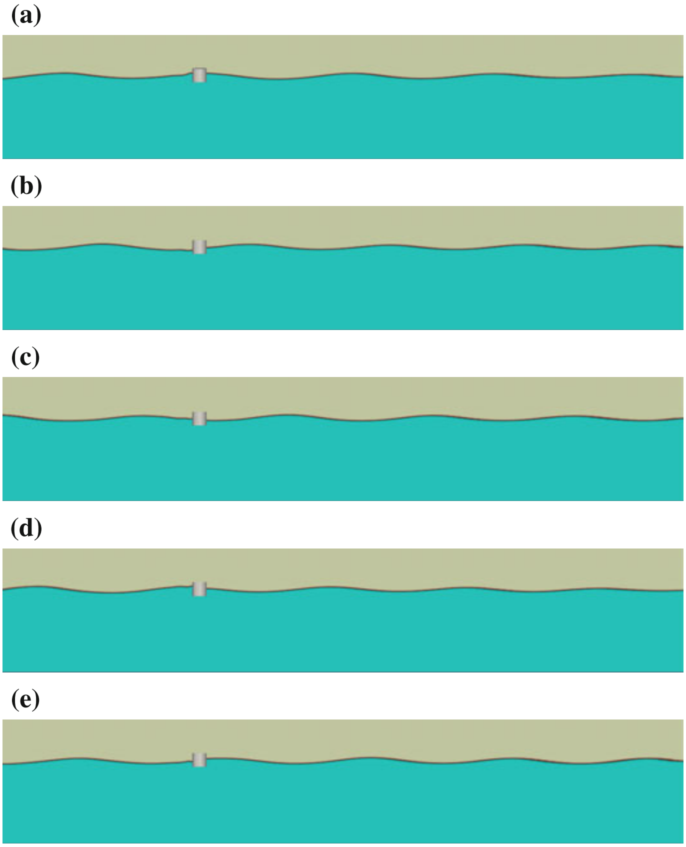


Fig. 11 The heaving semi-submerged cylinder at different time levels. **a** t = 42 s, **b** t = 43 s, **c** t = 44 s, **d** t = 45 s, **e** t = 46 s

3 Energy Extraction

Figure 10 shows the instantaneous free surface elevation around the heaving cylinder. The flow around the cylinder has clearly demonstrated the three-dimensional effect. The cylinder positions at different time levels are shown in Fig. 11. The velocity profiles beneath the cylinder are also plotted in Fig. 12.

In this simulation since the power absorption coefficient is a constant value, the PTO is considered as an ideal linear damper. But the non-linear dampers can be also incorporated into the model without changing the algorithm structure. Actually the energy absorption depends on not only the damping coefficient but also the wave conditions, dimensions of the converter, etc. The instantaneous velocity and power absorbed by the floating body is shown in Figs. 13 and 14 with the wave parameters given by Table 1.

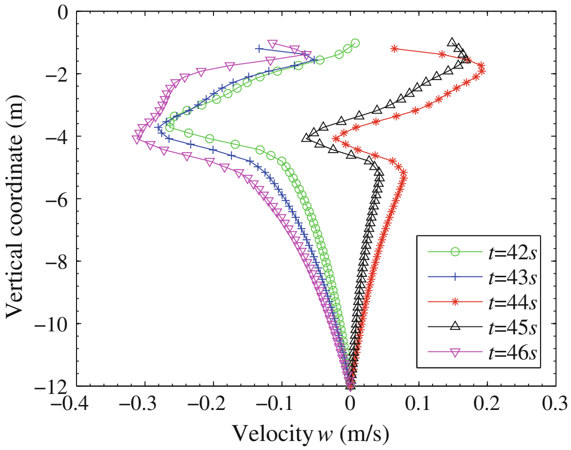


Fig. 12 Vertical velocity profiles along the vertical direction beneath the cylinder at different time levels

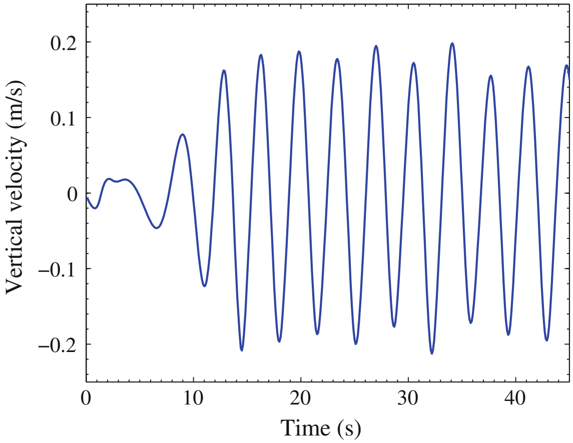
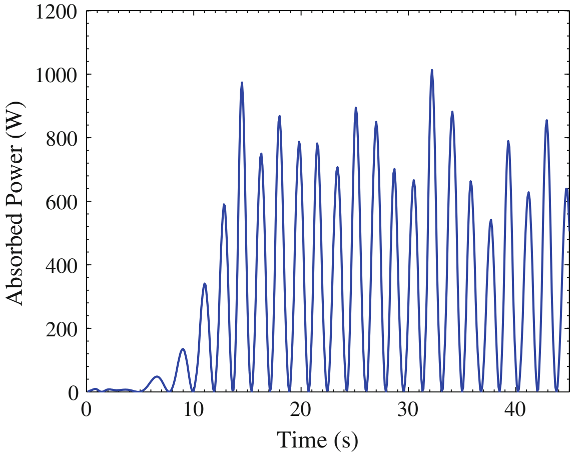


Fig. 13 Time evolution of vertical velocity of the heaving cylinder

Fig. 14 Instantaneous power P_{PTO} absorbed by the heaving cylinder



Theoretically the converter can achieve an optimal conversion efficiency when its natural frequency is equal to the frequency of the waves. At other frequencies, much less efficiency is obtained. Therefore in practice, control strategies can be used to improve the conversion efficiency considerably.

4 Conclusions

A numerical model is presented in this paper to demonstrate the ability of CFD techniques for the simulation of wave energy conversion. The solution is obtained with the viscous Navier-Stokes equations and the free surface is captured by the efficient VOF method. A numerical wave tank has been established by specifying the wave boundary conditions according to the wave theory. The fluid-structure interaction is taken into consideration for predicting the motion of the heaving cylinder. The numerical results in the present study show the validity of the proposed model.

Acknowledgments The first author greatly acknowledges the China Scholarship Council for the financial support.

References

1. Börner, T., Alam, M.-R.: Real Time Hybrid Modeling for Ocean Wave Energy Converters. Renewable and Sustainable Energy Reviews **43**, 784–795 (2015)
2. Cordonnier, J., Gorintin, F., Cagny, A.De, Clément, A.H., Babarit, A.: SEAREV: Case study of the development of a wave energy converter. Renewable Energy **80**, 40–52 (2015)
3. Falcao, A.F.O.: Wave energy utilization: A review of the technologies. Renewable and sustainable energy reviews **14**, 899–918 (2010)

4. Ouahsine, A., Smaoui, H., Meftah, K., Sergent, P.: Numerical study of coastal sandbar migration by hydro-morphodynamical coupling. Environmental Fluid Mechanics **13**, 169–187 (2013)
5. Ji, S., Ouahsine, A., Smaoui, H., Sergent, P.: 3D Modeling of sediment movement by ships-generated wakes in confined shipping channel. International Journal of Sediment Research **29**, 49–58 (2014)
6. Hirt, C.W., Nichols, B.D.: Volume of luid (VOF) method for the dynamics of free boundaries. Journal of Computational Physics. **39**, 201–225 (1981)
7. Cai, S.-G., Ouahsine, A., Sergent, P., Villon, P.: 3D Numerical analysis of wave energy conversion of a semi-submerged heaving cylinder. In: Ibrahimbegovic, A et al. Proceedings of the 2nd International Conference on Multi-scale Computational Methods for Solids and Fluids. June 10–12, Sarajevo, Bosnia and Herzegovina, 38–39 Eccomas Thematic Conference (2015)
8. Genest, R., Bonnefoy, Clément, A.H., Babarit, A.: Effect of non-ideal power take-off on the energy absorption of a reactively controlled one degree of freedom wave energy converter. Applied Ocean Research. **48**, 236–243 (2014)
9. Le Méhauté, B.: An introduction to hydrodynamics and water waves. Springer, Berlin (1976)
10. Issa, R.I.: Solution of the implicitly discretised fluid flow equations by operator-splitting. Journal of Computational Physics **62**, 40–65 (1985)
11. Barton, I.E.: Comparison of SIMPLE- and PISO type algorithms for transient flows. International Journal for Numerical Method in Fluids **26**, 459–483 (1998)
12. Cai, S.-G., Ouahsine, A., Smaoui, H., Favier, J., Hoarau, Y.: An efficient implicit direct forcing immersed boundary method for incompressible flows. Journal of Physics: Conference Series. **574**, 012165 (2014)
13. Cai, S.-G., Ouahsine, A., Favier, J., Hoarau, Y., Smaoui, H.: Immersed boundary method for the simulation of lid-driven cavity flow with an embedded cylinder. COUPLED PROBLEMS 2015 - Proceedings of the 6th International Conference on Coupled Problems in Science and Engineering. Venice, Italy, 1130–1137 (2015)
14. McCormick, M.E.: Ocean wave energy conversion. Dover, New York (2007)

Multiscale Modeling of Imperfect Interfaces and Applications

S. Dumont, F. Lebon, M.L. Raffa, R. Rizzoni and H. Welemane

Abstract Modeling interfaces between solids is of great importance in the fields of mechanical and civil engineering. Solid/solid interface behavior at the microscale has a strong influence on the strength of structures at the macroscale, such as gluing, optical systems, aircraft tires, pavement layers and masonry, for instance. In this lecture, a deductive approach is used to derive interface models, i.e. the thickness of the interface is considered as a small parameter and asymptotic techniques are introduced. A family of imperfect interface models is presented taking into account cracks at microscale. The proposed models combine homogenization techniques for micro-cracked media both in three-dimensional and two-dimensional cases, which leads to a cracked orthotropic material, and matched asymptotic method. In particular, it is shown that the Kachanov type theory leads to soft interface models and, alternatively,

S. Dumont
Université de Nîmes Place Gabriel Péri, 30000 Nîmes, France

S. Dumont
Laboratoire de Mécanique et d'Acoustique (LMA) - CNRS UPR 7051, 4 Impasse
Nikola Tesla, CS 40006-13453 Marseille Cedex 13, France

F. Lebon (✉)
Laboratoire de Mécanique et d'Acoustique (LMA) - CNRS UPR 7051,
Aix-Marseille University, Centrale Marseille, 4 Impasse Nikola Tesla,
CS 40006 - 13453 Marseille Cedex 13, France
e-mail: lebon@lma.cnrs-mrs.fr

M.L. Raffa
Department of Civil Engineering and Computer Science Engineering (DICII),
University of Rome "Tor Vergata", via del Politecnico 1 - 00133, Rome, Italy

M.L. Raffa
Laboratoire de Mécanique et d'Acoustique (LMA) - CNRS UPR 7051,
Aix-Marseille University, 4 Impasse Nikola Tesla CS 40006 - 13453,
Marseille Cedex 13, France

R. Rizzoni
Department of Engineering, University of Ferrara, Via Saragat 1, 44122, Ferrara, Italy

H. Welemane
University of Toulouse, National Engineering School of Tarbes,
47, avenue d'Azereix, BP 1629, 65016 Tarbes Cedex, France

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_5

that Goidescu et al. theory leads to stiff interface models. A fully nonlinear variant of the model is also proposed, derived from the St. Venant-Kirchhoff constitutive equations. Some applications to elementary masonry structures are presented.

1 Introduction

The study of interfaces between deformable solids significantly developed thanks to the rising interest of scientists and industries in mechanics of composite materials. Those first studies, in particular, focused on the presence of matrix-fiber interfaces in composite media and their effect on the determination of the effective thermoelastic properties of this kind of materials. Within the framework of these theories on mechanical behavior of composites, a commonly adopted assumption was the requirement of the continuity in terms of stresses and displacements at the interfaces among the principal constituents. The stress-based interface condition origins from the local equilibrium and the displacement-based interface condition derives from the hypothesis of perfect bonding. Such an interface condition was defined as *perfect interface*. Nevertheless, the assumption of perfect interfaces is established to be inappropriate in many mechanical problems. Indeed, the interface between two bodies or two parts of a body, defined as *adherents*, is a favorable zone to complex physico-chemical reactions and vulnerable to mechanical damage. Goland and Reissner [22], in the forties, were surely the first to model a thin *adhesive* as a weak interface, i.e. they were the first to assume that the adherents were linked by a continuous distribution of springs. Such an interface, is defined as *spring type*. Goland and Reissner have noted that the thinness suggests to consider constant stresses in the adhesive, and some years later, Gilibert and Rigolot [19] found a rational justification of this fact by means of the asymptotic expansion method, assuming that the thickness and the elastic properties of the adhesive had the same order of smallness ε. During the eighties and nineties, the relaxation of the perfect interface approximation was largely investigated, aiming principally to applicate these theories to composite materials with coated fibers or particles [3, 26], or in the case of decohesion and nucleation problems in cohesive zones [44, 45]. One of the first definition of *imperfect interface* was certainly due to Hashin and Benveniste [3, 26]. Particularly, Hashin concentrates his research in the case of composite material with thin layer or coating enveloping its reinforcing constituents (fibers). Such an interfacial layer is generally referred to as *interphase*, and its presence can be due to chemical interaction between the constituents or it may be introduced by design aiming to improve the mechanical properties of the composite. Several investigations in literature, before and after the work of Hashin, modeled this kind of problem with the so-called *three-phase-material theory*. Such a description requires, obviously, the knowledge of the interphase properties. These constitutive informations are rarely available, primarily, because the interphase material properties are in situ properties which are not necessarily equal to those of the bulk material, and additionally, the interphase may vary within a transition zone from one constituent to another.

Accordingly, in most cases, the interphase properties are unmeasurable. The *Imperfect interface theory* was formulated by Hashin [26–29] in order to overcome these challenges. This alternative model was based to the main idea that if the interphase has significant effects on the overall response, then its properties must be significantly different from those of the constituents, in general, much more flexible. To this aim, the attempts for explicit modeling of the three-dimensional interphase are highly reduced by replacing it with a two-dimensional imperfect interface. In particular, within the Hashin imperfect interface model [26] the discontinuity in terms of displacements is allowed, instead, the continuity in terms of stresses, according local equilibrium, is preserved. Hashin [26], as Goland [22] before him, made the simplest assumption that the displacement discontinuity is linearly proportional to the traction vector:

$$\boldsymbol{\sigma}^{(\Omega_1)}\mathbf{n} = \boldsymbol{\sigma}^{(\Omega_2)}\mathbf{n} = \mathbf{D}\,[\mathbf{u}] \quad [\mathbf{u}] = \mathbf{u}^{(\Omega_1)} - \mathbf{u}^{(\Omega_2)}$$

where $\boldsymbol{\sigma}^{(\Omega_1),(\Omega_2)}\mathbf{n}$ is the interfacial stress vector relative to the solids Ω_1 and Ω_2 in contact; $[\mathbf{u}]$ and $\mathbf{u}^{(\Omega_1),(\Omega_2)}$ are the displacement jump vector and displacement vector, respectively; $\mathbf{D}$ is a matrix which contains the spring constant type material parameters, in normal and tangential directions; these latter have dimension of stress divided by length. In the following, these parameters are referred as *interface stiffnesses*. It is worth to point out that infinite values of the interface stiffnesses imply vanishing of displacement jumps and therefore perfect interface conditions. At the other asymptotic limit, zero values of the stiffnesses imply vanishing of interface tractions and therefore, debonding conditions. Any finite positive values of the interface stiffnesses define an imperfect interface.

Hashin, with his pioneering work, determined the effective elastic properties and thermal expansions coefficients both for unidirectional fiber composites with imperfect interfaces conditions [26] and for composites with spherical inclusions and particles [27, 28]. Moreover, he demonstrated that the three-phase-material approach was a special case of the imperfect interface theory. It is worth remarking that Hashin, as first, showed that the interface stiffnesses (he referred them as *interface parameters*) can be simply related to the interphase properties and geometry [26].

Hashin and Benveniste, independently, generalized the classical extremum principles of the theory of elasticity for composite bodies to the case of an imperfect interface described by linear relations between interface displacement jumps and tractions [3, 29].

In the work of Bövik [9], the idea to use a simple tool that is the Taylor expansion of the relevant physical fields in a thin interphase, combined with the use of surface differential operators on a curved surface, has been applied to achieve the representation of a thin interphase by an imperfect interface. The idea of a Taylor expansion was also adopted by Hashin to derive the spring-type interface model for soft elastic interfaces [28] and for interphases of arbitrary conductivity and elastic moduli [30]. More recently, Gu [23, 24] derived an imperfect interface model for coupled multifield phenomena (thermal conductivity, elasticity and piezoelectricity) by applying Taylor expansion to an arbitrarily curved thin interphase between two adjoining phases; he also introduced some new coordinate-free interfacial operators.

All the above cited imperfect interface models are derived by assuming an isotropic interphase.

In a quite recent work Benveniste [4], provided a generalization of the Bövik model to an arbitrarily curved three-dimensional thin anisotropic layer between two anisotropic media. Benveniste model is carried out in the setting of unsteady heat conduction and dynamic elasticity. The derived interface model consists of expressions for the jumps in the physical fields, i.e. temperature and normal heat flux in conduction, and displacements and tractions in elasticity, across the interface.

Additionally, derivations of spring-type interface models by using asymptotic methods, for different geometrical configurations, have been given, among other by Klarbring [36, 37] and Geymonat [18].

A much less studied imperfect interface condition is the one obtained starting from a stiff and thin interphase, the so called *stiff* interface (or equivalently *hard* interface). Differently from the soft case, the hard interface is characterized by a jump of the traction vector across the interface and by continuity of displacements. Benveniste and Miloh [5], generalizing the study made by Caillerie [10] for curved interfaces, demonstrate that depending on its degree of stiffness with respect to the bodies in contact, a stiff thin interphase translates itself into a much richer class of imperfect interfaces than a soft interphase does. Within their study, a thin curved isotropic layer of constant thickness between two elastic isotropic media in a two-dimensional plane strain setting, is considered. The properties of the curved layer are allowed to vary in the tangential direction. It is shown that depending on the softness or stiffness of the interphase with respect to the neighboring media, as determined by the magnitude of the non-dimensional Lamé parameters $\bar{\lambda_c}$ and $\bar{\mu_c}$, there exists seven distinct regimes of interface conditions according the following expressions:

$$\bar{\lambda}_i = \frac{\tilde{\lambda}_i}{\varepsilon^N} \qquad \bar{\mu}_i = \frac{\tilde{\mu}_i}{\varepsilon^N}$$

where $\tilde{\lambda}_c$ and $\tilde{\mu}_c$ are non-dimensional constant Lamé parameters of the material interphase, ε in the non-dimensional interphase thickness and N is a negative or positive integer or zero. Accordingly with the above definition these regimes may be distinguished in: (a) vacuous contact type interface for $N \leqslant -2$, (b) spring type interface for $N = -1$, (c) ideal contact type interface for $N = 0$, (d) membrane type interface for $N = 1$, (e) inextensible membrane type interface for $N = 2$, (f) inextensible shell type interface for $N = 3$, (g) rigid contact type interface for $N \geqslant 4$. The first two conditions are characteristic of a soft interphase whereas the last four are characteristic of a stiff interphase. The cases (a), (c) and (g) are the classical ones: in case (a) the tractions vanish (debonding), in case (c) the displacements and tractions are continuous (perfect interface condition), and in case (g) the displacements vanish. Benveniste and Miloh [5], for the first time, derived the interface conditions for the hard cases (d), (e) and (f), by applying a formal asymptotic expansion for the displacements and stresses fields in the thin layer interphase.

In the present chapter, two kind of imperfect interface conditions are essentially referred to: the soft interphase case which brings to a spring-type, both linear and nonlinear, interface, and the hard interphase case which brings to a more general interface model that includes, as will be shown in the next section, the perfect interface conditions. In order to make an analogy with the Benveniste's classifications, the cases with $N = -1$ and $N = 0$ will be analyzed. It is worth remarking some differences between the hard interface case considered in this work and that defined by Benveniste and Miloh for $N = 0$. In fact in the work, the case $N = 0$, according formers papers [14, 38, 39, 52, 53] will be studied within the framework of higher order theory. This choice, extensively detailed in the following, leads to the evidence that the case $N = 0$ is an effective imperfect interface condition, i.e. stress vector and displacement jump vector in the one-order formulation have been found to be non-null. As a result, the perfect interface has been established to be a particular case of the hard interface condition at the zero-order [53], in what follows this evidence will be analytically derived within the asymptotic framework.

The imperfect interface models, object of the present chapter, are consistently derived by coupling a homogenization approach for microcracked media under the non-interacting approximation (NIA) [21, 34, 35, 57, 60], and arguments of asymptotic analysis [38, 39, 52, 53]. Such a method, is defined *imperfect interface approach*.

The text is organized as follows: Sects. 2 and 3 are devoted to detail the framework of the imperfect interface approach and to enforce it in order to derive several interface models, particularly, in Sect. 3 a nonlinear imperfect interface model is presented; Sect. 4 is consecrated to a simple numerical application useful to validate these interface models; finally, in Sect. 5 some conclusions are outlined.

2 Imperfect Interface Approach

In this section, it is shown how matched asymptotic expansion method coupled together an homogenization technique for microcracked media, give birth to both soft and hard imperfect interface laws.

2.1 Matched Asymptotic Expansion Method

It is worth pointing out that the application of the asymptotic methods to obtain governing equations of imperfect interface starting from thin layer problems in the elasticity framework, has a consistent mathematical background [10–13, 26–28, 30, 55, 56]. Ould Khaoua among other, in his doctoral thesis [46], studied the elastic equilibrium problem $\mathscr{P}^{\varepsilon}$ under the hypothesis of small perturbations. The author demonstrates that the solution of the reference problem (i.e. with an elastic thin layer of thickness ε) $\mathscr{P}^{\varepsilon}$, that is expressed in terms of both stress and displacement fields

$(\boldsymbol{\sigma}^{\varepsilon}, \mathbf{u}^{\varepsilon})$, for $\varepsilon \to 0$ admits a limit $(\boldsymbol{\sigma}, \mathbf{u})$ and that this limit solution is also the solution of the limit problem $\mathscr{P}$ ($\mathscr{P}^{\varepsilon} \to \mathscr{P}$ for $\varepsilon \to 0$). Additionally, Ould Khaoua [46] has found, as Hashin [26] before, that the mechanical and geometrical characteristics of the layer (interphase) are retained in the interface stiffnesses of the soft interface governing equations.

The matched asymptotic expansion method [38–40, 52, 53], adopted in this work, is detailed in the following paragraphs.

2.1.1 General Notations

With reference to [53], let the problem general notations be detailed. A thin layer $\mathscr{B}^{\varepsilon}$ with cross-section $\mathscr{S}$ and uniform small thickness $\varepsilon \ll 1$ is considered, $\mathscr{S}$ being an open bounded set in $\mathbb{R}^2$ with a smooth boundary. In the following $\mathscr{B}^{\varepsilon}$ and $\mathscr{S}$ will be called *interphase* and *interface*, respectively. The interphase lies between two bodies, named as *adherents*, occupying the reference configurations $\Omega^{\varepsilon}{}_{\pm} \subset \mathbb{R}^3$. In such a way, the interphase represents the adhesive joining the two bodies Ω^{ε}_{+} and Ω^{ε}_{-}. Let $\mathscr{S}^{\varepsilon}{}_{\pm}$ be taken to denote the plane interfaces between the interphase and the adherents and let $\Omega^{\varepsilon} = \Omega^{\varepsilon}{}_{\pm} \cup \mathscr{S}^{\varepsilon}{}_{\pm} \cup \mathscr{B}^{\varepsilon}$ denote the composite system comprising the interphase and the adherents.

It is assumed that the adhesive and the adherents are perfectly bonded and thus, the continuity of the displacement and stress vector fields across $\mathscr{S}^{\varepsilon}{}_{\pm}$ is ensured.

An orthonormal Cartesian basis $(O, \mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3)$ is introduced and let (x_1, x_2, x_3) be taken to denote the three coordinates of a particle. The origin lies at the center of the interphase midplane and the x_3-axis runs perpendicular to the open bounded set $\mathscr{S}$, as illustrated in Fig. 1.

The materials of the composite system are assumed to be homogeneous and initially linearly elastic and let $\mathbb{A}_{\pm}, \mathbb{B}^{\varepsilon}$ be the fourth-rank elasticity tensors of the

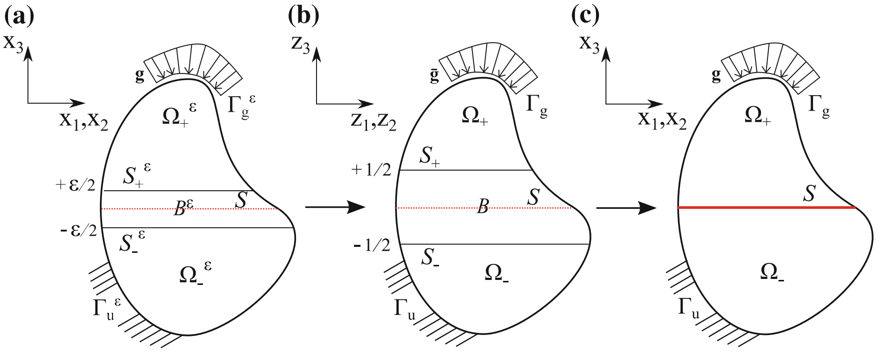


Fig. 1 Asymptotic procedure—Synoptic sketch of three steps performed in the matched asymptotic expansion approach: **a** the reference configuration with the ε-thick interphase; **b** the rescaled configuration; **c** the final configuration with the zero-thick interface

adherents and of the interphase, respectively. The tensors $\mathbb{A}_{\pm}$, $\mathbb{B}^{\varepsilon}$ are assumed to be symmetric, with the minor and major symmetries, and positive definite. The adherents are subjected to a body force density $\mathbf{f}^{\pm} : \Omega^{\varepsilon}{}_{\pm} \mapsto \mathbb{R}^3$ and to a surface force density $\mathbf{g}^{\pm} : \Gamma_g^{\varepsilon} \mapsto \mathbb{R}^3$ on $\Gamma_g^{\varepsilon} \subset (\partial\Omega^{\varepsilon}{}_{+} \backslash \mathscr{S}_{+}^{\varepsilon}) \cup (\partial\Omega^{\varepsilon}{}_{-} \backslash \mathscr{S}_{-}^{\varepsilon})$. Body forces are neglected in the adhesive.

On $\Gamma_u^{\varepsilon} = (\partial\Omega^{\varepsilon}{}_{+} \backslash \mathscr{S}_{+}^{\varepsilon}) \cup (\partial\Omega^{\varepsilon}{}_{-} \backslash \mathscr{S}_{-}^{\varepsilon}) \backslash \Gamma_g^{\varepsilon}$, homogeneous boundary conditions are prescribed:

$$\mathbf{u}^{\varepsilon} = \mathbf{0} \quad \text{on } \Gamma_u^{\varepsilon}, \tag{1}$$

where $\mathbf{u}^{\varepsilon} : \Omega^{\varepsilon} \mapsto \mathbb{R}^3$ is the displacement field defined on Ω^{ε}. Γ_g^{ε}, Γ_u^{ε} are assumed to be located far from the interphase, in particular, the external boundaries of the interphase $\mathscr{B}^{\varepsilon}$ $(\partial\mathscr{S} \times (-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}))$ are assumed to be stress-free. The fields of the external forces are endowed with sufficient regularity to ensure the existence of equilibrium configuration.

2.1.2 Rescaling Process

The rescaling phase in the asymptotic process represents a mathematical construct [12], not a physically-based configuration of the studied composed system. This standard step is used in the asymptotic expansion method, in order to eliminate the dependency on the small parameter ε of the integration domains. This rescaling procedure can be seen as a change of spatial variables $\hat{\mathbf{p}} : (x_1, x_2, x_3) \rightarrow (z_1, z_2, z_3)$ in the interphase [12]:

$$z_1 = x_1, \qquad z_2 = x_2, \qquad z_3 = \frac{x_3}{\varepsilon} \tag{2}$$

resulting

$$\frac{\partial}{\partial z_1} = \frac{\partial}{\partial x_1}, \qquad \frac{\partial}{\partial z_2} = \frac{\partial}{\partial x_2}, \qquad \frac{\partial}{\partial z_3} = \varepsilon \frac{\partial}{\partial x_3}. \tag{3}$$

Moreover, in the adherents the following change of variables $\bar{\mathbf{p}} : (x_1, x_2, x_3) \rightarrow (z_1, z_2, z_3)$ is also introduced:

$$z_1 = x_1, \qquad z_2 = x_2, \qquad z_3 = x_3 \pm \frac{1}{2}(1 - \varepsilon) \tag{4}$$

where the plus (minus) sign applies whenever $x \in \Omega^{\varepsilon}{}_{+}$ $(x \in \Omega^{\varepsilon}{}_{-})$, with

$$\frac{\partial}{\partial z_1} = \frac{\partial}{\partial x_1}, \qquad \frac{\partial}{\partial z_2} = \frac{\partial}{\partial x_2}, \qquad \frac{\partial}{\partial z_3} = \frac{\partial}{\partial x_3} \tag{5}$$

After the change of variables (2), the interphase occupies the domain

$$\mathscr{B} = \{(z_1, z_2, z_3) \in \mathbb{R}^3 : (z_1, z_2) \in \mathscr{S}, |z_3| < \frac{1}{2}\} \tag{6}$$

and the adherents occupy the domains $\Omega_{\pm} = {\Omega^{\varepsilon}}_{\pm} \pm \frac{1}{2}(1-\varepsilon)\mathbf{i}_3$, as shown in Fig. 1. The sets $\mathscr{S}_{\pm} = \{(z_1, z_2, z_3) \in \mathbb{R}^3 : (z_1, z_2) \in \mathscr{S}, z_3 = \pm\frac{1}{2}\}$ are taken to denote the interfaces between $\mathscr{B}$ and $\Omega_{\pm}$ and $\Omega = \Omega_{+} \cup \Omega_{-} \cup \mathscr{B} \cup \mathscr{S}_{+} \cup \mathscr{S}_{-}$ is the rescaled configuration of the composite body. Lastly, Γ_u and Γ_g indicates the images of Γ_u^{ε} and Γ_g^{ε} under the change of variables, and $\bar{\mathbf{f}}^{\pm} := \mathbf{f}^{\pm} \circ \bar{\mathbf{p}}^{-1}$ and $\bar{\mathbf{g}}^{\pm} := \mathbf{g}^{\pm} \circ \bar{\mathbf{p}}^{-1}$ the rescaled external forces.

2.1.3 Concerning Kinematics

Following the approach proposed by [38, 53], let focus on the kinematics of the elastic problem. After taking $\hat{\mathbf{u}}^{\varepsilon} = \mathbf{u}^{\varepsilon} \circ \hat{\mathbf{p}}^{-1}$ and $\bar{\mathbf{u}}^{\varepsilon} = \mathbf{u}^{\varepsilon} \circ \bar{\mathbf{p}}^{-1}$ to denote the displacement fields from the rescaled adhesive and adherents, respectively, the asymptotic expansions of the displacement fields with respect to the small parameter ε take the form:

$$\mathbf{u}^{\varepsilon}(x_1, x_2, x_3) = \mathbf{u}^0 + \varepsilon\mathbf{u}^1 + \varepsilon^2\mathbf{u}^2 + o(\varepsilon^2) \tag{7a}$$

$$\hat{\mathbf{u}}^{\varepsilon}(z_1, z_2, z_3) = \hat{\mathbf{u}}^0 + \varepsilon\hat{\mathbf{u}}^1 + \varepsilon^2\hat{\mathbf{u}}^2 + o(\varepsilon^2) \tag{7b}$$

$$\bar{\mathbf{u}}^{\varepsilon}(z_1, z_2, z_3) = \bar{\mathbf{u}}^0 + \varepsilon\bar{\mathbf{u}}^1 + \varepsilon^2\bar{\mathbf{u}}^2 + o(\varepsilon^2) \tag{7c}$$

Interphase:

The displacement gradient tensor of the field $\hat{\mathbf{u}}^{\varepsilon}$ in the rescaled interphase is computed as:

$$\hat{\mathbf{H}} = \varepsilon^{-1}\begin{bmatrix} 0 & \hat{u}^0_{\alpha,3} \\ 0 & \hat{u}^0_{3,3} \end{bmatrix} + \begin{bmatrix} \hat{u}^0_{\alpha,\beta} & \hat{u}^1_{\alpha,3} \\ \hat{u}^0_{3,\beta} & \hat{u}^1_{3,3} \end{bmatrix} + \varepsilon\begin{bmatrix} \hat{u}^1_{\alpha,\beta} & \hat{u}^2_{\alpha,3} \\ \hat{u}^1_{3,\beta} & \hat{u}^2_{3,3} \end{bmatrix} + O(\varepsilon^2) \tag{8}$$

where $\alpha = 1, 2$, so that the strain tensor can be obtained as:

$$\mathbf{e}(\hat{\mathbf{u}}^{\varepsilon}) = \varepsilon^{-1}\hat{\mathbf{e}}^{-1} + \hat{\mathbf{e}}^0 + \varepsilon\hat{\mathbf{e}}^1 + O(\varepsilon^2) \tag{9}$$

with:

$$\hat{\mathbf{e}}^{-1} = \begin{bmatrix} 0 & \frac{1}{2}\hat{u}^0_{\alpha,3} \\ \frac{1}{2}\hat{u}^0_{\alpha,3} & \hat{u}^0_{3,3} \end{bmatrix} = Sym(\hat{\mathbf{u}}^0_{,3} \otimes \mathbf{i}_3) \tag{10}$$

$$\hat{\mathbf{e}}^k = \begin{bmatrix} Sym(\hat{u}^k_{\alpha,\beta}) & \frac{1}{2}(\hat{u}^k_{3,\alpha} + \hat{u}^{k+1}_{\alpha,3}) \\ \frac{1}{2}(\hat{u}^k_{3,\alpha} + \hat{u}^{k+1}_{\alpha,3}) & \hat{u}^{k+1}_{3,3} \end{bmatrix} = Sym(\hat{\mathbf{u}}^k_{,1} \otimes \mathbf{i}_1 + \hat{\mathbf{u}}^k_{,2} \otimes \mathbf{i}_2 + \hat{\mathbf{u}}^{k+1}_{,3} \otimes \mathbf{i}_3) \tag{11}$$

where $Sym(\cdot)$ gives the symmetric part of the enclosed tensor and $k = 0, 1$, and $\otimes$ is the dyadic product between vectors such as: $(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i\, b_j$. Moreover, the following notation for derivatives is adopted: $f_{,j}$ denoting the partial derivative of f with respect to z_j.

Adherents:

The displacement gradient tensor of the field $\bar{\mathbf{u}}^{\varepsilon}$ in the adherents is computed as:

$$\bar{\mathbf{H}} = \begin{bmatrix} \bar{u}^0_{\alpha,\beta} & \bar{u}^0_{\alpha,3} \\ \bar{u}^0_{3,\beta} & \bar{u}^0_{3,3} \end{bmatrix} + \varepsilon \begin{bmatrix} \bar{u}^1_{\alpha,\beta} & \bar{u}^1_{\alpha,3} \\ \bar{u}^1_{3,\beta} & \bar{u}^1_{3,3} \end{bmatrix} + O(\varepsilon^2) \tag{12}$$

so that the strain tensor can be obtained as:

$$\mathbf{e}(\bar{\mathbf{u}}^{\varepsilon}) = \varepsilon^{-1}\bar{\mathbf{e}}^{-1} + \bar{\mathbf{e}}^{0} + \varepsilon\bar{\mathbf{e}}^{1} + O(\varepsilon^2) \tag{13}$$

with:

$$\bar{\mathbf{e}}^{-1} = \mathbf{0} \tag{14}$$

$$\bar{\mathbf{e}}^{k} = \begin{bmatrix} Sym(\bar{u}^k_{\alpha,\beta}) & \frac{1}{2}(\bar{u}^k_{3,\alpha} + \bar{u}^k_{\alpha,3}) \\ \frac{1}{2}(\bar{u}^k_{3,\alpha} + \bar{u}^k_{\alpha,3}) & \bar{u}^k_{3,3} \end{bmatrix} = Sym(\bar{\mathbf{u}}^k_{,1} \otimes \mathbf{i}_1 + \bar{\mathbf{u}}^k_{,2} \otimes \mathbf{i}_2 + \bar{\mathbf{u}}^k_{,3} \otimes \mathbf{i}_3) \tag{15}$$

with $k = 0, 1$.

2.1.4 Concerning Equilibrium

With reference to the work by [38, 53], the stress fields in the rescaled adhesive and adherents, $\hat{\boldsymbol{\sigma}}^{\varepsilon} = \boldsymbol{\sigma} \circ \hat{\mathbf{p}}^{-1}$ and $\bar{\boldsymbol{\sigma}}^{\varepsilon} = \boldsymbol{\sigma} \circ \bar{\mathbf{p}}^{-1}$ respectively, can be represented as asymptotic expansions:

$$\boldsymbol{\sigma}^{\varepsilon} = \boldsymbol{\sigma}^{0} + \varepsilon\boldsymbol{\sigma}^{1} + O(\varepsilon^2) \tag{16a}$$

$$\hat{\boldsymbol{\sigma}}^{\varepsilon} = \hat{\boldsymbol{\sigma}}^{0} + \varepsilon\hat{\boldsymbol{\sigma}}^{1} + O(\varepsilon^2) \tag{16b}$$

$$\bar{\boldsymbol{\sigma}}^{\varepsilon} = \bar{\boldsymbol{\sigma}}^{0} + \varepsilon\bar{\boldsymbol{\sigma}}^{1} + O(\varepsilon^2) \tag{16c}$$

Equilibrium Equations in the Interphase:

As body forces are neglected in the adhesive, the equilibrium equation is:

$$div\hat{\boldsymbol{\sigma}}^{\varepsilon} = \mathbf{0}. \tag{17}$$

Substituting the representation from (16b) into the equilibrium equation (17) and using (3), it becomes:

$$\begin{aligned} 0 &= \hat{\sigma}^{\varepsilon}_{i\alpha,\alpha} + \varepsilon^{-1}\hat{\sigma}^{\varepsilon}_{i3,3} \\ &= \varepsilon^{-1}\hat{\sigma}^{0}_{i3,3} + \hat{\sigma}^{0}_{i\alpha,\alpha} + \hat{\sigma}^{1}_{i3,3} + \varepsilon\hat{\sigma}^{1}_{i\alpha,\alpha} + O(\varepsilon) \end{aligned} \tag{18}$$

where $\alpha = 1, 2$. Equation (18) has to be satisfied for any value of ε, leading to:

$$\hat{\sigma}^{0}_{i3,3} = 0 \tag{19}$$

$$\hat{\sigma}^{0}_{i1,1} + \hat{\sigma}^{0}_{i2,2} + \hat{\sigma}^{1}_{i3,3} = 0 \tag{20}$$

where $i = 1, 2, 3$.

Equation (19) shows that $\hat{\sigma}^{0}_{i3}$ is not dependent on z_3 in the adhesive, and thus it can be written:

$$\left[\hat{\sigma}^{0}_{i3}\right] = 0 \tag{21}$$

where $[\bullet]$ denotes the jump between $z_3 = \frac{1}{2}$ and $z_3 = -\frac{1}{2}$.

In view of (21), Eq. (20), for $i = 3$, can be rewritten in the integrated form

$$[\hat{\sigma}^{1}_{33}] = -\hat{\sigma}^{0}_{13,1} - \hat{\sigma}^{0}_{23,2} \tag{22}$$

Equilibrium Equations in the Adherents:

The equilibrium equation in the adherents is:

$$\mathrm{div}\bar{\boldsymbol{\sigma}}^{\varepsilon} + \bar{\mathbf{f}} = \mathbf{0} \tag{23}$$

Substituting the representation form (16c) into the equilibrium Eq. (23) and taking into account that it has to be satisfied for any value of ε, it leads to:

$$\mathrm{div}\bar{\boldsymbol{\sigma}}^{0} + \bar{\mathbf{f}} = \mathbf{0} \tag{24}$$

$$\mathrm{div}\bar{\boldsymbol{\sigma}}^{1} = \mathbf{0} \tag{25}$$

2.1.5 Matching External and Internal Expansions

Due to the perfect bonded assumption between $\mathscr{B}^{\varepsilon}$ and $\Omega^{\varepsilon}_{\pm}$, the continuity conditions at $\mathscr{S}^{\varepsilon}_{\pm}$ for the fields $\mathbf{u}^{\varepsilon}$ and $\boldsymbol{\sigma}^{\varepsilon}$ lead to matching relationships between external and internal expansions [38, 53]. In particular, in terms of displacements the following relationship have to be satisfied:

$$\mathbf{u}^{\varepsilon}(\mathbf{x}_{\alpha}, \pm\frac{\varepsilon}{2}) = \hat{\mathbf{u}}^{\varepsilon}(\mathbf{z}_{\alpha}, \pm\frac{1}{2}) = \bar{\mathbf{u}}^{\varepsilon}(\mathbf{z}_{\alpha}, \pm\frac{1}{2}) \tag{26}$$

where $\mathbf{x}_\alpha := (x_1, x_2)$, $\mathbf{z}_\alpha := (z_1, z_2) \in \mathscr{S}$. Expanding the displacement in the adherent, $\mathbf{u}^\varepsilon$, in Taylor series along the x_3-direction and taking into account the asymptotic expansion (7a), it results:

$$\begin{aligned}\mathbf{u}^\varepsilon(\mathbf{x}_\alpha, \pm\frac{\varepsilon}{2}) &= \mathbf{u}^\varepsilon(\mathbf{x}_\alpha, 0^\pm) \pm \frac{\varepsilon}{2}\mathbf{u}^\varepsilon_{,3}(\mathbf{x}_\alpha, 0^\pm) + \cdots \\ &= \mathbf{u}^0(\mathbf{x}_\alpha, 0^\pm) + \varepsilon\mathbf{u}^1(\mathbf{x}_\alpha, 0^\pm) \pm \frac{\varepsilon}{2}\mathbf{u}^0_{,3}(\mathbf{x}_\alpha, 0^\pm) + \cdots \end{aligned} \tag{27}$$

Substituting Eqs. (7b) and (7c) together with formula (27) into continuity condition (26), it holds true:

$$\begin{aligned}\mathbf{u}^0(\mathbf{x}_\alpha, 0^\pm) + \varepsilon\mathbf{u}^1(\mathbf{x}_\alpha, 0^\pm) \pm \frac{\varepsilon}{2}\mathbf{u}^0_{,3}(\mathbf{x}_\alpha, 0^\pm) + \cdots &= \hat{\mathbf{u}}^0(\mathbf{z}_\alpha, \pm\frac{1}{2}) + \varepsilon\hat{\mathbf{u}}^1(\mathbf{z}_\alpha, \pm\frac{1}{2}) + \cdots \\ &= \bar{\mathbf{u}}^0(\mathbf{z}_\alpha, \pm\frac{1}{2}) + \varepsilon\bar{\mathbf{u}}^1(\mathbf{z}_\alpha, \pm\frac{1}{2}) + \cdots \end{aligned} \tag{28}$$

After identifying the terms in the same powers of ε, Eq. (28) gives:

$$\mathbf{u}^0(\mathbf{x}_\alpha, 0^\pm) = \hat{\mathbf{u}}^0(\mathbf{z}_\alpha, \pm\frac{1}{2}) = \bar{\mathbf{u}}^0(\mathbf{z}_\alpha, \pm\frac{1}{2}) \tag{29}$$

$$\mathbf{u}^1(\mathbf{x}_\alpha, 0^\pm) \pm \frac{1}{2}\mathbf{u}^0_{,3}(\mathbf{x}_\alpha, 0^\pm) = \hat{\mathbf{u}}^1(\mathbf{z}_\alpha, \pm\frac{1}{2}) = \bar{\mathbf{u}}^1(\mathbf{z}_\alpha, \pm\frac{1}{2}) \tag{30}$$

Following a similar analysis for the stress vector, analogous results are obtained [38, 53]:

$$\sigma^0_{i3}(\mathbf{x}_\alpha, 0^\pm) = \hat{\sigma}^0_{i3}(\mathbf{z}_\alpha, \pm\frac{1}{2}) = \bar{\sigma}^0_{i3}(\mathbf{z}_\alpha, \pm\frac{1}{2}) \tag{31}$$

$$\sigma^1_{i3}(\mathbf{x}_\alpha, 0^\pm) \pm \frac{1}{2}\sigma^0_{i3,3}(\mathbf{x}_\alpha, 0^\pm) = \hat{\sigma}^1_{i3}(\mathbf{z}_\alpha, \pm\frac{1}{2}) = \bar{\sigma}^1_{i3}(\mathbf{z}_\alpha, \pm\frac{1}{2}) \tag{32}$$

for $i = 1, 2, 3$.

Using the above results, it is possible to rewrite Eqs. (21) and (22) in the following form:

$$\begin{aligned}&[[\sigma^0_{i3}]] = 0, \quad i = 1, 2, 3 \\ &[[\sigma^1_{33}]] = -\sigma^0_{13,1} - \sigma^0_{23,2} - \langle\langle\sigma^0_{33,3}\rangle\rangle\end{aligned} \tag{33}$$

where $[[f]] := f(\mathbf{x}_\alpha, 0^+) - f(\mathbf{x}_\alpha, 0^-)$ is taken to denote the jump across the surface $\mathscr{S}$ of a generic function f defined on the limit configuration obtained as $\varepsilon \to 0$, as schematically outlined in Fig. 1, while it is set $\langle\langle f\rangle\rangle := \frac{1}{2}(f(\mathbf{x}_\alpha, 0^+) + f(\mathbf{x}_\alpha, 0^-))$.

It is worth to point out that all the equations written so far are independent of the constitutive behavior of the material.

2.1.6 Concerning Constitutive Equations

The specific constitutive behavior of the materials is now introduced [38, 53]. In particular, the linearly elastic constitutive laws for the adherents and the interphase, relating the stress with the strain, are given by the equations:

$$\bar{\boldsymbol{\sigma}}^{\varepsilon} = \mathbb{A}_{\pm}(\mathbf{e}(\bar{\mathbf{u}}^{\varepsilon})) \tag{34a}$$

$$\hat{\boldsymbol{\sigma}}^{\varepsilon} = \mathbb{B}^{\varepsilon}(\mathbf{e}(\hat{\mathbf{u}}^{\varepsilon})) \tag{34b}$$

where $A^{\pm}_{ijkl}$, B^{ε}_{ijkl} are the elasticity tensor of the adherents and of the interphase, respectively.

It is worth pointing out that in order to achieve the interface law via this asymptotic approach, the only assumption on the constitutive behavior of constituents, to do necessarily, is that of linear elastic materials. Thereby, no assumption is herein made on the anisotropy of both constituents and on their soundness.

In what follows, within the framework of the imperfect interface approach it has been shown that it is possible to account for different interphase anisotropy conditions and for damage phenomena in the interphase.

In the following section, reference is made to the analysis of interphase behavior, detailing both the soft and hard interphase cases.

2.1.7 Internal/Interphase Analysis

Recalling the seven-regimes distinguish made by Benveniste and Miloh [5] (see Sect. 1), basically two of these typologies of interphase are considered in the present work. The first interphase type, called *soft* interphase, is defined as an interphase material whose elastic properties are linearly rescaled with respect to the interphase thickness ε. The second type, referred as *hard* interphase, is characterized by elastic moduli, which, on the contrary, do not depend on the thickness ε. It is worth pointing out that these hypothesis are referred to the stiffness or the softness of the interphase with respect to the neighboring media (adherents) and it does not depend on the constitutive assumptions (in terms of anisotropy) made on the interphase material. Moreover, the soft interphase behavior is, generally, the simplest constitutive hypothesis made to describe an adhesive layer (e.g. glue). Nevertheless, such an assumption can be an useful strategy in order to take into account for contact zone or thin zones between solids in which interacting phenomena occur.

The *soft interface* definition, as above explained, concerns the capacity to have a non-negligible displacement jump $[[\mathbf{u}]]$ through a surface between two bodies in contact [5, 26, 41], this kind of interface has been also referred as spring-type model. The *hard interface* definition, instead, concerns the capacity to have non-negligible displacement jump $[[\mathbf{u}]]$ and stress jump $[[\boldsymbol{\sigma}]]$ through a surface between two bodies in contact.

The matched asymptotic expansion method applied to soft and hard interphases gives rise to soft and hard interface laws, respectively.

These two cases are relevant for the development of the interface laws classically used in technical problems. Moreover, models of perfect and imperfect interfaces, which are currently used in finite element codes, are known to arise from the hard and the soft interface conditions expanded at the first (zero) order [4, 5, 10, 36, 38]. The interface laws at the higher order, both in the soft and in the hard cases, are object of recent studies [53] which are recalled in the following.

Soft Interphase Analysis:

Assuming that the interphase is soft, let the interphase elasticity tensor $\mathbb{B}^{\varepsilon}$ be defined as [53]:

$$\mathbb{B}^{\varepsilon} = \varepsilon \mathbb{B} \tag{35}$$

where tensor $\mathbb{B}$ does not depend on ε. Referring to Voigt notation rule, its components can be expressed as:

$$K_{ki}^{jl} := B_{ijkl} \tag{36}$$

Taking into account relations (9), (16b) and (35), the stress-strain law (34b) takes the following form:

$$\hat{\boldsymbol{\sigma}}^{0} + \varepsilon \hat{\boldsymbol{\sigma}}^{1} = \mathbb{B}(\hat{\mathbf{e}}^{-1} + \varepsilon \hat{\mathbf{e}}^{0}) + o(\varepsilon) \tag{37}$$

As Eq. (37) is true for any value of ε, the following expressions are derived:

$$\hat{\boldsymbol{\sigma}}^{0} = \mathbb{B}(\hat{\mathbf{e}}^{-1}) \tag{38a}$$

$$\hat{\boldsymbol{\sigma}}^{1} = \mathbb{B}(\hat{\mathbf{e}}^{0}) \tag{38b}$$

Substituting Eq. (36) into Eq. (38a) it results:

$$\hat{\sigma}_{ij}^{0} = B_{ijkl} \hat{e}_{kl}^{-1} = K_{ki}^{jl} \hat{e}_{kl}^{-1} \tag{39}$$

and using Eq. (10), it follows that:

$$\hat{\boldsymbol{\sigma}}^{0} \mathbf{i}_{j} = \mathbf{K}^{3j} \hat{\mathbf{u}}_{,3}^{0} \tag{40}$$

for $j = 1, 2, 3$. Integrating Eq. (40) written for $j = 3$, with respect to z_3, it results:

$$\hat{\boldsymbol{\sigma}}^{0} \mathbf{i}_{3} = \mathbf{K}^{33} \left[\hat{\mathbf{u}}^{0}\right] \tag{41}$$

which represents the classical law for a soft interface at the zero-order.

Recalling a recent study by Rizzoni et al. [53], it is possible to formulate the soft interface law at the one-order. Accordingly, by substituting the expression (36) into (38b) and by using formula (11) written for $k = 0$, one has:

$$\hat{\boldsymbol{\sigma}}^1\mathbf{i}_j = \mathbf{K}^{1j}\hat{\mathbf{u}}^0_{,1} + \mathbf{K}^{2j}\hat{\mathbf{u}}^0_{,2} + \mathbf{K}^{3j}\hat{\mathbf{u}}^1_{,3} \tag{42}$$

for $j = 1, 2, 3$. Moreover, by taking into account formula (40), written for $j = 1, 2$, the equilibrium Eq. (20) explicitly becomes:

$$(\mathbf{K}^{31}\hat{\mathbf{u}}^0_{,3})_{,1} + (\mathbf{K}^{32}\hat{\mathbf{u}}^0_{,3})_{,2} + (\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3)_{,3} = \mathbf{0} \tag{43}$$

and thus, integrating with respect to z_3 between $-\frac{1}{2}$ and $\frac{1}{2}$, it gives:

$$\left[\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3\right] = -\mathbf{K}^{31}\left[\hat{\mathbf{u}}^0\right]_{,1} - \mathbf{K}^{32}\left[\hat{\mathbf{u}}^0\right]_{,2} \tag{44}$$

It is worth remarking that the stress components $\hat{\sigma}^0_{i3}$ (with $i = 1, 2, 3$) are independent of z_3, because of the Eq. (19). Consequently, taking into account Eq. (40) written for $j = 3$, the derivatives $\hat{u}^0_{i,3}$ are also independent of z_3; thus, the displacement components $\hat{u}^0_i$ are a linear functions of z_3. Therefore, Eq. (44) reveals that the stress components $\hat{\sigma}^1_{i3}$, with $i = 1, 2, 3$, are linear functions in z_3, allowing to write the following representation form for the stress components:

$$\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3 = \left[\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3\right]z_3 + \langle\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3\rangle \tag{45}$$

where $\langle f\rangle(\mathbf{z}_\alpha) := \frac{1}{2}\big(f(\mathbf{z}_\alpha, \frac{1}{2}) + f(\mathbf{z}_\alpha, -\frac{1}{2})\big)$. Substituting Eq. (42) written for $j = 3$ into expression (45) and integrating with respect to z_3 it yields:

$$\langle\hat{\boldsymbol{\sigma}}^1\mathbf{i}_3\rangle = \mathbf{K}^{\alpha 3}\langle\hat{\mathbf{u}}^0\rangle_{,\alpha} + \mathbf{K}^{33}\left[\hat{\mathbf{u}}^1\right] \tag{46}$$

where the sum over $\alpha = 1, 2$ is implied. Combining Eqs. (44)–(46), it results:

$$\begin{aligned}\hat{\boldsymbol{\sigma}}^1(\mathbf{z}_\alpha, \pm\frac{1}{2})\mathbf{i}_3 = \mathbf{K}^{33}[\hat{\mathbf{u}}^1](\mathbf{z}_\alpha) + \frac{1}{2}(\mathbf{K}^{\alpha 3} \mp \mathbf{K}^{3\alpha})\hat{\mathbf{u}}^0_{,\alpha}(\mathbf{z}_\alpha, \frac{1}{2})\\ + \frac{1}{2}(\mathbf{K}^{\alpha 3} \pm \mathbf{K}^{3\alpha})\hat{\mathbf{u}}^0_{,\alpha}(\mathbf{z}_\alpha, -\frac{1}{2})\end{aligned} \tag{47}$$

The soft interface laws at zero-order and at one-order, expressed by Eqs. (41) and (47) respectively, have to be formulated in their final form in terms of the stresses and displacements fields in the final configuration (see Fig. 1c). To this aim, using the matching relations (29)–(32), the final formulations of the soft interface laws at zero-order and at one-order, respectively, are the following [53]:

$$\boldsymbol{\sigma}^0(\cdot, 0)\mathbf{i}_3 = \mathbf{K}^{33}[[\mathbf{u}^0]], \tag{48}$$

$$\begin{aligned}\boldsymbol{\sigma}^1(\cdot, 0^\pm)\mathbf{i}_3 = \mathbf{K}^{33}([[\mathbf{u}^1]] + \langle\langle\mathbf{u}^0_{,3}\rangle\rangle) + \frac{1}{2}(\mathbf{K}^{\alpha 3} \mp \mathbf{K}^{3\alpha})\mathbf{u}^0_{,\alpha}(\cdot, 0^+)\\ + \frac{1}{2}(\mathbf{K}^{\alpha 3} \pm \mathbf{K}^{3\alpha})\mathbf{u}^0_{,\alpha}(\cdot, 0^-) \mp \frac{1}{2}\boldsymbol{\sigma}^0_{,3}(\cdot, 0^\pm)\mathbf{i}_3\end{aligned} \tag{49}$$

where the symbol $(\cdot)$ represents the coordinates (x_1, x_2) in a generic point of the system $\Omega_+ \cup \Omega_-$ in the final configuration. In detail, Eq. (48) represents the classical spring-type interface law, derived from an interphase characterized by a finite stiffness. Moreover, Eq. (49) allows to evaluate the stress vector at the higher (one) order, highlighting that the stress vector $\boldsymbol{\sigma}^1(\cdot, 0^{\pm})\mathbf{i}_3$ depends not only on displacement jump at one-order but also on the displacement and stress fields evaluated at the zero-order and their derivatives.

In order to have a complete expression of the effective stress field in the reference configuration (see Fig. 1a), Eqs. (16b) and (7c) must be substituted in Eqs. (48) and (49). Finally, it results:

$$\begin{aligned}\boldsymbol{\sigma}^{\varepsilon}(\cdot, 0^{\pm})\mathbf{i}_3 \approx \mathbf{K}^{33}[[\mathbf{u}^{\varepsilon}]] + \varepsilon\Big(&\mathbf{K}^{33}\langle\langle\mathbf{u}^{\varepsilon}_{,3}\rangle\rangle \\ &+\frac{1}{2}(\mathbf{K}^{\alpha 3} \mp \mathbf{K}^{3\alpha})\mathbf{u}^{\varepsilon}_{,\alpha}(\cdot, 0^{+}) \\ &+\frac{1}{2}(\mathbf{K}^{\alpha 3} \pm \mathbf{K}^{3\alpha})\mathbf{u}^{\varepsilon}_{,\alpha}(\cdot, 0^{-}) \mp \frac{1}{2}\boldsymbol{\sigma}^{\varepsilon}_{,3}(\cdot, 0^{\pm})\mathbf{i}_3\Big)\end{aligned} \tag{50}$$

It is worth remarking that Eq. (50) improves the classic interface law at zero-order by linearly linking the stress vector and the relative displacement via a higher order term, involving the in-plane first derivatives of the displacement. Moreover, (50) allows to clearly quantify the error committed in the interface constitutive equation by modeling a ε-thick layer with a soft interface law at the zero-order (first right-side term in Eq. (50)). In particular, if the in-plane gradient of displacement and/or the out-of-plane gradient of stress are relevant, they can be neglected in the interface constitutive law.

Hard Interphase Analysis:

For a hard interphase, the elasticity tensor $\mathbb{B}^{\varepsilon}$ takes the following form [14, 38, 53]:

$$\mathbb{B}^{\varepsilon} = \mathbb{B} \tag{51}$$

where the tensor $\mathbb{B}$ does not depend on ε, and $\mathbf{K}^{jl}$ is still taken to denote the matrices such that $K^{jl}_{ki} := B_{ijkl}$ (Voigt notation).

Taking into account relations (9) and (16b), the stress-strain Eq. (34b) takes the following form:

$$\hat{\boldsymbol{\sigma}}^0 + \varepsilon\hat{\boldsymbol{\sigma}}^1 = \mathbb{B}(\varepsilon^{-1}\hat{\mathbf{e}}^{-1} + \hat{\mathbf{e}}^0 + \varepsilon\hat{\mathbf{e}}^1) + o(\varepsilon) \tag{52}$$

As Eq. (52) is true for any value of ε, the following conditions are derived:

$$\mathbf{0} = \mathbb{B}(\hat{\mathbf{e}}^{-1}) \tag{53a}$$

$$\hat{\boldsymbol{\sigma}}^{0} = \mathbb{B}(\hat{\mathbf{e}}^{0}) \tag{53b}$$

Taking into account Eq. (10) and the positive definiteness of the tensor $\mathbb{B}$, relation (53a) gives:

$$\hat{\mathbf{u}}^{0}_{,3} = 0 \Rightarrow [\hat{\mathbf{u}}^{0}] = \mathbf{0} \tag{54}$$

which corresponds to the kinematics of the perfect interface.

Substituting Eq. (11) written for $k = 0$ into (53b) one has:

$$\hat{\boldsymbol{\sigma}}^{0}\mathbf{i}_j = \mathbf{K}^{1j}\hat{\mathbf{u}}^{0}_{,1} + \mathbf{K}^{2j}\hat{\mathbf{u}}^{0}_{,2} + \mathbf{K}^{3j}\hat{\mathbf{u}}^{1}_{,3} \tag{55}$$

for $j = 1, 2, 3$. Integrating Eq. (55) written for $j = 3$, with respect to z_3, it results:

$$[\hat{\mathbf{u}}^{1}] = (\mathbf{K}^{33})^{-1}\Big(\hat{\boldsymbol{\sigma}}^{0}\mathbf{i}_3 - \mathbf{K}^{\alpha 3}\hat{\mathbf{u}}^{0}_{,\alpha}\Big) \tag{56}$$

Recalling the Eq. (55) (written for $j = 1, 2$), equilibrium equation Eq. (20) explicitly becomes:

$$(\mathbf{K}^{11}\hat{\mathbf{u}}^{0}_{,1} + \mathbf{K}^{21}\hat{\mathbf{u}}^{0}_{,2} + \mathbf{K}^{31}\hat{\mathbf{u}}^{1}_{,3})_{,1} + (\mathbf{K}^{12}\hat{\mathbf{u}}^{0}_{,1} + \mathbf{K}^{22}\hat{\mathbf{u}}^{0}_{,2} + \mathbf{K}^{32}\hat{\mathbf{u}}^{1}_{,3})_{,2} + (\hat{\boldsymbol{\sigma}}^{1}\mathbf{i}_3)_{,3} = \mathbf{0} \tag{57}$$

and thus, integrating with respect to z_3 between $-1/2$ and $1/2$ and using (56), it gives:

$$\begin{aligned}\Big[\hat{\boldsymbol{\sigma}}^{1}\mathbf{i}_3\Big] &= \Big(- \mathbf{K}^{\alpha\beta}\hat{\mathbf{u}}^{0}_{,\beta} - \mathbf{K}^{3\alpha}[\hat{\mathbf{u}}^{1}]\Big)_{,\alpha} \\ &= \Big(- \mathbf{K}^{\alpha\beta}\hat{\mathbf{u}}^{0}_{,\beta} - \mathbf{K}^{3\alpha}(\mathbf{K}^{33})^{-1}\Big(\hat{\boldsymbol{\sigma}}^{0}\mathbf{i}_3 - \mathbf{K}^{\beta 3}\hat{\mathbf{u}}^{0}_{,\beta}\Big)\Big)_{,\alpha}\end{aligned} \tag{58}$$

with the Greek indexes ($\alpha, \beta = 1, 2$) are related, as usual, to the in-plane (x_1, x_2) quantities.

It is worth noting that in Eq. (58) higher order effects occur and they are related to the appearance of in-plane derivatives, which are usually neglected in the classical first (zero) order theories of interfaces [14, 38, 53]. These *new* terms are related to second-order derivatives and as a consequence, indirectly, to the curvature of the deformed interface. By non-neglecting these terms it is possible to model a membrane effect in the adhesive [53].

In the hard case also, it is possible to derive a final form of the interface laws in terms of the stresses and displacements fields in the final configuration (Fig. 1c). Using matching relations (29)–(32) the interface laws, calculated both at zero-order and at one-order, can be rewritten as follows [14, 53]:

$$[[\mathbf{u}^0]] = \mathbf{0} \tag{59}$$

$$[[\mathbf{u}^1]] = -(\mathbf{K}^{33})^{-1}\Big(\boldsymbol{\sigma}^0\mathbf{i}_3 - \mathbf{K}^{\alpha 3}\mathbf{u}^0_{,\alpha}\Big) - \langle\langle \mathbf{u}^0_{,3}\rangle\rangle \tag{60}$$

$$[[\boldsymbol{\sigma}^0\,\mathbf{i}_3]] = \mathbf{0} \tag{61}$$

$$[[\boldsymbol{\sigma}^1\,\mathbf{i}_3]] = \Big(- \mathbf{K}^{\alpha\beta}\mathbf{u}^0_{,\beta} + \mathbf{K}^{3\alpha}(\mathbf{K}^{33})^{-1}\Big(\boldsymbol{\sigma}^0\mathbf{i}_3 - \mathbf{K}^{\beta 3}\mathbf{u}^0_{,\beta}\Big)\Big)_{,\alpha} - \langle\langle \boldsymbol{\sigma}^0_{,3}\,\mathbf{i}_3\rangle\rangle \tag{62}$$

Equations (59) and (61) represent the classical perfect interface law characterized by the continuity of the displacement and stress vector fields [5]. Additionally, Eqs. (60) and (62) are imperfect interface conditions, allowing jumps in the displacement and in the stress vector fields at the higher (one) order across the interface $\mathscr{S}$ [53]. Moreover, Eqs. (60) and (62) highlight that these jumps depend on the displacement and the stress fields at the zero-order and on their first and second order derivatives [14].

As done in the soft case, the constitutive law for the hard interface written in terms of displacement jumps and stresses in the reference configuration (Fig. 1a) can be derived (with reference to [14, 53]). By considering the expansions (16a) and (7a) combined with Eqs. (59)–(62). The obtained imperfect interface laws reads as:

$$[[\mathbf{u}^\varepsilon]] \approx \; -\varepsilon\Big((\mathbf{K}^{33})^{-1}\Big(\boldsymbol{\sigma}^\varepsilon\mathbf{i}_3 + \mathbf{K}^{\alpha 3}\mathbf{u}^\varepsilon_{,\alpha}\Big) - \langle\langle \mathbf{u}^\varepsilon_{,3}\rangle\rangle\Big) \tag{63}$$

$$\begin{aligned}[[\boldsymbol{\sigma}^\varepsilon\,\mathbf{i}_3]] \approx \varepsilon\Big(\big(- \mathbf{K}^{\alpha\beta}\mathbf{u}^\varepsilon_{,\beta} + \mathbf{K}^{3\alpha}(\mathbf{K}^{33})^{-1}\big(\boldsymbol{\sigma}^\varepsilon\mathbf{i}_3 - \mathbf{K}^{\beta 3}\mathbf{u}^\varepsilon_{,\beta}\big)\big)_{,\alpha} \\ -\langle\langle \boldsymbol{\sigma}^\varepsilon_{,3}\,\mathbf{i}_3\rangle\rangle\Big)\end{aligned} \tag{64}$$

2.2 Homogenization in Non-interacting Approximation (NIA) for Microcracked Media

The class of inhomogeneities considered in the paper is that of planar microcracks, both in the two-dimensional framework (rectilinear cracks) and in the three-dimensional framework (penny-shaped cracks). The considered imperfect interphase $\mathscr{B}^\varepsilon$, defined as the thin layer having $\mathscr{S}$ as the middle section and ε as the uniform small thickness, is weakened by non-interacting penny-shaped microcracks of radius b. Cracks are assumed to be characterized by a periodic transversally isotropic distribution with symmetry axis $\mathbf{i}_3$. Moreover, the non-interacting approximation (NIA) is enforced [57], accordingly, each crack does not experience mechanical interaction by surrounding cracks. Within the NIA framework, one recalls that the microcracks contribution to the material effective properties is obtained as a summation over the contribution of a single crack (or a family of cracks with characteristic length and orientation [34]). As a result, a ε-thick representative elementary volume (REV) of the interphase comprising a single crack can be conveniently introduced as sketched in Fig. 2. Note that in the case of a family of parallel cracks (with same orientation), it is possible to identify this family by an equivalent crack with average radius.

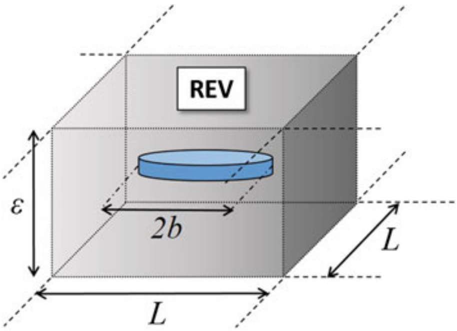


Fig. 2 *REV* with a crack—sketch of the ε-thick representative elementary volume (REV) taken into account in the 3-d homogenization process

The non-interacting approximation is particularly useful for cracked materials, basically, for two reasons:

- it appears to be relatively accurate to high crack density, where local interaction effects become substantial, this evidence can be due to the fact that presence of cracks does not change the average stress in the matrix;
- substantial progress has been made in analyzing shape factors for cracks of complex shapes.

It is worth remarking that, in the following text, for the sake of briefness, the word *crack* is often used instead of *microcracks*; however the whole formulation that will be discussed belongs to a micromechanics framework.

Mathematically, a crack is characterized by a surface of discontinuity experienced by displacements (or temperature) when external fields are applied. Property contribution tensors have rather specific structure for cracks. A complicating factor is that cracks often have *irregular* shapes (including non-planar and intersected configurations). Nevertheless, this shortcoming is not taken into account in the present paper.

NIA formulation have two dual forms, which in the following will be referred as *stress-based approach* and *strain-based approach*. They correspond to obtain the property contribution tensor via a summation of compliance or stiffness contributions of individual inhomogeneities, respectively. In the following sections, the general formulation of these homogenization approaches for microcracked media in the NIA framework is outlined.

2.2.1 Stress-Based Approach

Generally, the additional strain tensor (averaged over the domain Ω of volume V) due to the presence of a pore is given by the following integral over the pore boundary $\partial\Omega_p$:

$$\Delta\varepsilon = -\frac{1}{V}\int_{\partial\Omega_p} (\mathbf{u} \otimes^s \mathbf{n})\, \mathrm{d}S \tag{65}$$

where $\otimes^s$ is the symmetric tensorial product, $\mathbf{u}$ is the displacement vector, $\mathbf{n}$ is a unit normal to $\partial\Omega_p$ directed inward the pore.

Let $\boldsymbol{v}^+$ and $\boldsymbol{v}^-$ be the displacements at the crack boundaries Γ^+ and Γ^- with $\Gamma = \Gamma^+ \cup \Gamma^-$. Denote also as $\mathbf{u}_{cod} = \langle \boldsymbol{v}^+ - \boldsymbol{v}^- \rangle = [\int_\Gamma (\boldsymbol{v}^+ - \boldsymbol{v}^-) \mathrm{d}\Gamma]/|\Gamma|$ the average measure of the displacement jump across the crack, in the following referred to as *crack opening displacement* (COD) vector. Where $|\Gamma|$ is the measure of the crack surface. In this case, Eq. (65) takes the form:

$$\Delta\varepsilon = \frac{1}{V}\int_{\Gamma^+} ([\boldsymbol{v}] \otimes^s \mathbf{n})\, \mathrm{d}S \tag{66}$$

where $\mathbf{n}$ is the normal unit vector of the crack surface and $[\boldsymbol{v}] = (\boldsymbol{v}^+ - \boldsymbol{v}^-)$ is the displacement discontinuity vector along Γ. Calculation of the integral in terms of remotely applied stress $\boldsymbol{\sigma}^0 \equiv \boldsymbol{\sigma}$ would yield the $\mathbb{H}$-tensor of the crack, defined as:

$$\Delta\boldsymbol{\varepsilon} = \frac{V^p}{V}\left(\mathbb{H} : \boldsymbol{\sigma}^0\right) \tag{67}$$

For a flat (planar) crack ($\mathbf{n}$ is constant along Γ), the additional strain $\Delta\varepsilon$ becomes:

$$\Delta\varepsilon = \frac{1}{V}(\mathbf{u}_{cod} \otimes^s \mathbf{n})\, \Gamma \tag{68}$$

Equations (66) and (68) are an immediate consequence of a footnote remark in the famous work by Hill [31].

Let recall that under the approximation of non-interacting cracks, each crack is embedded into the $\boldsymbol{\sigma}$-field and it does not experience any influence of other cracks. As a result, for a flat crack of any shape, a second-rank crack compliance tensor $\mathbf{B}$ can be introduced that relates vector $\mathbf{u}_{cod}$ to the vector of uniform traction $\mathbf{T}_n = \boldsymbol{\sigma} \cdot \mathbf{n}$ induced at the crack site by the far-field $\boldsymbol{\sigma}$ [34, 35, 43, 60]:

$$\mathbf{u}_{cod} = \mathbf{T}_n^T \cdot \mathbf{B} \tag{69}$$

Therefore, according to the hypothesis of linear elasticity of materials and absence of friction along crack faces, the average COD vector for each crack is expressed in terms of the vector of uniform traction $\mathbf{T}_n$.

Since $\mathbf{B}$ is a symmetric tensor (as follows from application of the Betti reciprocity theorem to the normal and shear tractions on a crack), three orthogonal principal directions of the crack compliance exist: application of a uniform traction in one of them does not generate components of vector $\mathbf{u}_{cod}$ in the other two directions. If the matrix is isotropic, $\mathbf{n}$ is one of them and the other two, $\mathbf{t}$ and $\mathbf{s}$, lie in the crack plane, as follows:

$$\mathbf{B} = B_{nn}(\mathbf{n} \otimes \mathbf{n}) + B_{tt}(\mathbf{t} \otimes \mathbf{t}) + B_{ss}(\mathbf{s} \otimes \mathbf{s}) \tag{70}$$

As a definition, let introduce the average, over in-plane directions $\boldsymbol{\tau}$, shear crack compliance that is of importance for the effective elastic properties of a solid with multiple cracks [57]:

$$B_T = \frac{(B_{tt} + B_{ss})}{2} \tag{71}$$

It is worth to remark that the **B** tensor has to be specialized with respect to the bulk material properties.

Within the NIA framework, the problem of *quantitative characterization* of microstructures is reduced to find the proper microstructural parameter of inhomogeneities in whose terms the effective property of interest, compliance tensor has to be expressed [35]. Generally, the concentration parameters of inhomogeneities in the context of the elastic properties are better identified via the structure of the additional elastic potential Δf.

For flat cracks, recalling Eshelby's theory [16], the elastic potential $f(\boldsymbol{\sigma})$ of the effective microcracked material, written in terms of *microstructural* quantities defined on the crack surfaces Γ^i, is [34, 57, 59]:

$$f(\boldsymbol{\sigma}) = f_0(\boldsymbol{\sigma}) + \Delta f = \frac{1}{2}\boldsymbol{\sigma} : \mathbb{S}_0 : \boldsymbol{\sigma} + \frac{1}{2V}\sum_i (\mathbf{T}_n^T \cdot \mathbf{u}_{cod})^i \, \Gamma^i \tag{72}$$

where $f_0(\boldsymbol{\sigma})$ is, as usually, the potential of the bulk matrix (interphase) and the perturbation term Δf is obtained as a sum of the contributions of individual cracks, i.e. $\sum_i$ is a summation over the families of microcracks of length $2\,l^i$ and normal vector $\mathbf{n}^i$. Recall that the tensor $\mathbb{S}_0$ appearing in Eq. (72) is the compliance tensor of the virgin interphase.

In the important case of randomly oriented circular cracks (penny-shaped) of radii b^i, their concentration is characterized by the crack density parameter introduced by Bristow [7]:

$$\rho = \frac{1}{V}\sum_i b^{(i)3} \tag{73}$$

that in the two-dimensional case of randomly oriented rectilinear cracks of mean-length l^i becomes:

$$\rho = \frac{1}{A}\sum_i l^{(i)2} \tag{74}$$

This parameter was generalized by Budiansky and O'Connell [8] to the elliptical in-plane shapes, of areas $S^{(i)}$ and perimeters $P^{(i)}$ (provided aspect ratios of ellipses are identical) as:

$$\rho = \frac{2}{\pi}\frac{1}{V}\sum_i \left(\frac{S^2}{P}\right)^i \tag{75}$$

For non-random crack orientations, the *crack density tensor* was introduced by Kachanov [34]:

$$\boldsymbol{\alpha} = \frac{1}{V}\sum_i (b^3\mathbf{n}\otimes\mathbf{n})^i \qquad \left(\boldsymbol{\alpha} = \frac{1}{A}\sum_i (l^2\mathbf{n}\otimes\mathbf{n})^i \text{ in 2-D case}\right) \tag{76}$$

with $\rho = \mathrm{Tr}\,\boldsymbol{\alpha}$. Kachanov introduced also a fourth-rank density tensor in three-dimensional case

$$\frac{1}{V}\sum_i (b^3\mathbf{n}\otimes\mathbf{n}\otimes\mathbf{n}\otimes\mathbf{n})^i \tag{77}$$

which in general causes a *small* deviation from orthotropy.

As an example, in the three-dimensional case of an isotropic material weakened by open penny-shaped cracks the elastic potential f is [35]:

$$f = f_0 + \frac{8(1-\nu_0^2)}{3\left(1-\frac{\nu_0}{2}\right)E_0}\left[(\boldsymbol{\sigma} * \boldsymbol{\sigma}) : \boldsymbol{\alpha} - \frac{\nu_0}{2}\left(\boldsymbol{\sigma} : \frac{1}{V}\sum_i (b^3\mathbf{n}\otimes\mathbf{n}\otimes\mathbf{n}\otimes\mathbf{n})^i : \boldsymbol{\sigma}\right)\right] \tag{78}$$

2.2.2 Strain-Based Approach

Goidescu-type formulation is developed within the framework of 2-D homogenization problems [20, 21]. It extends the micromechanical approach proposed by Andrieux et al. [2] and leads to a closed-form expression of the macroscopic free energy of a 2D orthotropic elastic medium weakened by arbitrarily oriented microcracks in the dilute limit assumption. It exists a large amount of literature about the homogenization of microcracked media following a strain-based approach [25, 31–33, 47]. It is worth to recall that within the framework of this approach the stiffness contribution tensor $\Delta\mathbb{C}$ are derived starting from a free energy $\mathscr{W}$. As done above for the stress-based approach, let the general background be outlined. Particularly, reference is herein made to the two-dimensional formulation by [20, 21].

Let consider a RVE of total area $\mathscr{A}$, the bulk matrix is assumed to be weakened by an array of $\mathscr{N}$ families of flat microcracks with arbitrary orientation relative to orthotropic axes and mean length $2l^i$, which occupy the domain ω. As a general recall [33], the macroscopic stress $\boldsymbol{\Sigma}$ and strain $\mathbf{E}$ tensors and the macroscopic free energy $\mathscr{W}$ on a cell $\mathscr{A}$ are respectively defined as average values of microscopic stress $\boldsymbol{\sigma}$ and strain $\boldsymbol{\varepsilon}$ fields and local free energy. Let denote by $\bar{\mathscr{A}} = \mathscr{A} - \omega$ the area of the matrix phase, $\mathbf{v}(\mathbf{x})$ the outward unit normal to ω and $\mathbf{T}(\mathbf{x}, \mathbf{v}(\mathbf{x}))$ the traction along the crack faces for any point $\mathbf{x} \in \omega$. Decomposition of local fields over the RVE and application of the divergence theorem allow to relate macroscopic and microscopic quantities [31]. For the macroscopic stress $\boldsymbol{\Sigma}$, one has:

$$\boldsymbol{\Sigma} = \langle\boldsymbol{\sigma}\rangle_{\mathscr{A}} = \langle\boldsymbol{\sigma}\rangle_{\bar{\mathscr{A}}} + \frac{\mathscr{N}}{2}\int_{\omega}\left(\mathbf{T}(\mathbf{x}, \mathbf{v}(\mathbf{x})) \otimes^s \mathbf{x}\right)\,\mathrm{d}x = \langle\boldsymbol{\sigma}\rangle_{\bar{\mathscr{A}}} \tag{79}$$

For the macroscopic strain $\mathbf{E}$ one has, recalling Eq. (66):

$$\mathbf{E} = \langle \boldsymbol{\varepsilon} \rangle_{\mathscr{A}} = \langle \boldsymbol{\varepsilon} \rangle_{\bar{\mathscr{A}}} + \frac{\mathscr{N}}{2} \int_{\omega^+} \big([\boldsymbol{\upsilon}(\mathbf{x})] \otimes^s \mathbf{n} \big) \mathrm{d}x \tag{80}$$

with the surface integral operator $\langle \bullet \rangle_{\mathscr{M}} = \frac{1}{|\mathscr{M}|} \int_{\mathscr{M}} (\bullet) \mathrm{d}S$, and $\mathbf{v}(\mathbf{x}) = \mathbf{n}$ for $\mathbf{x} \in \omega$ the unit vector normal to the cracks, supposed to be constant along ω for flat and regular cracks. It is worth recalling that both Eqs. (66) and (80) are generalization of the Hill lemma for continuous media and they are directly derived from a footnote remark in his work [31]. From Eq. (80) is pointed out that the average strain field on the solid part $\langle \boldsymbol{\varepsilon} \rangle_{\bar{\mathscr{A}}}$ is therefore not sufficient to describe $\mathbf{E}$, the contribution of displacements jump $[\boldsymbol{\upsilon}]$ on the cracks must be taken into account in its expression.

The macroscopic free energy of the material is a finite quantity exclusively defined on the matrix part of the material, that is:

$$\mathscr{W} = \frac{1}{2} \langle \boldsymbol{\varepsilon} : \mathbb{C}_0 : \boldsymbol{\varepsilon} \rangle_{\bar{\mathscr{A}}} \tag{81}$$

For microcracked media it has been established, among other by Telega [58], that the following equation holds:

$$\mathscr{W} = \frac{1}{2} \int_{\partial \bar{\mathscr{A}}} \boldsymbol{\upsilon}(\mathbf{x}) \cdot \boldsymbol{\sigma}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}) \, \mathrm{d}x = \frac{1}{2} \boldsymbol{\Sigma} : \mathbf{E} - \frac{\mathscr{N}}{2} \int_{\omega^+} [\boldsymbol{\upsilon}(\mathbf{x})] \cdot \boldsymbol{\sigma}(\mathbf{x}) \cdot \mathbf{n} \, \mathrm{d}x \tag{82}$$

with $\partial \bar{\mathscr{A}} = \partial \mathscr{A} \cup \omega$ the boundary of the solid matrix.

Let consider an uniform boundary condition applied on the boundary $\partial \mathscr{A}$ of the RVE, given in terms of stresses as follows:

$$\boldsymbol{\sigma}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}) = \boldsymbol{\Sigma} \cdot \mathbf{v}(\mathbf{x}) \ \forall \mathbf{x} \in \partial \mathscr{A} \tag{83}$$

Within the framework of the strain-type approach, in order to derive the local fields involved and to determine the effective microcracks contribution, the elastic problem $\mathscr{P}$ is decomposed into two sub-problems $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(2)}$ [2]:

- in the sub-problem $\mathscr{P}^{(1)}$, the displacement field $\mathbf{u}^{(1)}$ corresponds to that of the homogeneous virgin material subjected to uniform stress conditions; accordingly the related local stress $\boldsymbol{\sigma}^{(1)}$ and strain $\boldsymbol{\varepsilon}^{(1)}$ fields are uniform and must comply with the average stress rule $\boldsymbol{\Sigma} = \langle \boldsymbol{\sigma}^{(1)} \rangle_{\mathscr{A}}$ and $\mathbf{E}^{(1)} = \langle \boldsymbol{\varepsilon}^{(1)} \rangle_{\mathscr{A}} = [\mathbb{C}_0]^{-1} : \boldsymbol{\Sigma}$
- for the sub-problem $\mathscr{P}^{(2)}$, the displacement field $\mathbf{u}^{(2)}$ is induced by the displacement jump $[\boldsymbol{\upsilon}]$ between the crack faces; the related local stress $\boldsymbol{\sigma}^{(2)}$ is in this case self-equilibrated, i.e. $\langle \boldsymbol{\sigma}^{(2)} \rangle_{\bar{\mathscr{A}}} = 0$ from (79); besides, since $\langle \boldsymbol{\varepsilon}^{(2)} \rangle_{\bar{\mathscr{A}}} = [\mathbb{C}_0]^{-1} : \langle \boldsymbol{\sigma}^{(2)} \rangle_{\bar{\mathscr{A}}} = 0$; the macroscopic strain reads from (80):

$$\mathbf{E}^{(2)} = \frac{\mathscr{N}}{2} \int_{\omega^+} ([\boldsymbol{\upsilon}(\mathbf{x})] \otimes \mathbf{n} + \mathbf{n} \otimes [\boldsymbol{\upsilon}(\mathbf{x})]) \mathrm{d}x \tag{84}$$

Introducing two scalar variables β and γ related to the normal $[\mathbf{u}_N(\mathbf{x})] = [\boldsymbol{\upsilon}(\mathbf{x})] \cdot \mathbf{n}$ and tangential $[\mathbf{u}_T(\mathbf{x})] = [\boldsymbol{\upsilon}(\mathbf{x})] \cdot \mathbf{t}$ average displacement jump components on the cracks faces:

$$\beta = \mathscr{N} \int_{\omega^+} [\mathbf{u}_N(\mathbf{x})] \; \mathrm{d}x \qquad \gamma = \mathscr{N} \int_{\omega^+} [\mathbf{u}_T(\mathbf{x})] \; \mathrm{d}x \tag{85}$$

the macroscopic strain in $\mathscr{P}^{(2)}$ reads as:

$$\mathbf{E}^{(2)} = \beta \, \mathbf{n} \otimes \mathbf{n} + \frac{\gamma}{2}(\mathbf{n} \otimes \mathbf{t} + \mathbf{t} \otimes \mathbf{n}) \tag{86}$$

where $(\mathbf{n}, \mathbf{t})$ define an integral orthonormal basis for the crack.

According with the decomposition, the overall macroscopic strain is:

$$\mathbf{E} = \mathbf{E}^{(1)} + \mathbf{E}^{(2)} \tag{87}$$

Moreover, the overall free energy per unit surface $\mathscr{W}$ defined by Eq. (81) with $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{u}^{(1)} + \mathbf{u}^{(2)})$ can be expressed as the sum of two terms [2]:

$$\mathscr{W} = \frac{1}{2}\langle(\boldsymbol{\varepsilon}^{(1)} + \boldsymbol{\varepsilon}^{(2)}) : \mathbb{C}_0 : (\boldsymbol{\varepsilon}^{(1)} + \boldsymbol{\varepsilon}^{(2)})\rangle_{\bar{\mathscr{A}}} = \mathscr{W}^{(1)} + \mathscr{W}^{(2)} \tag{88}$$

for which have been taken into account the uniformity of $\boldsymbol{\varepsilon}^{(1)}$ and the property $\langle\boldsymbol{\varepsilon}^{(2)}\rangle_{\bar{\mathscr{A}}} = 0$.

$\mathscr{W}^{(1)}$ is the free energy of the virgin material related to the problem $\mathscr{P}^{(1)}$:

$$\mathscr{W}^{(1)} = \frac{1}{2}\langle\boldsymbol{\varepsilon}^{(1)} : \mathbb{C}_0 : \boldsymbol{\varepsilon}^{(1)}\rangle_{\mathscr{A}} = \frac{1}{2}\mathbf{E}^{(1)} : \mathbb{C}_0 : \mathbf{E}^{(1)} \tag{89}$$

and the terms $\mathscr{W}^{(2)}$ is related to the contribution of the jump displacement in problem $\mathscr{P}^{(2)}$. It follows from Eq. (82) that:

$$\begin{aligned}
\mathscr{W}^{(2)} &= \frac{1}{2}\langle\boldsymbol{\varepsilon}^{(2)} : \mathbb{C}_0 : \boldsymbol{\varepsilon}^{(2)}\rangle_{\bar{\mathscr{A}}} = -\frac{\mathscr{N}}{2}\int_{\omega^+} [\boldsymbol{\upsilon}(\mathbf{x})] \cdot \boldsymbol{\sigma}^{(2)}(\mathbf{x}) \cdot \mathbf{n} \, \mathrm{d}x \\
&= -\frac{\mathscr{N}}{2}\int_{\omega^+} ([\mathbf{u}_N(\mathbf{x})] \, \mathbf{n} \cdot \boldsymbol{\sigma}^{(2)}(\mathbf{x}) \cdot \mathbf{n} + [\mathbf{u}_T(\mathbf{x})] \, \mathbf{n} \cdot \boldsymbol{\sigma}^{(2)}(\mathbf{x}) \cdot \mathbf{t}) \, \mathrm{d}x \\
&= -\frac{1}{2}(\beta \mathbf{n} \cdot \boldsymbol{\sigma}^{(2)} \cdot \mathbf{n} + \gamma \mathbf{n} \cdot \boldsymbol{\sigma}^{(2)} \cdot \mathbf{t})
\end{aligned} \tag{90}$$

for which: $\boldsymbol{\sigma}^{(2)}(\mathbf{x}) = \boldsymbol{\sigma}^{(2)} = const, \; \forall \mathbf{x} \in \omega$ for a dilute concentration of cracks.

Final expression of the free energy $\mathscr{W}$ of the microcracked material with open cracks parallel to $\mathbf{i}_1$ direction (for further details refer to [20, 21]) is:

$$\mathscr{W} = \mathscr{W}_0 - d\left[H_{nn}\,(\mathbf{N} : \mathbf{E})^2 + H_{tt}\,(\mathbf{T} : \mathbf{E})^2\right] \tag{91}$$

with

$$\mathscr{W}_0 = \frac{1}{2}\mathbf{E} : \mathbb{C}_0 : \mathbf{E} \tag{92}$$

be the overall free energy of the virgin initially-orthotropic material and $d = \mathscr{N} l^2$ be the microcracks density, where $\mathscr{N}$ is the number of cracks per unit surface area, and as usual, l is the half-length of a crack. Parameters $H_{nn} = C(1 + D)$ and $H_{tt} = C(1 - D)$ are identical to B_{nn} and B_{tt} respectively, of the stress-based approach of Kachanov type. Constants C and D can be expressed in terms of engineering mechanical parameters, or equivalently, in terms of the components of tensor $\mathbb{C}_0$ [21]. Moreover, second-order tensors $\mathbf{N} = \mathbb{C}_0 : (\mathbf{i}_3 \otimes \mathbf{i}_3)$ and $\mathbf{T} = \mathbb{C}_0 : (\mathbf{i}_1 \otimes^s \mathbf{i}_3)$ are used.

Finally, from Eq. (91) the effective stiffness tensor $\mathbb{C}$ of the microcracked material, is obtained.

It is worth noting that all the obtained coefficients K_{ki}^{jl} are of the form $f(\mathbb{C}_0) - d\,[g(\mathbb{C}_0)]$ with f, g generic functions. It is well highlighted a shortcoming of this kind of formulation in dilute limit assumption, that is severely limit values of the microcracks density d. Nevertheless, a great advantage of such a homogenization can be leading to coefficients which do not depend on the REV geometry.

From a computational point of view, the implementation of a hard interface model, also for a quite simple geometry, is not an easy issue due to the discontinuities both in terms of stresses and displacements at the interface. This aspect is not considered in the present work, nevertheless some numerical results are given in [14].

3 A St. Venant-Kirchhoff Imperfect Interface Model

In this section, a nonlinear-imperfect interface model is proposed. Within the framework of the detailed micromechanical approach, the model is formulated following the same procedure detailed in Sect. 2, in order to derive both soft and hard imperfect interface laws. In detail, the matched asymptotic expansion method [1, 4, 38–41, 52, 53] is extended to the finite strain theory [14, 15, 49]. Moreover, the homogenization method for microcracked media under the NIA [34, 35, 43, 57, 60] is applied to a damaged interphase comprising of a hyperelastic St. Venant-Kirchhoff initially orthotropic material [49].

3.1 Matched Asymptotic Expansion Method in Finite Strains

Let an orthonormal Cartesian basis $(O, \mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3)$ be introduced, with x_1, x_2 and x_3 be the corresponding coordinates of a particle belonging to the system Ω^ε. Refer to Sect. 2 for the notation and three-dimensional problem statement (see Fig. 1).

The equations governing the equilibrium problem of such a composite system are expressed as follows:

$$\begin{cases} (s^\varepsilon_{ij} + s^\varepsilon_{kj} u^\varepsilon_{i,k})_{,j} + f_i = 0 & \text{in } \Omega^\varepsilon_\pm \\ (s^\varepsilon_{ij} + s^\varepsilon_{kj} u^\varepsilon_{i,k}) n_j = p_i & \text{on } \Gamma_1 \\ (s^\varepsilon_{ij} + s^\varepsilon_{kj} u^\varepsilon_{i,k})_{,j} = 0 & \text{in } \mathscr{B}^\varepsilon \\ [[s^\varepsilon_{i3} + s^\varepsilon_{k3} u^\varepsilon_{i,k}]] = 0 & \text{on } \mathscr{S}^\varepsilon_\pm \\ [[u^\varepsilon_i]] = 0 & \text{on } \mathscr{S}^\varepsilon_\pm \\ u^\varepsilon_i = 0 & \text{on } \Gamma_0 \\ s^\varepsilon_{ij} = A^\pm_{ijhk} E_{hk}(u^\varepsilon) & \text{in } \Omega^\varepsilon_\pm \\ s^\varepsilon_{ij} = A^\varepsilon_{ijhk} E_{hk}(u^\varepsilon) & \text{in } \mathscr{B}^\varepsilon \end{cases} \tag{93}$$

where $\mathbf{s}^\varepsilon$ is the second Piola-Kirchhoff stress tensor, $\mathbf{E}(u^\varepsilon)$ is the Green-Lagrange strain tensor ($E_{ij}(u^\varepsilon) = \frac{1}{2}(u_{i,j} + u_{j,i} + u_{k,i}u_{k,j})$ with $i, j = 1, 2, 3$) and $\mathbb{A}^\pm$, $\mathbb{A}^\varepsilon$ are the elasticity tensors of the deformable adherents and of the adhesive, respectively. It is worth remarking that for the elastic tensor $\mathbb{A}^\varepsilon$ holds the following identity $\mathbb{A}^\varepsilon \equiv \mathbb{B}^\varepsilon$. Additionally, by the homogenization for microcracked media, detailed in the next section, the interphase elastic tensor is found to be consistent with the soft interphase assumption. Such a finding, allows to express its components through the following relationship:

$$A^\varepsilon_{ijkl} = \varepsilon \hat{A}_{ijkl} \tag{94}$$

Since the interphase is assumed to behave as a thin layer of thickness ε, it is natural to seek the solution of the equilibrium problem, expressed by Eq. (93), by using asymptotic expansions with respect to the small parameter ε [42]. In particular, the following asymptotic series with fractional powers are exploited [54]:

$$\begin{cases} \mathbf{u}^\varepsilon(x_1, x_2, x_3) = \mathbf{u}^0 + \varepsilon^{1/3}\mathbf{u}^1 + \varepsilon^{2/3}\mathbf{u}^2 + \varepsilon\ \mathbf{u}^3 + \varepsilon^{4/3}\mathbf{u}^4 + \varepsilon^{5/3}\mathbf{u}^5 + \varepsilon^2\mathbf{u}^6 + o(\varepsilon^2) \\ \mathbf{s}^\varepsilon(x_1, x_2, x_3) = \mathbf{s}^0 + \varepsilon^{1/3}\mathbf{s}^1 + \varepsilon^{2/3}\mathbf{s}^2 + \varepsilon\ \mathbf{s}^3 + \varepsilon^{4/3}\mathbf{s}^4 + \varepsilon^{5/3}\mathbf{s}^5 + \varepsilon^2\mathbf{s}^6 + o(\varepsilon^2) \end{cases} \tag{95}$$

It is worth remarking that such a choice of a fractional expansion is due to energy-based evidences [54]. In particular, from a quite simple mono-dimensional example, proposed in [54], it has been put in evidence that the solution in terms of displacement jump is proportional to $\varepsilon^{\frac{2}{3}}$.

In agreement with [12] and equivalently to what performed in the others models (see Sect. 2), also in this case, let the change of variable $\hat{\mathbf{g}} : (x_1, x_2, x_3) \rightarrow (z_1, z_2, z_3)$ be introduced in $\mathscr{B}^\varepsilon$, with $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_3/\varepsilon$. Moreover, let the change of variable $\bar{\mathbf{g}} : (x_1, x_2, x_3) \rightarrow (z_1, z_2, z_3)$ be introduced in $\Omega^\varepsilon_\pm$, with $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_3 \pm (1 - \varepsilon)/2$. As a result, the interphase $\mathscr{B}^\varepsilon$ and the adherents $\Omega^\varepsilon_\pm$ are scaled in domains of unitary thickness $\mathscr{B}$ and $\Omega_\pm$, respectively. In what follows, symbols $\bar{\cdot}$ and $\hat{\cdot}$ refer to rescaled quantities for $\mathscr{B}$ and $\Omega_\pm$, respectively. More precisely, $\hat{\mathbf{u}}^\varepsilon = \mathbf{u}^\varepsilon \circ \hat{\mathbf{g}}^{-1}$ and $\hat{\mathbf{s}}^\varepsilon = \mathbf{s}^\varepsilon \circ \hat{\mathbf{g}}^{-1}$ denote displacement and stress fields for $\mathscr{B}$, and $\bar{\mathbf{u}}^\varepsilon = \mathbf{u}^\varepsilon \circ \bar{\mathbf{g}}^{-1}$ and $\bar{\mathbf{s}}^\varepsilon = \mathbf{s}^\varepsilon \circ \bar{\mathbf{g}}^{-1}$ are displacement vector and stress tensor for $\Omega_\pm$, $\mathbf{u}^\varepsilon$ and $\mathbf{s}^\varepsilon$ being the corresponding fields on the system Ω^ε. The internal and external forces, $\mathbf{f}$ and $\mathbf{p}$, respectively, are assumed to be independent of ε. As a consequence, it is set $\bar{\mathbf{f}}(z_1, z_2, z_3) = \mathbf{f}(x_1, x_2, x_3)$ and $\bar{\mathbf{p}}(z_1, z_2, z_3) = \mathbf{p}(x_1, x_2, x_3)$. Moreover, under the change of variables, the domains Γ_0 and Γ_1 are transformed into the domains denoted by $\bar{\Gamma}_0$ and $\bar{\Gamma}_1$, respectively. As a result, the governing equations of the equilibrium problem, in the rescaled composite system, are expressed as follows:

$$\begin{cases} (\bar{s}_{ij} + \bar{s}_{kj}\bar{u}_{i,k})_{,j} + \bar{f}_i = 0 & \text{in } \Omega_\pm \\ (\bar{s}_{ij} + \bar{s}_{kj}\bar{u}_{i,k})n_j = \bar{p}_i & \text{on } \bar{\Gamma}_1 \\ (\hat{s}_{i\alpha} + \hat{s}_{k\alpha}\hat{u}_{i,k})_{,\alpha} + \frac{1}{\varepsilon}(\hat{s}_{i3} + \hat{s}_{k3}\hat{u}_{i,k})_{,3} = 0 & \text{in } \mathscr{B} \\ \bar{s}_{i3} + \bar{s}_{k3}\bar{u}_{i,k} = \hat{s}_{i3} + \hat{s}_{\alpha 3}\hat{u}_{i,\alpha} + \frac{1}{\varepsilon}\hat{s}_{33}\hat{u}_{i,3} & \text{on } \mathscr{S}_\pm \\ \bar{u}_i = \hat{u}_i & \text{on } \mathscr{S}_\pm \\ \bar{u}_i = 0 & \text{on } \bar{\Gamma}_0 \\ \bar{s}_{ij} = A^{\pm}_{ijhk}\bar{E}_{hk}(\bar{u}) & \text{in } \Omega_\pm \\ \hat{s}_{ij} = A^{\varepsilon}_{ijhk}\hat{E}_{hk}(\hat{u}) & \text{in } \mathscr{B} \end{cases} \tag{96}$$

where $\bar{\mathbf{E}}$, $\hat{\mathbf{E}}$ denote the rescaled Green-Lagrange strain tensors in the adherents and in the adhesive.

In view of Eq. (95) the relevant fields, in the rescaled adhesive and adherents, can be expressed as asymptotic expansions in the following way:

$$\begin{cases} \hat{\mathbf{s}}^\varepsilon(z_1, z_2, z_3) = \hat{\mathbf{s}}^0 + \varepsilon^{1/3}\hat{\mathbf{s}}^1 + \varepsilon^{2/3}\hat{\mathbf{s}}^2 + \varepsilon\hat{\mathbf{s}}^3 + \varepsilon^{4/3}\hat{\mathbf{s}}^4 + \varepsilon^{5/3}\hat{\mathbf{s}}^5 + \varepsilon^2\hat{\mathbf{s}}^6 + o(\varepsilon^2) \\ \bar{\mathbf{s}}^\varepsilon(z_1, z_2, z_3) = \bar{\mathbf{s}}^0 + \varepsilon^{1/3}\bar{\mathbf{s}}^1 + \varepsilon^{2/3}\bar{\mathbf{s}}^2 + \varepsilon\bar{\mathbf{s}}^3 + \varepsilon^{4/3}\bar{\mathbf{s}}^4 + \varepsilon^{5/3}\bar{\mathbf{s}}^5 + \varepsilon^2\bar{\mathbf{s}}^6 + o(\varepsilon^2) \\ \hat{\mathbf{u}}^\varepsilon(z_1, z_2, z_3) = \hat{\mathbf{u}}^0 + \varepsilon^{1/3}\hat{\mathbf{u}}^1 + \varepsilon^{2/3}\hat{\mathbf{u}}^2 + \varepsilon\hat{\mathbf{u}}^3 + \varepsilon^{4/3}\hat{\mathbf{u}}^4 + \varepsilon^{5/3}\hat{\mathbf{u}}^5 + \varepsilon^2\hat{\mathbf{u}}^6 + o(\varepsilon^2) \\ \bar{\mathbf{u}}^\varepsilon(z_1, z_2, z_3) = \bar{\mathbf{u}}^0 + \varepsilon^{1/3}\bar{\mathbf{u}}^1 + \varepsilon^{2/3}\bar{\mathbf{u}}^2 + \varepsilon\bar{\mathbf{u}}^3 + \varepsilon^{4/3}\bar{\mathbf{u}}^4 + \varepsilon^{5/3}\bar{\mathbf{u}}^5 + \varepsilon^2\bar{\mathbf{u}}^6 + o(\varepsilon^2) \end{cases} \tag{97}$$

In the following, the conditions holding in the rescaled interphase $\mathscr{B}$ are detailed. These latter are obtained by substituting the first of Eq. (97) into the equilibrium equation holding in the interphase (i.e., third equation of the system (96)) and by identifying, in a standard way, similar terms with respect to the power of the parameter ε:

- Power of $\varepsilon : -2$

$$(\hat{u}^0_{i,3}\hat{s}^0_{33})_{,3} = 0, \tag{98}$$

- Power of $\varepsilon : -5/3$

$$(\hat{u}^0_{i,3}\hat{s}^1_{33} + \hat{u}^1_{i,3}\hat{s}^0_{33})_{,3} = 0, \tag{99}$$

- Power of $\varepsilon : -4/3$

$$(\hat{u}^0_{i,3}\hat{s}^2_{33} + \hat{u}^1_{i,3}\hat{s}^1_{33} + \hat{u}^2_{i,3}\hat{s}^0_{33})_{,3} = 0, \tag{100}$$

- Power of $\varepsilon : -1$

$$\begin{aligned} &(\hat{u}^0_{i,3}\hat{s}^0_{3\alpha})_{,\alpha} + (\hat{s}^0_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^0_{3\alpha})_{,3} \\ &\quad + (\hat{u}^0_{i,3}\hat{s}^3_{33} + \hat{u}^1_{i,3}\hat{s}^2_{33} + \hat{u}^2_{i,3}\hat{s}^1_{33} + \hat{u}^3_{i,3}\hat{s}^0_{33})_{,3} = 0, \end{aligned} \tag{101}$$

- Power of $\varepsilon : -2/3$

$$
\begin{aligned}
&(\hat{u}^0_{i,3}\hat{s}^1_{3\alpha} + \hat{u}^1_{i,3}\hat{s}^0_{3\alpha})_{,\alpha} + (\hat{s}^1_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^1_{3\alpha} + \hat{u}^1_{i,\alpha}\hat{s}^0_{3\alpha})_{,3} \\
&\quad + (\hat{u}^0_{i,3}\hat{s}^4_{33} + \hat{u}^1_{i,3}\hat{s}^3_{33} + \hat{u}^2_{i,3}\hat{s}^2_{33} + \hat{u}^3_{i,3}\hat{s}^1_{33} + \hat{u}^4_{i,3}\hat{s}^0_{33})_{,3} = 0, \qquad (102)
\end{aligned}
$$

- Power of $\varepsilon : -1/3$

$$
\begin{aligned}
&(\hat{u}^0_{i,3}\hat{s}^2_{3\alpha} + \hat{u}^1_{i,3}\hat{s}^1_{3\alpha} + \hat{u}^2_{i,3}\hat{s}^0_{3\alpha})_{,\alpha} \\
&\quad + (\hat{s}^2_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^2_{3\alpha} + \hat{u}^1_{i,\alpha}\hat{s}^1_{3\alpha} + \hat{u}^2_{i,\alpha}\hat{s}^0_{3\alpha})_{,3} \\
&\quad + (\hat{u}^0_{i,3}\hat{s}^5_{33} + \hat{u}^1_{i,3}\hat{s}^4_{33} + \hat{u}^2_{i,3}\hat{s}^3_{33} + \hat{u}^3_{i,3}\hat{s}^2_{33} + \hat{u}^4_{i,3}\hat{s}^1_{33} + \hat{u}^5_{i,3}\hat{s}^0_{33})_{,3} = 0, \qquad (103)
\end{aligned}
$$

- Power of $\varepsilon : 0$

$$
\begin{aligned}
&(\hat{u}^0_{i,3}\hat{s}^3_{3\alpha} + \hat{u}^1_{i,3}\hat{s}^2_{3\alpha} + \hat{u}^2_{i,3}\hat{s}^1_{3\alpha} + \hat{u}^3_{i,3}\hat{s}^0_{3\alpha})_{,\alpha} + (\hat{s}^0_{i\alpha} + \hat{s}^0_{\alpha\beta}\hat{u}^0_{i,\beta})_{\alpha} \\
&\quad + (\hat{s}^3_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^3_{3\alpha} + \hat{u}^1_{i,\alpha}\hat{s}^2_{3\alpha} + \hat{u}^2_{i,\alpha}\hat{s}^1_{3\alpha} + \hat{u}^3_{i,\alpha}\hat{s}^0_{3\alpha})_{,3} \\
&\quad + (\hat{u}^0_{i,3}\hat{s}^6_{33} + \hat{u}^1_{i,3}\hat{s}^5_{33} + \hat{u}^2_{i,3}\hat{s}^4_{33} + \hat{u}^3_{i,3}\hat{s}^3_{33} + \hat{u}^4_{i,3}\hat{s}^2_{33} + \hat{u}^5_{i,3}\hat{s}^1_{33} + \hat{u}^6_{i,3}\hat{s}^0_{33})_{,3} = 0, \\
&\hfill (104)
\end{aligned}
$$

- …

By substituting the first two equations of (97) into the continuity condition of the traction vector holding through the rescaled interfaces $\mathscr{S}_{\pm}$ (i.e., fourth equation of system (96)), and by applying the usual identification procedure, the following relationships are obtained:

- Power of $\varepsilon : -1$

$$
0 = (\hat{u}^0_{i,3}\hat{s}^0_{33}) \qquad (105)
$$

- Power of $\varepsilon : -2/3$

$$
0 = (\hat{u}^0_{i,3}\hat{s}^1_{33} + \hat{u}^1_{i,3}\hat{s}^0_{33}) \qquad (106)
$$

- Power of $\varepsilon : -1/3$

$$
0 = (\hat{u}^0_{i,3}\hat{s}^2_{33} + \hat{u}^1_{i,3}\hat{s}^1_{33} + \hat{u}^2_{i,3}\hat{s}^0_{33}) \qquad (107)
$$

- Power of $\varepsilon : 0$

$$
(\bar{s}^0_{i3} + \bar{u}^0_{ik}\bar{s}^0_{k3}) = (\hat{s}^0_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^0_{\alpha 3} + \hat{u}^0_{i,3}\hat{s}^3_{33} + \hat{u}^1_{i,3}\hat{s}^2_{33} + \hat{u}^2_{i,3}\hat{s}^1_{33} + \hat{u}^3_{i,3}\hat{s}^0_{33}) \qquad (108)
$$

- Power of ε : $1/3$

$$\begin{aligned}&(\bar{s}^1_{i3} + \bar{u}^0_{i,k}\bar{s}^1_{k3} + \bar{u}^1_{i,k}\bar{s}^0_{k3})\\ &\quad = (\hat{s}^1_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^1_{\alpha3} + \hat{u}^1_{i,\alpha}\hat{s}^0_{\alpha3})\\ &\qquad + (\hat{u}^0_{i,3}\hat{s}^4_{33} + \hat{u}^1_{i,3}\hat{s}^3_{33}\\ &\qquad + \hat{u}^2_{i,3}\hat{s}^2_{33} + \hat{u}^3_{i,3}\hat{s}^1_{33} + \hat{u}^4_{i,3}\hat{s}^0_{33})\end{aligned} \tag{109}$$

- Power of ε : $2/3$

$$\begin{aligned}&(\bar{s}^2_{i3} + \bar{u}^0_{i,k}\bar{s}^2_{k3} + \bar{u}^1_{i,k}\bar{s}^1_{k3} + \bar{u}^2_{i,k}\bar{s}^0_{k3})\\ &\quad = (\hat{s}^2_{33} + \hat{u}^0_{i,\alpha}\hat{s}^2_{\alpha3} + \hat{u}^1_{i,\alpha}\hat{s}^1_{\alpha3} + \hat{u}^0_{i,\alpha}\hat{s}^2_{\alpha3})\\ &\qquad + (\hat{u}^0_{i,3}\hat{s}^5_{33} + \hat{u}^1_{i,3}\hat{s}^4_{33} + \hat{u}^2_{i,3}\hat{s}^3_{33} + \hat{u}^3_{i,3}\hat{s}^2_{33} + \hat{u}^4_{i,3}\hat{s}^1_{33} + \hat{u}^5_{i,3}\hat{s}^0_{33})\end{aligned} \tag{110}$$

- Power of ε : 1

$$\begin{aligned}&(\bar{s}^3_{i3} + \bar{u}^0_{i,k}\bar{s}^3_{k3} + \bar{u}^1_{i,k}\bar{s}^2_{k3} + \bar{u}^2_{i,k}\bar{s}^1_{k3} + \bar{u}^3_{i,k}\bar{s}^0_{k3})\\ &\quad = (\hat{s}^3_{i3} + \hat{u}^0_{i,\alpha}\hat{s}^3_{\alpha3} + \hat{u}^1_{i,\alpha}\hat{s}^2_{\alpha3} + \hat{u}^2_{i,\alpha}\hat{s}^1_{\alpha3} + \hat{u}^3_{i,\alpha}\hat{s}^0_{\alpha3})\\ &\qquad + (\hat{u}^0_{i,3}\hat{s}^6_{33} + \hat{u}^1_{i,3}\hat{s}^5_{33} + \hat{u}^2_{i,3}\hat{s}^4_{33} + \hat{u}^3_{i,3}\hat{s}^3_{33} + \hat{u}^4_{i,3}\hat{s}^2_{33} + \hat{u}^5_{i,3}\hat{s}^1_{33} + \hat{u}^6_{i,3}\hat{s}^0_{33})\end{aligned} \tag{111}$$

- ...

It is worth noting that the above equations hold both in $\mathscr{S}_+$ and in $\mathscr{S}_-$, for the sake of briefness they have been detailed only in one case. Moreover, by remarking that the left-hand sides in Eqs. (108)–(111) can be identified as the expansions of the $\mathbf{i}_3$ components of the first Piola-Kirchhoff stress tensor $\bar{P}_{i3} = (\bar{s}_{i3} + \bar{u}_{ik}\bar{s}_{k3})$ in the adherents, a significant simplification of these equations it is possible.

According to the soft-material-interphase assumption, by substituting Eq. (94) in the constitutive law holding in the interphase $\mathscr{B}$ of the rescaled domain (i.e., last equation of the problem (96)), written for $j = 3$, the following conditions are deduced:

- Power of ε : -1

$$0 = (\hat{u}^0_{k,3}\hat{u}^0_{k,3}) \tag{112}$$

- Power of ε : $-2/3$

$$0 = (\hat{u}^0_{k,3}\hat{u}^1_{k,3} + \hat{u}^1_{k,3}\hat{u}^0_{k,3}) \tag{113}$$

- Power of ε : $-1/3$

$$0 = (\hat{u}^0_{k,3}\hat{u}^2_{k,3} + \hat{u}^1_{k,3}\hat{u}^1_{k,3} + \hat{u}^2_{k,3}\hat{u}^0_{k,3}) \tag{114}$$

- Power of ε : 0

$$
\begin{aligned}
\hat{s}^0_{\alpha 3} &= \hat{A}_{33\alpha 3}\left[\hat{u}^0_{3,3} + (\hat{u}^0_{s,3}\hat{u}^3_{s,3} + \hat{u}^1_{s,3}\hat{u}^2_{s,3})\right] + \frac{1}{2}\hat{A}_{\beta 3\alpha 3}(\hat{u}^0_{\beta,3} + \hat{u}^0_{s,\beta}\hat{u}^0_{s,3}) \\
\hat{s}^0_{33} &= \hat{A}_{3333}\left[\hat{u}^0_{3,3} + (\hat{u}^0_{s,3}\hat{u}^3_{s,3} + \hat{u}^1_{s,3}\hat{u}^2_{s,3})\right] + \frac{1}{2}\hat{A}_{33\beta 3}(\hat{u}^0_{\beta,3} + \hat{u}^0_{s,\beta}\hat{u}^0_{s,3})
\end{aligned}
\tag{115}
$$

- Power of ε : 1/3

$$
\begin{aligned}
\hat{s}^1_{\alpha 3} =&\hat{A}_{33\alpha 3}\left[\hat{u}^1_{3,3} + (\hat{u}^0_{s,3}\hat{u}^4_{s,3} + \hat{u}^1_{s,3}\hat{u}^3_{s,3} + \frac{1}{2}\hat{u}^2_{s,3}\hat{u}^2_{s,3})\right] \\
&+ \frac{1}{2}\hat{A}_{\beta 3\alpha 3}(\hat{u}^1_{\beta,3} + \hat{u}^0_{s,\beta}\hat{u}^1_{s,3} + \hat{u}^1_{s,\beta}\hat{u}^0_{s,3}) \\
\hat{s}^1_{33} =&\hat{A}_{3333}\left[\hat{u}^1_{3,3} + (\hat{u}^0_{s,3}\hat{u}^4_{s,3} + \hat{u}^1_{s,3}\hat{u}^3_{s,3} + \frac{1}{2}\hat{u}^2_{s,3}\hat{u}^2_{s,3})\right] \\
&+ \frac{1}{2}\hat{A}_{33\beta 3}(\hat{u}^1_{\beta,3} + \hat{u}^0_{s,\beta}\hat{u}^1_{s,3} + \hat{u}^1_{s,\beta}\hat{u}^0_{s,3})
\end{aligned}
\tag{116}
$$

- ...

From Eqs. (112)–(115) it follows that:

$$\hat{\mathbf{u}}^0_{,3} = 0 \quad \text{in } \mathscr{B} \Rightarrow [\hat{\mathbf{u}}^0] = 0 \tag{117}$$

$$\hat{\mathbf{u}}^1_{,3} = 0 \quad \text{in } \mathscr{B} \Rightarrow [\hat{\mathbf{u}}^1] = 0 \tag{118}$$

$$\hat{s}^0_{\alpha 3} = 0 = \hat{s}^0_{33} \quad \text{in } \mathscr{B} \tag{119}$$

where it is set $[f](z_1, z_2) = f(z_1, z_2, 1/2) - f(z_1, z_2, -1/2)$ for $f : \mathscr{B} \mapsto \mathbb{R}^3$. By combining Eqs. (117)–(119) into Eqs. (98)–(101), the following relationship is obtained:

$$(\hat{u}^2_{i,3}\hat{s}^1_{33})_3 = 0 \quad \text{in } \mathscr{B} \tag{120}$$

which integrated with respect to z_3 gives

$$\hat{u}^2_{i,3}\hat{s}^1_{33} = const. = \bar{P}^0_{i3}|_{\mathscr{S}_\pm} \quad \text{in } \mathscr{B} \tag{121}$$

where $\bar{P}^0_{i3}|_{\mathscr{S}_\pm}$ is the common value taken at the interfaces $\mathscr{S}_\pm$ (cfr. Eq. (108)). Moreover, by substituting Eq. (116) and Eqs. (117)–(119) into Eq. (121) the following relationship is obtained:

$$\frac{1}{2}\hat{A}_{3333}(\mid \hat{u}^2_{i,3} \mid^2 \hat{u}^2_{i,3}) = \bar{P}^0_{i3} = \text{in } \mathscr{B} \tag{122}$$

By solving with respect to $\hat{\mathbf{u}}^2_{,3}$ and by integrating with respect to z_3 one has:

$$[\hat{\mathbf{u}}^2] = \frac{1}{(\frac{1}{2}\hat{A}_{3333})^{1/3}} \frac{1}{\mid \bar{\mathbf{P}}^0\mathbf{i}_3 \mid^{2/3}} \bar{\mathbf{P}}^0\mathbf{i}_3 \tag{123}$$

Thereby, substituting Eqs. (112)–(115) into Eqs. (108)–(110) it is obtained that:

$$[\bar{\mathbf{P}}^0\mathbf{i}_3] = 0 \tag{124}$$
$$[\bar{\mathbf{P}}^1\mathbf{i}_3] = 0 \tag{125}$$
$$[\bar{\mathbf{P}}^2\mathbf{i}_3] = 0 \tag{126}$$

with

$$\bar{P}^0_{i3} = (\bar{s}^0_{i3} + \bar{u}^0_{i,k}\bar{s}^0_{k3}) \tag{127}$$
$$\bar{P}^1_{i3} = (\bar{s}^1_{i3} + \bar{u}^0_{i,k}\bar{s}^1_{k3} + \bar{u}^1_{i,k}\bar{s}^0_{k3}) \tag{128}$$
$$\bar{P}^2_{i3} = (\bar{s}^2_{i3} + \bar{u}^0_{i,k}\bar{s}^2_{k3} + \bar{u}^1_{i,k}\bar{s}^1_{k3} + \bar{u}^2_{i,k}\bar{s}^0_{k3}) \tag{129}$$

The final step of the asymptotic expansion method consists in applying the matching conditions in order to find the proper interface law for the limit equilibrium problem, in which the interphase is replaced by the limit interface $\mathscr{S}$ and the adherents by the domains $\Omega^0_\pm = \{(x_1, x_2, x_3) \in \Omega : \pm x_3 > 0\}$. By taking into account the asymptotic expansion of the displacement field (95) and assuming that $\mathbf{u}^\varepsilon$ in the adherent can be expanded in a Taylor series representation along the x_3-direction, it results:

$$\begin{aligned} \mathbf{u}^\varepsilon(\bar{\mathbf{x}}, \pm\frac{\varepsilon}{2}) &= \mathbf{u}^\varepsilon(\bar{\mathbf{x}}, 0^\pm) \pm \frac{\varepsilon}{2}\mathbf{u}^\varepsilon_{,3}(\bar{\mathbf{x}}, 0^\pm) + \cdots \\ &= \mathbf{u}^0(\bar{\mathbf{x}}, 0^\pm) + \varepsilon^{1/3}\mathbf{u}^1(\bar{\mathbf{x}}, 0^\pm) + \varepsilon^{2/3}\mathbf{u}^2(\bar{\mathbf{x}}, 0^\pm) \\ &\quad + \varepsilon\big(\mathbf{u}^3(\bar{\mathbf{x}}, 0^\pm) \pm \frac{1}{2}\mathbf{u}^0_{,3}(\bar{\mathbf{x}}, 0^\pm)\big) + \cdots \end{aligned} \tag{130}$$

In view of the continuity of the displacements at the interfaces $\mathscr{S}^\varepsilon_\pm$ and $\mathscr{S}_\pm$ one has

$$\begin{aligned} &\mathbf{u}^0(\bar{\mathbf{x}}, 0^\pm) + \varepsilon^{1/3}\mathbf{u}^1(\bar{\mathbf{x}}, 0^\pm) + \varepsilon^{2/3}\mathbf{u}^2(\bar{\mathbf{x}}, 0^\pm) + \cdots \\ &= \hat{\mathbf{u}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2}) + \varepsilon^{1/3}\hat{\mathbf{u}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2}) + \cdots \\ &= \bar{\mathbf{u}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2}) + \varepsilon^{1/3}\bar{\mathbf{u}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2}) + \cdots \end{aligned} \tag{131}$$

and, identifying the terms in the same powers of ε, it is deduced that:

$$\begin{aligned}
\mathbf{u}^0(\bar{\mathbf{x}}, 0^{\pm}) &= \hat{\mathbf{u}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2}) = \bar{\mathbf{u}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2}) \\
\mathbf{u}^1(\bar{\mathbf{x}}, 0^{\pm}) &= \hat{\mathbf{u}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2}) = \bar{\mathbf{u}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2}) \\
\mathbf{u}^2(\bar{\mathbf{x}}, 0^{\pm}) &= \hat{\mathbf{u}}^2(\bar{\mathbf{z}}, \pm\frac{1}{2}) = \bar{\mathbf{u}}^2(\bar{\mathbf{z}}, \pm\frac{1}{2})
\end{aligned} \tag{132}$$

Analogous results can be obtained for the tractions vector, herein expressed in terms of the first Piola-Kirchhoff tensor:

$$\begin{aligned}
\mathbf{P}^0(\bar{\mathbf{x}}, 0^{\pm})\mathbf{i}_3 &= \hat{\mathbf{P}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3 = \bar{\mathbf{P}}^0(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3 \\
\mathbf{P}^1(\bar{\mathbf{x}}, 0^{\pm})\mathbf{i}_3 &= \hat{\mathbf{P}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3 = \bar{\mathbf{P}}^1(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3 \\
\mathbf{P}^2(\bar{\mathbf{x}}, 0^{\pm})\mathbf{i}_3 &= \hat{\mathbf{P}}^2(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3 = \bar{\mathbf{P}}^2(\bar{\mathbf{z}}, \pm\frac{1}{2})\mathbf{i}_3
\end{aligned} \tag{133}$$

Let the following notation be adopted: $[[\mathbf{f}]] := \mathbf{f}(\mathbf{x}, 0^+) - \mathbf{f}(\mathbf{x}, 0^-)$ with $\mathbf{f} : \Omega^0_+ \cup \Omega^0_- \mapsto \mathbb{R}^3$; accordingly, the proper contact conditions for the limit equilibrium problem, i.e. expressed in terms of the relevant fields defined on $\Omega^0_+ \cup \Omega^0_-$, can be obtained by using this relation into the interphase laws Eqs. (117), (118), (123), (124)–(126):

$$\begin{aligned}
[\bar{\mathbf{u}}^l] &= [[\mathbf{u}^l]] \quad l = 0, 1, 2 \\
[\bar{\mathbf{P}}^l\mathbf{i}_3] &= [[\mathbf{P}^l\mathbf{i}_3]] \quad l = 0, 1, 2
\end{aligned} \tag{134}$$

By applying the matching relations (134) and taking into account Eqs. (132) and (133), the interface laws for the soft interphase can be rewritten in the limit configuration $\Omega^0_+ \cup \Omega^0_- \cup \mathscr{S}$ in a form involving only the fields in the adherents:

$$[[\mathbf{u}^0]] = 0 \quad [[\mathbf{P}^0\mathbf{i}_3]] = 0 \tag{135}$$

$$[[\mathbf{u}^1]] = 0 \quad [[\mathbf{P}^1\mathbf{i}_3]] = 0 \tag{136}$$

$$[[\mathbf{u}^2]] = \frac{1}{(\frac{1}{2}\hat{A}_{3333})^{1/3}} \frac{1}{|\mathbf{P}^0\mathbf{i}_3|^{2/3}}\mathbf{P}^0\mathbf{i}_3 \quad [[\mathbf{P}^2\mathbf{i}_3]] = 0 \tag{137}$$

which are the final expressions of the interface conditions for the proposed St. Venant-Kirchhoff anisotropic model. It is worth remarking that the imperfect interface condition, prescribing a jump of the displacement, appears at the second order. By taking into account the expansions (95) and the relations (127)–(129), one finds:

$$\mathbf{P}^{\varepsilon}\mathbf{i}_3 = \mathbf{P}^0\mathbf{i}_3 + O(\varepsilon^{1/3}) \tag{138}$$

$$[[\mathbf{P}^{\varepsilon}\mathbf{i}_3]] = \varepsilon^{2/3}[[\mathbf{P}^2\mathbf{i}_3]] + O(\varepsilon) \tag{139}$$

$$[[\mathbf{u}^{\varepsilon}]] = \varepsilon^{2/3}[[\mathbf{u}^2]] + O(\varepsilon) \tag{140}$$

which, substituted into (135)–(137), give

$$[[\mathbf{P}^{\varepsilon}\mathbf{i}_3]] = 0 + o(\varepsilon) \tag{141}$$

$$\mathbf{P}^{\varepsilon}\mathbf{i}_3 = \frac{A^{\varepsilon}_{3333}}{2\,\varepsilon^3}\,|[[\mathbf{u}^{\varepsilon}]]|^2\,[[\mathbf{u}^{\varepsilon}]] + o(\varepsilon^{1/3}) \tag{142}$$

Note that, to the aim to fully express the interface law Eq. (142) within the interphase domain, Eq. (94) is taken into account: $\hat{A}_{3333} = \frac{A^{\varepsilon}_{3333}}{\varepsilon}$.

Finally, the imperfect interface law can be rewritten in terms of the Piola stress vector $\mathbf{P}\,\mathbf{i}_3$ and the displacement jump $[[\mathbf{u}]]$ in the limit configuration (Fig. 1c):

$$\mathbf{P}\,\mathbf{i}_3 = \frac{\hat{A}_{3333}}{2\,\varepsilon^2}\,|[[\mathbf{u}]]|^2\,[[\mathbf{u}]] \tag{143}$$

Remark that Eq. (143) is the relevant expression, from a computational point of view, of the interface law for the proposed St. Venant-Kirchhoff model. Thereby, it represents the transmission condition for the stress vector $\mathbf{P}\,\mathbf{i}_3$ across the interface $\mathscr{S}$. As a definition, $\hat{A}_{3333}$ is the interface stiffness for this soft nonlinear interface. Moreover, Eq. (143) highlights that this transmission condition is an effectively nonlinear imperfect interface law. It is worth remarking that such an imperfect interface condition, in order to be numerically implemented needs to fix a value for the thickness ε. This fact can represent a shortcoming for the proposed model. Nevertheless, such a parameter, in many cases can be measurable, for instance in the case of the glue layers in bonding problems.

3.2 *Homogenization of the Microcracked Interphase*

The interface law (see Eq. 142) is a function of the elastic constant A^{ε}_{3333} of the interphase material. This latter is assumed to be orthotropic with principal axes $(\mathbf{i}_1, \mathbf{i}_3)$ and weakened by one family of parallel rectilinear microcracks with length $2l$ and orientation $\phi = (\mathbf{i}_1, \mathbf{t}) = 0°$. In order to recover the elastic constant A^{ε}_{3333} and consecutively the interface stiffness $\hat{A}_{3333}$, the homogenization technique for microcracked orthotropic media in the two-dimensional context, refer to [17] for further details. Therefore, in what follows only the relevant relations are outlined.

Recall the expression of the effective compliance tensor $\mathbb{S}$ of the microcracked interphase, obtained through a stress-based homogenization in NIA:

$$(\mathbb{S})_{ijkl} = (\mathbb{S}_0)_{ijkl} + (\Delta\mathbb{S})_{ijkl} \tag{144}$$

where $\varDelta\mathbb{S}$ is the contribution compliance tensor associated to the perturbative term in the complementary elastic potential $\varDelta f$, and accounting for the crack features. $\mathbb{S}_0$ is the compliance of the undamaged initially orthotropic interphase material. As a result, the elasticity tensor $\mathbb{A}^\varepsilon$ can be easily derived as: $\mathbb{A}^\varepsilon = \mathbb{S}^{-1}$. Note that the tensor $\mathbb{A}^\varepsilon$ depends on the interphase thickness ε through the microcrack density ρ:

$$\rho = \frac{l^2}{|REA|} = \frac{l^2}{\varepsilon\, L} \tag{145}$$

with L a characteristic dimension of the interphase, and ε is the interphase thickness, as usual. As a result, the elastic constant A^ε_{3333} reads as:

$$A^\varepsilon_{3333} = \frac{E_1 E_3}{E_1 + 2B_{nn}\rho E_1 E_3 - E_3 \nu_{13}^2} \tag{146}$$

with

$$B_{nn} = \frac{\pi}{2\sqrt{E_3}} \sqrt{\frac{2}{\sqrt{E_1 E_3}} + \frac{1}{G_{13}} - \frac{2\nu_{13}}{E_1}} \tag{147}$$

where E_1, E_3, G_{13} and ν_{13} are the elastic constants of the undamaged interphase. It is worth pointing out that these latter can be obtained in terms of the elastic properties of the two adherents, as the result of a homogenization step performed on the undamaged ε-thick representative elementary volume [17, 50, 51].

Finally, the interface stiffness $\hat{A}_{3333}$, derived from Eq. (146), is expressed by the following relationship:

$$\hat{A}_{3333} = \frac{L}{2\, B_{nn}\, l^2} \tag{148}$$

4 Numerical Applications

In this section a numerical benchmark is proposed in order to validate the imperfect-nonlinear-interface model of the St. Venant-Kirchhoff type detailed in Sect. 3 and to compare it with the spring-like model described in Sect. 2. A quite simple three-dimensional geometry is treated, in particular an unit brick (210 mm × 100 mm × 50 mm) joined with a mortar joint (210 mm × 100 mm × 10 mm). The composite system is assumed to be fixed on a flat rigid plane. The geometry and the boundary conditions are outlined in Fig. 3. This simple academic model has been chosen to focus, in a more accurate way, on the behavior of the brick/mortar interface.

With respect to the hypothesis on the constitutive behavior of the principal constituents, i.e. brick and mortar, a linear and a nonlinear isotropic cases have been distinguished. In the first case, the materials are assumed to be linearly elastic with parameters: Young modulus $E_b = 13 \times 10^3$ MPa and Poisson ratio $\nu_b = 0.2$ for the brick, and Young modulus $E_m = 4 \times 10^3$ MPa and Poisson ratio $\nu_m = 0.2$ for the

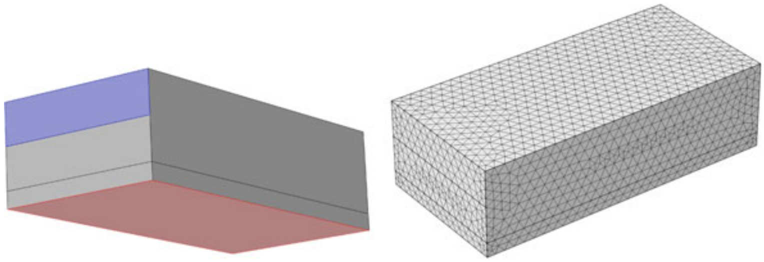


Fig. 3 Geometry, boundary conditions and mesh detail—sketch of the three-dimensional model (on the *left*). The surface loaded with the incremental displacement is represented in *blue* and the *red* surface is fixed. On the *right* side, a detail of the free *tetrahedral* mesh is represented (color online)

mortar, respectively. Instead, in the nonlinear case, both brick and mortar behave as hyperelastic materials of the St. Venant-Kirchhoff type with Lamé constants: $\lambda_b = 3.6 \times 10^3$ MPa and $\mu_b = 5.4 \times 10^3$ MPa for the brick, and $\lambda_m = 1.1 \times 10^3$ MPa and $\mu_m = 1.6 \times 10^3$ MPa for the mortar, respectively.

The interphase, localized at the brick/mortar interface level, is assumed to be a thin stratified layer comprising of brick and mortar material characteristics. In all numerical models proposed, the interphase is treated via the imperfect interface approach. In other words, it is a third material supposed to be initially-transversely isotropic, whose elastic constants E_1, E_3, G_{13} and ν_{13} are derived starting from the mechanical properties of the constituents (E_b, E_m, ν_b, ν_m). In order to obtain its elastic constants a preliminary standard homogenization for stratified is performed on the undamaged ε-thick representative elementary volume [48, 50, 51]. Moreover, this interphase is assumed to be microcracked.

As a result, the following effective elastic constants for the virgin interphase material are obtained: $E_1 = 8.5 \times 10^3$ MPa, $E_3 = 6.3 \times 10^3$ MPa, $G_{13} = 5 \times 10^3$ MPa and $\nu_{13} = 0.2$.

Two imperfect interface models are taken into account, the nonlinear St. Venant-Kirchhoff interface, and the soft interface model obtained in Sect. 2. Note that it is possible to applicate them in a three-dimensional context under the hypothesis of isotropic interface (i.e. the tangential interface stiffness is assumed to be isotropic in the interface plane). The nonlinear imperfect interface is modeled according Eq. (149):

$$\mathbf{P}\,\mathbf{i}_3 = \frac{\hat{A}_{3333}}{2\,\varepsilon^2}\,|[[\mathbf{u}]]|^2\,[[\mathbf{u}]] \tag{149}$$

in which the interface stiffness $\hat{A}_{3333}$ is given by Eq. (148). Assuming the following values: $L = 210$ mm, $l = L/100 = 2$ mm and $\varepsilon = 0.2$ mm, the stiffness results in: $\hat{A}_{3333} = \frac{L}{2\,B_{nn}\,l^2} = 5.9 \times 10^4$ N/mm^3.

Concerning the linear-interface case, let the imperfect interface law be recalled:

$$\boldsymbol{\sigma}\mathbf{i}_3 = \mathbf{K}^{33}[[\mathbf{u}]] \tag{150}$$

where the interface stiffnesses in normal and in tangential-to-the-interface directions comprised in the two-rank matrix $\mathbf{K}^{33}$, are expressed by Eq. (151):

$$K_T = \frac{L}{B_{tt} l^2}, \qquad K_N = \frac{L}{2\, B_{nn} l^2} \tag{151}$$

accordingly they result in: $K_N = 5.9 \times 10^4$ N/mm^3 and $K_T = 1.4 \times 10^5$ N/mm^3.

Several numerical simulations are carried out, aimed at validating the proposed interface models; in particular, in the model with the linearly elastic constituents, in what follows denoted as *linear model*, both the linear and the nonlinear interface laws, are implemented. In the *nonlinear model*, i.e. the one with the hyperelastic St. Venant-Kirchhoff constituents, only the nonlinear interface law is enforced. Additionally, the linear and the nonlinear models are also implemented with perfect interface condition (i.e. $[[\mathbf{u}]] = 0,\ [[\boldsymbol{\sigma}\mathbf{i}_3]] = 0,\ [[\mathbf{P}\mathbf{i}_3]] = 0$), in order to have some reference data, in the following they are referred as LP and NLP, respectively. Moreover, in what follows the linear model with linear interface will be called L^2, the linear model with nonlinear interface will be called LNL and the nonlinear model with nonlinear interface will be called NL^2.

All analysis are performed with the software COMSOL Multiphysics® 4.3 on a processor Intel(R) Core(TM) i3-2350M 2.3 GHz CPU. A free tetrahedral mesh of fine size is chosen in all the analysis cases for the whole domain, moreover, the brick/mortar interface is modeled through interface finite elements of zero thickness, as represented in Fig. 3. The implemented numerical models aim to reproduce a push-out test on a single brick in a quasi-static loading process. The tests are performed in displacement-controlled mode with an imposed displacement of a maximum value equal to 5 mm. The degrees of freedom and the solution times expressed in seconds are summarized in Table 1, for all the analysis cases.

It is worth noting that it could be possible to reduce the degrees of freedom and, consequently, the CPU times, by applying some considerations about the geometrical symmetries of the considered system. Nevertheless, the remarkable aspect is the large difference in terms of CPU time among linear and nonlinear calculations, independently if the nonlinearity is localized at the interface level (LNL) or in the constituents (NL^2). Moreover, both in the linear and in the nonlinear model, the introduction of the linear and nonlinear-imperfect-interface conditions, i.e. L^2 and

Table 1 Values of degrees of freedom (dof) and solution times (in seconds) for all the analyzed numerical models

Model	dof	CPU time (s)
LP	2,13,621	309
L^2	2,19,822	326
LNL	2,19,822	9716
NLP	2,13,621	8831
NL^2	2,19,822	10,317

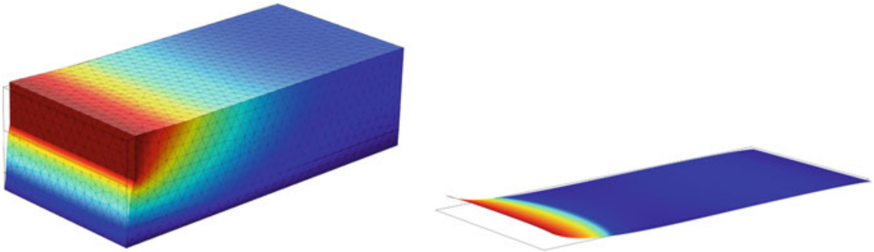


Fig. 4 Deformed shape in LNL model—final deformed shape relative to LNL model (on the *right*). The x_1-component of the displacement field is mapped in *colors*. Final deformed shape of the interface in the same model (on the *left*) with color map of the x_3-component of the displacement-jump vector, the maximum value of the displacement is 0.12 mm (a factor scale of 5 is applied) (color online)

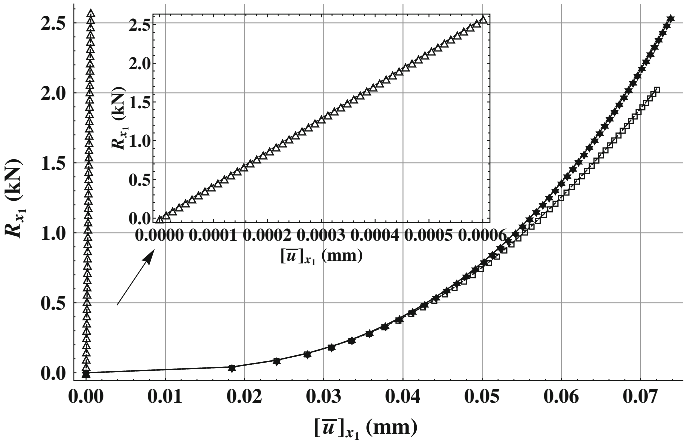


Fig. 5 Comparison in terms of reaction force—reaction force in the x_1-direction averaged over the loaded boundary surface versus the x_1-component of the average displacement-jump vector, at the final step (the maximum value of the imposed displacement is 5 mm). Comparison among: L^2 (- - △ - -); LNL (–★–); and NL^2 (- - □ - -). A zoom of the curve relative to L^2 is represented

NL^2 respectively, does not produce a significant increment of the CPU times with respect to the perfect-interface cases.

The numerical simulations stop when the imposed displacement reaches is maximum value (5 mm). In Fig. 4 a deformed shape at the final configuration is shown and the distribution of the displacement field is color-mapped.

The curves shown in Fig. 5 represent the x_1-component of the reaction force (i.e., in the acting direction of the imposed displacement) averaged on the loaded boundary, plotted with respect to the x_1-component of the displacement jump averaged over the interface surface, for all the analyzed cases. Interestingly, both models LNL and NL^2, allow to take into account for larger deformations (about one order of magnitude) at the interface level, than the L^2 model.

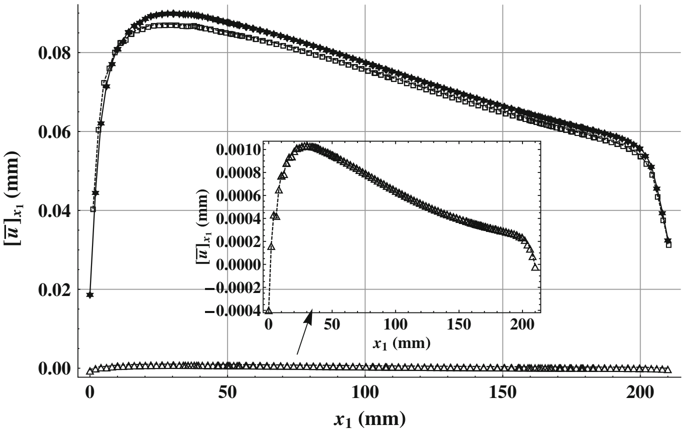


Fig. 6 Comparison in terms of displacement jump in x_1-direction—final distribution of the x_1-component of the average displacement-jump vector along the interface in the x_1-direction (recall that the maximum value of the imposed displacement is 5 mm). Comparison among: L^2 (- - △ - -); LNL (–★–); and NL^2 (- - □ - -). A zoom of the curve relative to L^2 is represented

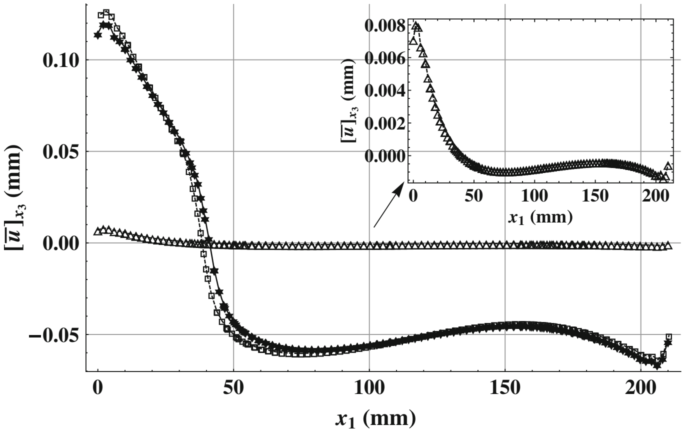


Fig. 7 Comparison in terms of displacement jump in x_3-**direction**—final distribution of the x_3-component of the average displacement-jump vector along the interface in the x_1-direction (recall that the maximum value of the imposed displacement is 5 mm). Comparison among: L^2 (- - △ - -); LNL (–★–); and NL^2 (- - □ - -). A zoom of the curve relative to L^2 is represented

A comparison of Figs. 6 and 7 put in evidence this aspect. The figures represent the distribution, at the final configuration, of the x_1-component and of the x_3-component of the displacement-jump vector, respectively, along a cut line obtained from the intersection of the interface plane with the plane of symmetry.

Furthermore, Figs. 5, 6 and 7 highlight that is not very useful to model the adherents as hyperelastic materials in order to take into account the geometrical nonlinearities, i.e. large deformations, in terms of global response. In fact, the implementation

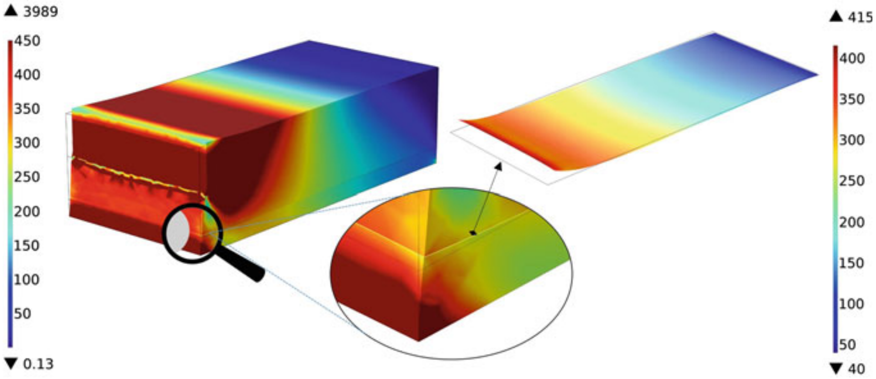


Fig. 8 Von Mises stresses in LNL model—Von Mises stress (MPa) in LNL, with a particular of the stress distribution at the interface level

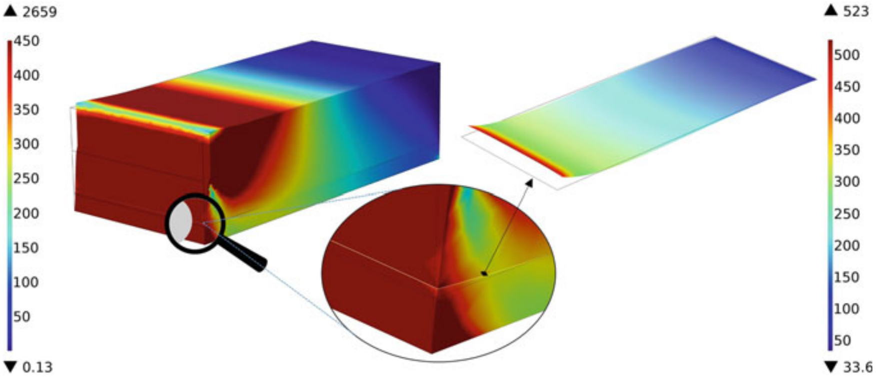


Fig. 9 Von Mises stresses in NL^2 model—Von Mises stress (MPa) in NL^2, with a particular of the stress distribution at the interface level

of a nonlinear imperfect interface, as the proposed St. Venant Kirchhoff model, in a linearly elastic composite system (LNL), seems to sufficiently catch the nonlinear-interface behavior as the fully nonlinear model (NL^2), reaching the same order of magnitude in terms of displacement jumps.

Figures 8 and 9 show the distribution of the Von Mises stresses in LNL and in NL^2 respectively, in both cases a detail of the interface zone is represented. It is worth noting the difference in terms of magnitude of the stresses. In particular, in NL^2 model the Von Mises stresses are significantly smaller than in LNL. Moreover, by analyzing the particular of the interfaces in both model, a significant difference in terms of stress distribution can be appreciated.

5 Conclusions

In the first part of the present paper, the principal tools on which the *imperfect interface approach* is founded, have been introduced. After a brief recall on modeling background of imperfect interfaces, the matched asymptotic expansion method and the homogenization for microcracked media in NIA framework, have been extensively detailed.

The matched asymptotic expansion formulation, based on a higher order theory [53], has been formulated for both soft and hard interface cases. Thereby, the interface laws until the second (one) order have been derived, in both soft and hard interface conditions [53]. Such an asymptotic method, within the imperfect interface approach, is coupled to another tool, that is a micromechanical homogenization technique. In particular, a homogenization for microcracked media in the NIA framework has been chosen in order to take into account for damage in interphase. Two dual approaches in NIA have been presented, the stress-based and the strain-based approach.

In the hard imperfect interface model, the matched asymptotic technique has been expanded until the order one, within the higher order theory framework, recovering an imperfect interface law in terms of stresses and displacements jumps. This resulting interface law, is a challenging issue from a computational point of view. Moreover, a homogenization technique in the strain-based approach under the hypothesis of *dilute concentration*, is adopted [20, 21]. This homogenization technique leads to an expression of the effective elastic coefficients of the type: $f(\mathbb{C}_0) - d\,[g(\mathbb{C}_0)]$ with f, g generic functions. From this expression, it is well highlighted that the values of density d are severely limited. It is worth remarking that there exist other homogenization techniques in dilute concentration approximation that overcome this shortcoming, for instance the dilute estimate scheme by [6], for which the stiffness coefficients, in the initially-isotropic (E^0, ν^0) interphase case, are given by:

$$\left\{\begin{aligned} C_\rho &= \frac{3(2\nu^0-1)+16((1-\nu^0)^2)\rho}{3(2\nu^0-1)+32(\nu^0)^2(1-\nu^0)\rho} \\ E_1^1 &= C_\rho E^0, \\ E_3^1 &= E^0, \\ \mu_{13}^1 &= C_\rho \mu^0, \\ \nu_{31}^1 &= C_\rho \nu^0. \end{aligned}\right. \tag{152}$$

Nevertheless, a great advantage of the Goidescu homogenization can be to lead to coefficients which do not depend on the REV geometry, because of the chosen form of the microstructural parameter d. For the St. Venant-Kirchhoff type interface (Sect. 3), a new matched asymptotic technique, based on fractional expansions of the relevant fields, has been proposed. This asymptotic procedure has been formulated by extending the asymptotic method to the finite strain theory [49, 54]. Also in this case, a homogenization has been performed to treat the microcracked interphase, i.e. the NIA building block in a stress-based approach.

Finally, a simple three-dimensional benchmark is proposed, in which three modeling cases have been compared in order to validate the proposed models.

The chosen application domain is masonry. The first two models, defined as *linear*, have been conceived with linearly elastic adherents (brick and mortar), and with two different interface conditions. In the first case, the brick/mortar interface has been modeled with the linear spring-type interface law, and in the second case, the St. Venant-Kirchhoff nonlinear interface law has been implemented. The third model, defined as *nonlinear*, is a fully nonlinear one, in which the adherents are assumed to be St. Venant-Kirchhoff hyperelastic material and the interface has been modeled with the St. Venant-Kirchhoff nonlinear interface law. Some comparisons have been carried out in terms of displacement jumps and of stresses distribution along the interface. The soundness and the consistency of the proposed interface models are highlighted, both from a theoretical and a numerical points of view. Moreover, it has been established that the linear model with the nonlinear interface is able to catch the large displacements occurring at the interface level as much as the fully nonlinear model, additionally, the computational cost in the first case is smaller.

References

1. R. Abdelmoula, M. Coutris, and J.J. Marigo. **Asymptotic behavior of an elastic thin layer**. *C. R. Acad. Sci.. Serie II F. B-Mec. Phys. Astr.*, **326**(4):237–242, 1998.
2. S. Andrieux, Y. Bamberger, and J.J. Marigo. **A model of micro-cracked material for concretes and rocks**. *J. Méc. Theor. Appl.*, **5**(3):471–513, 1986.
3. Y. Benveniste. **The effective mechanical behaviour of composite materials with imperfect contact between the constituents**. *Mech. Mat.*, **4**(2):197–208, 1985.
4. Y. Benveniste. **A general interface model for a three-dimensional curved thin anisotropic interphase between two anisotropic media**. *J. Mech. Phys. Solid*, **54**(4):708–734, 2006.
5. Y. Benveniste and T. Miloh. **Imperfect soft and stiff interfaces in two-dimensional elasticity**. *Mech. Mat.*, **33**(6):309–323, 2001.
6. P.G. Bornert, T. Bretheau, and P. Gilormini. *Homogénéisation en mécanique des matériaux, Tome 1 : Matériaux aléatoires élastiques et milieux périodiques*. Hermes Sciences, Paris, 2001.
7. J.R. Bristow. **Microcracks, and the Static and Dynamic Elastic Constants of Annealed and Heavily Cold-worked Metals**. *Br. J. Appl. Phys.*, **11**(2):81–85, 1960.
8. B. Budiansky and R.J. O'Connel. **Elastic moduli of a cracked solid**. *Int. J. Solid Struct.*, **12**:81–97, 1976.
9. P. Bövik. **On the modelling of thin interface layers in elastic and acoustic scattering problems**. *Quart. J. Mech. Appl. Math.*, **47**(1):17–42, 1994.
10. D. Caillerie. **Behavior at limit of a thin inclusion of high stiffness in an elastic body**. *C. R. Hebd. Seances Acad. Sci. Serie A*, **287**(8):675–678, 1978.
11. P.G. Ciarlet. **Recent Progress in the Two-Dimensional Approximation of Three - Dimensional Plate Models in Nonlinear Elasticity**. In Eduardo L. Ortiz, editor, *Numerical Approximation of Partial Differential Equations Selection of Papers Presented at the International Symposium on Numerical Analysis held at the Polytechnic University of Madrid*, **133** of *North-Holland Mathematics Studies*, pages 3–19. North-Holland, 1987.
12. P.G. Ciarlet. **Mathematical Elasticity**. In Philippe G. Ciarlet, editor, *Mathematical Elasticity Volume II: Theory of Plates*, **27** of *Studies in Mathematics and Its Applications*, pages vii–xi. Elsevier, 1997.
13. P.G. Ciarlet and P. Destuynder. **A justification of a nonlinear model in plate theory**. *Comput. Meth. Appl. Mech. Eng.*, **17–18, Part 1**(0):227–258, 1979.

14. S. Dumont, F. Lebon, and R. Rizzoni. **An asymptotic approach to the adhesion of thin stiff films**. *Mech. Res. Commun.*, **58**(0):24–35, 2014.
15. S. Dumont, F. Lebon, M.L. Raffa, and R. Rizzoni. **Towards nonlinear imperfect interface models including micro-cracks and smooth roughness**. *Annal Solid Struct. Mech.*, In press.
16. J.D. Eshelby. **Elastic inclusions and inhomogeneities.** In *Progr. Solid mech.*, **2**, page 87–140. North-Holland, Amsterdam, 1961.
17. F. Fouchal, F. Lebon, M.L. Raffa, and G. Vairo. **An interface model including cracks and roughness applied to masonry**. *Open Civ. Eng. J.*, **8**:263–271, 2014.
18. G. Geymonat and F. Krasucki. **Analyse asymptotique du comportement en flexion de deux plaques collées**. *C. R. Acad. Sci. - Series IIB - Mech.-Phys.-Chem.-Astr.*, **325**(6):307–314, 1997.
19. Y. Gilibert and A. Rigolot. **Asymptotic analysis of double adhesive bonded joints loaded in shear tension (in French)**. *J. Méc. Appl.*, **3**(3):341–372, 1979.
20. C. Goidescu. *Caractérisation et modélisation de l'endommagement par microfissuration des composites stratifiés - Apports des mesures de champs et de l'homogénéisation*. PhD thesis, Institut National Polytechnique de Toulouse (INP Toulouse), 2011.
21. C. Goidescu, H. Welemane, D. Kondo, and C. Gruescu. **Microcracks closure effects in initially orthotropic materials**. *Eur. J. Mech. A-Solid*, **37**:172–184, 2013.
22. M. Goland and E. Reissner. **The stresses in cemented joints**. *J. Appl. Mech.*, **11**: A17–A27, 1944.
23. S.T. Gu and Q.C. He. **Interfacial discontinuity relations for coupled multifield phenomena and their application to the modeling of thin interphases as imperfect interfaces**. *J. Mech. Phys. Solid*, **59**(7):1413–1426, 2011.
24. S.T. Gu. *Contributions to the modelling of imperfect interfaces and to the homogenization of heterogeneous materials. Ph.D. Thesis (in French)*. PhD thesis, Université Paris-Est Marne-la-Vallée, France, 2008.
25. Z. Hashin. **The differential scheme and its application to cracked materials**. *J. Mech. Phys. Solid*, **36**(6):719–734, 1988.
26. Z. Hashin. **Thermoelastic properties of fiber composites with imperfect interface**. *Mech. Mat.*, **8**(4):333–348, 1990.
27. Z. Hashin. **Thermoelastic properties of particulate composites with imperfect interface**. *J. Mech. Phys. Solid*, **39**(6):745–762, 1991.
28. Z. Hashin. **The Spherical Inclusion With Imperfect Interface**. *J. Appl. Mech.*, **58**(2): 444–449, June 1991.
29. Z. Hashin. **Extremum principles for elastic heterogenous media with imperfect interfaces and their application to bounding of effective moduli**. *J. Mech. Phys. Solid*, **40**(4):76–781, 1992.
30. Z. Hashin. **Thin interphase/imperfect interface in elasticity with application to coated fiber composites**. *J. Mech. Phys. Solid*, **50**(12):2509–2537, 2002.
31. R. Hill. **Elastic Properties of Reinforced Solids - Some Theoretical Principles**. *J. Mech. Phys. Solid*, **11**(5):357–372, 1963.
32. R. Hill. **A Self-consistent Mechanics of Composite Materials**. *J. Mech. Phys. Solid*, **13**(4):213–222, 1965.
33. H. Horii and S. Nemat- Nasser. **Overall moduli of solids with microcracks: Load-induced anisotropy**. *J. Mech. Phys. Solid*, **31**(2):155–171, 1983.
34. M. Kachanov. **Elastic solids with many cracks and related problems**. *Adv. Appl. Mech.*, **30**:259–445, 1994.
35. M. Kachanov and I. Sevostianov. **On quantitative characterization of microstructures and effective properties**. *Int. J. Solid Struct.*, **42**(2):309–336, 2005.
36. A. Klarbring. **Derivation of a model of adhesively bonded joints by the asymptotic expansion method**. *Int. J. Eng. Sci.*, **29**(4):493–512, 1991.
37. A. Klarbring and A.B. Movchan. **Asymptotic modelling of adhesive joints**. *Mech. Mat.*, **28**(1–4):137–145, 1998.

38. F. Lebon and R. Rizzoni. **Asymptotic analysis of a thin interface: The case involving similar rigidity**. *Int. J. Eng. Sci.*, **48**(5):473–486, 2010.
39. F. Lebon and R. Rizzoni. **Asymptotic behavior of a hard thin linear elastic interphase: An energy approach**. *Int. J. Solid Struct.*, **48**(3–4):441–449, 2011.
40. F. Lebon and F. Zaittouni. **Asymptotic modelling of interfaces taking contact conditions into account: Asymptotic expansions and numerical implementation**. *Int. J. Eng. Sci.*, **48**(2):111–127, 2010.
41. F. Lebon, A. Ould Khaoua, and C. Licht. **Numerical study of soft adhesively bonded joints in finite elasticity**. *Computat. Mech.*, **21**:134–140, 1998.
42. F. Lebon, R. Rizzoni, and S. Ronel- Idrissi. **Asymptotic analysis of some non-linear soft thin layers**. *Comput. & Struct.*, **82**(23–26):1929–1938, 2004.
43. C. Mauge and M. Kachanov. **Effective elastic properties of an anisotropic material with arbitrarily oriented interacting cracks**. *J. Mech. Phys. Solid*, **42**(4):561–584, 1994.
44. A. Needleman. **An analysis of decohesion along an imperfect interface**. *Int. J. Fract.*, **42**(1):21–40, 1990.
45. A. Needleman. **Micromechanical modelling of interfacial decohesion**. *Ultramicroscopy*, **40**(3):203–214, 1992.
46. A. Ould Khaoua. *Etude théorique et numérique de problemes de couches minces en elasticité*. PhD thesis, Université de Montpellier Sciences et Techniques du Languedoc, 1995.
47. P. Ponte Castañeda and J.R. Willis. **The effect of spatial distribution on the effective behavior of composite materials and cracked media**. *J. Mech. Phys. Solid*, **43**(12):1919–1951, 1995.
48. M.L. Raffa, F. Lebon, E. Sacco, and H. Welemane. **A multi-level interface model for damaged masonry**. In *B.H.V. Topping, P. Iványi, (Editors), "Proceedings of the Fourteenth International Conference on Civil, Structural and Environmental Engineering Computing", Civil-Comp Press, Stirlingshire, UK, Paper 64, 2013.*, 2013.
49. M.L. Raffa, F. Lebon, and R. Rizzoni. **On modeling brick/mortar interface via a St. Venant-Kirchhoff orthotropic soft interface. Part I: theory**. *Int. J. Mason. Res. Innov.*, In press.
50. A. Rekik and F. Lebon. **Identification of the representative crack length evolution in a multi-level interface model for quasi-brittle masonry**. *Int. J. Solid Struct.*, **47**(22–23): 3011–3021, 2010.
51. A. Rekik and F. Lebon. **Homogenization methods for interface modeling in damaged masonry**. *Adv. Eng. Softw.*, **46**(1):35–42, 2012.
52. R. Rizzoni and F. Lebon. **Imperfect interfaces as asymptotic models of thin curved elastic adhesive interphases**. *Mech. Res. Commun.*, **51**(0):39–50, 2013.
53. R. Rizzoni, S. Dumont, F. Lebon, and E. Sacco. **Higher order model for soft and hard elastic interfaces**. *Int. J. Solid Struct.*, **51**(23-24):4137–4148, 2014.
54. R. Rizzoni, Dumont S., and F. Lebon. **On Saint Venant - Kirchhoff Imperfect Interfaces**. Submitted.
55. E. Sanchez- Palencia. *Non-homogenous media and vibration theory*. Lecture notes in physics. Springer-Verlag, Berlin; New York, 1980.
56. E. Sanchez-Palencia and J. Sanchez-Hubert. *Introduction aux méthodes aymptotiques et à l' homogénéisation*. Masson, 1992.
57. I. Sevostianov and M. Kachanov. **Non-interaction Approximation in the Problem of Effective Properties**. In Mark Kachanov and Igor Sevostianov, editors, *Effective Properties of Heterogeneous Materials*, **193** of *Solid Mechanics and Its Applications*, pages 1–95. Springer Netherlands, 2013.
58. J.J. Telega. **Homogenization of fissured elastic solids in the presence of unilateral conditions and friction**. *Computat. Mech.*, **6**(2):109–127, 1990.
59. I. Tsukrov and M. Kachanov. **Anisotropic material with arbitrarily oriented cracks and elliptical holes: Effective elastic moduli**. *Int. J. Fract.*, **92**(1):L9–L14, 1998.
60. I. Tsukrov and M. Kachanov. **Effective moduli of an anisotropic material with elliptical holes of arbitrary orientational distribution**. *Int. J. Solid Struct.*, **37**(41):5919–5941, 2000.

A Stochastic Multi-scale Approach for Numerical Modeling of Complex Materials—Application to Uniaxial Cyclic Response of Concrete

Pierre Jehel

Abstract In complex materials, numerous intertwined phenomena underlie the overall response at macroscale. These phenomena can pertain to different engineering fields (mechanical, chemical, electrical), occur at different scales, can appear as uncertain, and are nonlinear. Interacting with complex materials thus calls for developing nonlinear computational approaches where multi-scale techniques that grasp key phenomena at the relevant scale need to be mingled with stochastic methods accounting for uncertainties. In this chapter, we develop such a computational approach for modeling the mechanical response of a representative volume of concrete in uniaxial cyclic loading. A mesoscale is defined such that it represents an equivalent heterogeneous medium: nonlinear local response is modeled in the framework of Thermodynamics with Internal Variables; spatial variability of the local response is represented by correlated random vector fields generated with the Spectral Representation Method. Macroscale response is recovered through standard homogenization procedure from Micromechanics and shows salient features of the uniaxial cyclic response of concrete that are not explicitly modeled at mesoscale.

1 Introduction

Widely-used materials in engineering practice such as polymer, composite, steel, concrete, are characterized by engineering parameters for design purposes, while these latter homogeneous macroscopic mechanical properties actually result from heterogeneous structures at lower scales. Material can be qualified as complex as their macroscopic behavior result from numerous multi-scale intertwined phenomena that

P. Jehel (✉)
MSSMat, CNRS, CentraleSupélec, Université Paris-Saclay, Grande Voie des Vignes,
92290 Châtenay-Malabry, France
e-mail: pierre.jehel@centralesupelec.fr

P. Jehel
Department of Civil Engineering and Engineering Mechanics, Columbia University,
630 SW Mudd, 500 West 120th Street, New York, NY 10027, USA

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_6

have nonlinear and uncertain evolution throughout loading history. Modifications in the underlying structures of this category of materials can result in dramatic changes in mechanical behavior at the relevant macroscopic scale for engineering applications. Micro-cracks coalescence in the constitutive material of a structure challenges its capacity for meeting the performance level targeted during its design process. Alkali-aggregate reaction in concrete microscopic structure can lead to hazardous loss of bearing capacity in reinforced concrete structures. Accounting for phenomena at lower scales to reliably predict macroscopic response of heterogeneous structures is one of the challenges numerical multi-scale simulation techniques have been developed for over the past ([1, 8, 15, 18] among many others).

In continuum mechanics, explicitly accounting for relevant mechanisms and structures in heterogeneous scales underlying macroscopic scale provides the rationale for representing characteristic features of homogenized material behavior laws at macroscale that can then be used for engineering design. Micromechanics has been developed to extract macroscopic local continuum properties from microscopically heterogeneous media through the concept of Representative Volume Element (RVE). An RVE for a material point at macroscale is statistically representative of the microscopic structure in a neighborhood of this point [23]. Also, Thermodynamics with Internal Variables provides a robust framework for modeling material response at macroscale according to a set of internal variables that carry information about the history of evolution mechanisms at lower scales without explicitly representing them [9, 21]. Other strategies to derive macroscopic mechanical properties of heterogeneous materials have been developed based on the introduction of a mesoscale, that is a scale that bridges the micro- and macroscales. In [24], heterogeneities are represented by random fields introduced at a mesoscale, which defines so-called Statistical Volume Elements that tends to become RVEs as mesoscale grows; effective properties at macroscale are retrieved according to two hierarchies of scale-dependent bounds obtained from either homogenous displacements or homogenous tensions applied on the boundary of the mesoscale. In [2], a mesoscale is explicitly constructed for representing the macroscopic behavior of heterogeneous quasi-brittle materials. This mesoscale consists of a 3D finite element mesh composed of truss elements cut by inclusions. Truss element kinematics is enriched to account for discontinuities in the strain field due to the presence of inclusions along truss elements as well as discontinuities in the displacement field to account for possible cracks in the matrix, in the inclusions, or at their interface. With the improvement of computational ressources, stochastic homogenization of random heterogeneous media can now be achieved without introducing a mesoscale. In [5], an efficient numerical strategy is presented to obtain effective tensors of random materials by coupling random micro-structures to tentative effective models within the Arlequin framework for model superposition [1]. In [30], micro-structures composed of a medium with randomly distributed inclusions of random shapes are generated and their behaviors are simulated with the extended finite element method (XFEM); homogenized properties at macroscale are then derived through the computation of mean response using Monte Carlo simulations.

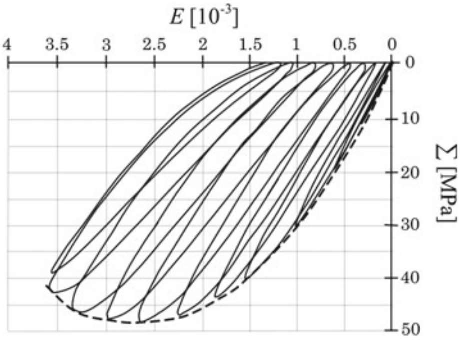


Fig. 1 Strain-stress concrete experimental response in pseudo-static cyclic uniaxial compressive loading (adapted from [29])

In the work presented in this chapter, we focus on the numerical representation of the homogenized one-dimensional response of a concrete specimen in cyclic compressive loading, as it can be observed in lab tests. Figure 1 illustrates the main features of such an homogenized response: a backbone curve (dashed line) that is a nonlinear strain hardening phase ($0 \leq E \leq 2.7 \times 10^{-3}$) followed by a strain softening phase where strength degradation is observed; unloading-reloading cycles show that stiffness decreases while loading increases, hysteresis loops are generated. This typical response is observed at macroscale and results from numerous underlying mechanisms of physical or chemical nature at many different scales. For designing concrete structures, an equivalent homogeneous concrete model is sought, which has to represent concrete mechanical behavior in different loading conditions while accounting for mechanisms at lower scales [37]. Heterogeneities can be observed in concrete at different scales: aggregates of different sizes are distributed in a cement paste; the so-called interfacial transition zone where the aggregates are bound to the cement paste plays a key role in the concrete mechanical properties [38]; cement paste is composed of water, voids and of the products of the complete or partial hydration of the clinker particles, which generates a microscopic structure composed of numerous intertwined phases.

This chapter presents the basic ingredients of a stochastic multi-scale approach developed to represent the macroscopic compressive cyclic response of a concrete specimen while attempting not to sacrifice too much of the complexity of this material. To that aim, two scales are considered: the macroscale where an equivalent homogenous concrete model capable of representing the main features that are shown in Fig. 1 is retrieved, and a mesoscale where heterogeneous local nonlinear response is assumed. Local response at mesoscale is modeled in the framework of Thermodynamics with Internal Variables and is seen as the homogenized response of mechanisms that occur at the micro- or nano-underlying scales. Spatial variability at mesoscale is introduced using stochastic vector fields. Homogenized macroscopic response is recovered using standard averaging method from micromechanics.

The chapter is organized as follows. In the next section, the ingredients of the proposed stochastic multi-scale modeling are presented. First, the averaging method for computing the homogenized model response at macroscale is recalled.

Then, the model of the mechanical local behavior of a material point at mesoscale is constructed. Finally, the Spectral Representation Method for generating stochastic vector fields that model heterogeneity at mesoscale is presented. In a third section, the numerical implementation of the approach in the framework of the finite element method is detailed. Before the conclusion, numerical applications are presented to demonstrate the capability of the proposed approach (i) for yielding homogeneous material behavior at macroscale without stochastic homogenization and (ii) for representing salient features of macroscopic 1D concrete response in uniaxial cyclic compressive loading.

2 Multi-scale Stochastic Approach for Modeling Concrete

Figure 2 presents the three following concepts, which further developments are based on:

- *Actual heterogeneous medium* (A-mesoscale): Concrete is made of aggregates distributed in a cement paste. Aggregates, cement and interface between both of them exhibit different mechanical responses. In the cement paste, micro- and nano-structures also exist.
- *Equivalent heterogeneous medium* (E-mesoscale): The proposed approach does not consist in explicitly generating a multi-phase medium with random distribution of aggregates of random geometry in a cement paste with known mechanical behavior for each phase. The approach followed here consists in generating a random medium at each point of which the mechanical response obeys a prescribed behavior that has uncertain parameters and that is the homogenized response of mixtures of aggregates and cement where mechanisms at lower scales are also involved but not explicitly modeled.
- *Equivalent homogeneous medium* (macroscale): Homogenization of E-mesoscale yields homogenized homogeneous concrete response. It will be shown in the numerical applications that one realization only of the random E-mesoscale can be sufficient to retrieve homogeneous properties at macroscale.

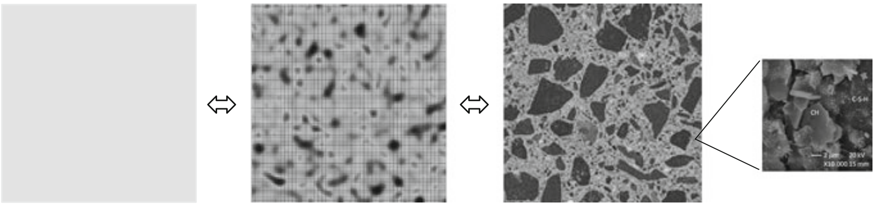


Fig. 2 From *left* to *right*: equivalent homogeneous concrete (macroscale), equivalent heterogeneous concrete (E-mesoscale, 5 cm × 5 cm-square), actual heterogeneous concrete (A-mesoscale, 5 cm × 5 cm-square), and zoom on the underlying microstructure in the cement paste (20 µm × 20 µm-square observed through Scanning Electron Microscope, courtesy Trigo [38])

2.1 Homogenized Material Behavior at Macroscale

We consider a material elementary domain (ED) that occupies a spatial domain $\mathscr{R} \subset \mathbb{R}^3$. The boundary $\partial\mathscr{R}$ of the ED has outward normal $\mathbf{n}$, tension $\bar{\mathbf{t}}$ can be imposed on the part of the boundary $\partial_\sigma\mathscr{R}$ while displacement $\bar{\mathbf{u}}$ can be imposed on $\partial_u\mathscr{R}$, where $\partial_\sigma\mathscr{R} \cup \partial_u\mathscr{R} = \partial\mathscr{R}$ and $\partial_\sigma\mathscr{R} \cap \partial_u\mathscr{R} = \emptyset$. There are no external forces other than $\bar{\mathbf{t}}$ applied on the ED and no dynamic effects are considered either. Then, the displacement vector field $\mathbf{u}$, and the strain and stress tensor fields, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$, satisfy at any pseudo-time $t \in [0,\ T]$:

$$\begin{aligned}
&\mathbf{div}\,\boldsymbol{\sigma}(\mathbf{x},t) = \mathbf{0} && \forall \mathbf{x} \in \mathscr{R}\\
&\boldsymbol{\varepsilon}(\mathbf{x},t) = \mathrm{sym}\,[\boldsymbol{\nabla}(\mathbf{u}(\mathbf{x},t))] && \forall \mathbf{x} \in \mathscr{R}\\
&\boldsymbol{\sigma}(\mathbf{x},t) = \hat{\boldsymbol{\sigma}}(\boldsymbol{\varepsilon}(\mathbf{x},t)) && \forall \mathbf{x} \in \mathscr{R}\\
&\boldsymbol{\sigma}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = \bar{\mathbf{t}}(\mathbf{x},t) && \forall \mathbf{x} \in \partial_\sigma\mathscr{R}\\
&\mathbf{u}(\mathbf{x},t) = \bar{\mathbf{u}}(\mathbf{x},t) && \forall \mathbf{x} \in \partial_u\mathscr{R}
\end{aligned} \tag{1}$$

$\mathrm{sym}\,[\boldsymbol{\nabla}(\cdot)] := \frac{1}{2}\left(\boldsymbol{\nabla}(\cdot) + \boldsymbol{\nabla}^T(\cdot)\right)$ is the symmetric part of the gradient tensor $\boldsymbol{\nabla}(\cdot)$, the superscript $(\cdot)^T$ denoting the transpose operation. In the set of equations above, small strains are assumed and behavior law $\hat{\boldsymbol{\sigma}}(\boldsymbol{\varepsilon})$ can be nonlinear.

We classically assume that any macroscopic quantity Q is connected to its E-mesoscopic counterpart q through domain averaging over the ED:

$$Q(\mathbf{X}) := \langle q\rangle(\mathbf{X}) = \frac{1}{|\mathscr{R}|}\int_{\mathscr{R}} q(\mathbf{x};\mathbf{X})d\mathbf{x} \tag{2}$$

$|\mathscr{R}| = \int_{\mathscr{R}} d\mathbf{x}$ is the measure of the spatial domain occupied by the ED centered at material point $\mathbf{X}$ of the macroscale, and $\mathbf{x}$ denotes a material point of the E-mesoscale.

In all what follows, we will assume linear displacements imposed all over the boundary of $\mathscr{R}$:

$$\mathbf{u}(\mathbf{x},t) = \mathbf{E}'(\mathbf{X},t)\cdot\mathbf{x};\quad \forall \mathbf{x} \in \partial\mathscr{R}\ ,\ \forall t \in [0,\ T] \tag{3}$$

Hence, $\partial_u\mathscr{R} = \partial\mathscr{R}$ and $\partial_\sigma\mathscr{R} = \emptyset$. With this assumption, it can be shown (see e.g. [23, Chap. 1] or [39]) that:

$$\mathbf{E}'(\mathbf{X},t) = \mathbf{E}(\mathbf{X},t) := \langle\boldsymbol{\varepsilon}(\mathbf{x},t)\rangle;\quad \mathbf{x} \in \mathscr{R}(\mathbf{X}) \tag{4}$$

and also, because it is assumed there is no external forces applied on $\mathscr{R}(\mathbf{X})$:

$$\boldsymbol{\Sigma}(\mathbf{X},t) = \frac{1}{|\mathscr{R}|}\int_{\partial\mathscr{R}} \mathrm{sym}\ [\mathbf{t}(\mathbf{x},t)\otimes\mathbf{x}]\ d\partial\mathscr{R} \tag{5}$$

where $\mathbf{t}(\mathbf{x},t) := \boldsymbol{\sigma}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x})$ are the tension forces developed over $\partial_u\mathscr{R}$.

Note that other boundary conditions could be considered. In any case, it is in general not possible to derive the strain or stress fields at E-mesoscale from the macroscopic quantities, and consequently simplifying assumptions as in (3) are made. Whether it is displacements or forces that are imposed on $\mathscr{R}$, and whether these latter conditions are linear or periodic, this can influence the homogenized macroscopic response of the ED. However, this is out of the scope of this work where the consequences of assuming linear displacements imposed on $\partial\mathscr{R}$ will not be discussed.

With the boundary conditions (3) applied on $\mathscr{R}$, Hill's lemma can be proved:

$$\langle \boldsymbol{\sigma} : \boldsymbol{\varepsilon} \rangle = \langle \boldsymbol{\sigma} \rangle : \langle \boldsymbol{\varepsilon} \rangle := \boldsymbol{\Sigma} : \mathbf{E} \tag{6}$$

which means that the medium recovered at macroscale through homogenization is energetically equivalent to the heterogeneous medium considered at E-mesoscale.

The possibly nonlinear material response at mesoscale is expressed as:

$$\dot{\boldsymbol{\sigma}} = \boldsymbol{\lambda} : \dot{\boldsymbol{\varepsilon}} \tag{7}$$

where $\boldsymbol{\lambda}$ is the tangent modulus at E-mesoscale and the superimposed dot denotes partial derivative with respect to pseudo-time. Thanks to Hill's lemma, we then have the following two equivalent definitions for the tangent modulus $\mathbf{L}$ at the homogenized macroscale:

$$\langle \boldsymbol{\varepsilon} : \boldsymbol{\lambda} : \dot{\boldsymbol{\varepsilon}} \rangle = \mathbf{E} : \mathbf{L} : \dot{\mathbf{E}} \qquad \Leftrightarrow \qquad \dot{\boldsymbol{\Sigma}} = \mathbf{L} : \dot{\mathbf{E}} \tag{8}$$

2.2 Material Behavior Law at E-mesoscale

We assume a coupled damage-plasticity model to be suitable for representing material response at any material point $\mathbf{x}$ at E-mesoscale (see Fig. 2). This choice is motivated by the fact that concrete A-mesoscale is composed of both a ductile cement matrix that can be represented by a plastic model, and brittle aggregates that are confined in the cement paste and whose compressive response can be more realistically represented by a damage model. Hereafter, we develop a model in a way that allows for explicitly controlling the coupling of damage and plasticity. Indeed, as illustrated in Fig. 3, the response of material points at mesoscale can be either better represented by a damage model alone, or a plasticity model alone, or by the appropriate coupling of both models. This is developed in the framework of thermodynamics with internal variables [9, 21] where the internal variables carry the history of irreversible mechanisms occurring in the material at lower (micro- and nano-) scales.

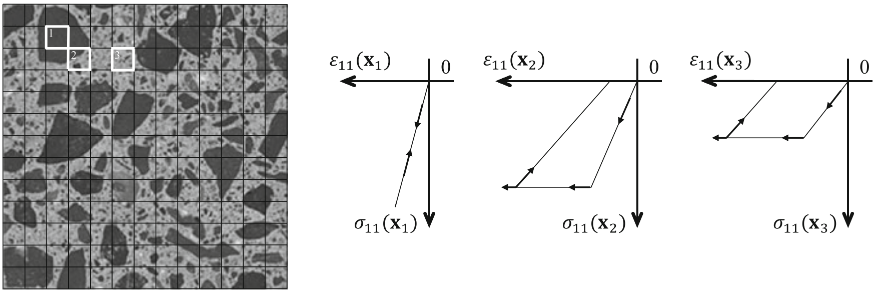

1
2
3
$\varepsilon_{11}(\mathbf{x}_1)$
0
$\sigma_{11}(\mathbf{x}_1)$
$\varepsilon_{11}(\mathbf{x}_2)$
0
$\sigma_{11}(\mathbf{x}_2)$
$\varepsilon_{11}(\mathbf{x}_3)$
0
$\sigma_{11}(\mathbf{x}_3)$

Fig. 3 Each point at E-mesoscale has a different behavior due to the heterogenous structure of concrete. Behavior laws at E-mesoscale are homogenized responses of aggregate-cement paste mixtures with also heterogenous microstructures in the cement paste. At point $\mathbf{x}_1$ there is an aggregate solely; at point $\mathbf{x}_2$ there is mixture of large and small aggregates in the cement paste; at point $\mathbf{x}_3$ there is mainly cement with some small aggregates

2.2.1 Basic Ingredients

The three basic ingredients for developing this model of local behavior law at E-mesoscale are as follows:

- Total deformation ε is split into damage (ε^d) and plastic (ε^p) parts:

$$\varepsilon := \varepsilon^d + \varepsilon^p \tag{9}$$

- Stored energy function is defined as:

$$\psi(\varepsilon, \mathbf{D}, \varepsilon^p) := \boldsymbol{\sigma} : \varepsilon^d - \frac{1}{2}\boldsymbol{\sigma} : \mathbf{D} : \boldsymbol{\sigma} \tag{10}$$

  with $\mathbf{D}$ the fourth-order damage compliance tensor. $\mathbf{D}$ and ε^p are the internal variables that drive the evolution of the material. Also, denoting by $\mathbf{C}$ the elasticity tensor, we set initially, as the material is undamaged, $\mathbf{D}^{-1} = \mathbf{C}$. The elements of $\mathbf{C}$ are parameters of the model.
- A criterium function is introduced as:

$$\phi(\boldsymbol{\sigma}) := h(\boldsymbol{\sigma}) - \sigma_y \leq 0. \tag{11}$$

  It defines the limit states between the states where there is no evolution of the internal variables ($\phi < 0$) and those where there is evolution ($\phi = 0$). The so-called yield stress $\sigma_y > 0$ is a scalar parameter.

More general models coupling damage and plasticity with hardening or softening could be defined. Then, other internal variables would be introduced (see e.g. [14, 17, 20]).

2.2.2 Material Dissipation and State Equation

Then, the material dissipation reads:

$$\begin{aligned}\mathscr{D} &:= \boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}} - \dot{\psi} \geq 0 \\ &= \dot{\boldsymbol{\sigma}} : \left(\mathbf{D} : \boldsymbol{\sigma} - \boldsymbol{\varepsilon}^d\right) + \frac{1}{2}\boldsymbol{\sigma} : \dot{\mathbf{D}} : \boldsymbol{\sigma} + \boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}}^p \geq 0\end{aligned} \tag{12}$$

$\mathscr{D}$ should be non-negative to comply with the principles of thermodynamics. In case there is no evolution of the internal variables, that is for loading steps that do not generate any change of state in the material, there is no evolution of the internal variables: $\dot{\mathbf{D}} = \dot{\boldsymbol{\varepsilon}}^p = \mathbf{0}$ and the process is assumed to be non-dissipative, that is $\mathscr{D}$ is null. According to Eq. (12), it then comes the state equation:

$$\boldsymbol{\varepsilon}^d := \mathbf{D} : \boldsymbol{\sigma} \tag{13}$$

Equation (13) is to this damage model what the more classical constitutive relation $\boldsymbol{\sigma} := \mathbf{C} : \boldsymbol{\varepsilon}^e$ is to linear elasticity model.

Introducing this latter state equation into Eq. (12), we can rewrite:

$$\mathscr{D} = \frac{1}{2}\boldsymbol{\sigma} : \dot{\mathbf{D}} : \boldsymbol{\sigma} + \boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}}^p \geq 0 \tag{14}$$

from where we define $\mathscr{D}^d := \frac{1}{2}\boldsymbol{\sigma} : \dot{\mathbf{D}} : \boldsymbol{\sigma} \geq 0$ and $\mathscr{D}^p := \boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}}^p \geq 0$.

2.2.3 Evolution of the Internal Variables

Following what has been done to derive the equations of mechanical models with plasticity solely [12], the evolution of the internal variables is obtained appealing to the principle of maximum dissipation. Accordingly, among all the admissible stresses, that is $\boldsymbol{\sigma}$ such that $\phi(\boldsymbol{\sigma}) \leq 0$, it is those that maximize the material dissipation $\mathscr{D}$ that have to be retained. This can be cast into a minimization problem with constraint $\phi \leq 0$ [19]. Lagrange multiplier method can be used to solve it with the so-called Lagrangian reading:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\sigma}, \dot{\gamma}) &:= -\mathscr{D} + \dot{\gamma}\,\phi \\ &= (-\mathscr{D}^d + \dot{\gamma}^d\,\phi) + (-\mathscr{D}^p + \dot{\gamma}^p\,\phi)\end{aligned} \tag{15}$$

Here, we have split the total Lagrange multiplier $\dot{\gamma} \geq 0$ so that $\dot{\gamma} = \dot{\gamma}^d + \dot{\gamma}^p$ with two Lagrange multipliers defined as $\dot{\gamma}^d := r\,\dot{\gamma}$ and $\dot{\gamma}^p := (1-r)\,\dot{\gamma}$ where r is to be taken in the range $[0,\ 1]$. r is a damage-plasticity coupling parameter: if $r = 0$, $\dot{\gamma}^d = 0$ and there is plasticity evolution only; if $r = 1$, only damage evolves in the material; and for any other r in-between, there is coupled evolution of both damage and plasticity.

In turn, the Lagrangian is also split into damage and plasticity parts:

$$\mathscr{L}^d(\boldsymbol{\sigma}, \dot{\gamma}^d) := -\frac{1}{2}\boldsymbol{\sigma} : \dot{\mathbf{D}} : \boldsymbol{\sigma} + \dot{\gamma}^d\,\phi \quad ; \quad \mathscr{L}^p(\boldsymbol{\sigma}, \dot{\gamma}^p) := -\boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}}^p + \dot{\gamma}^p\,\phi \tag{16}$$

Both parts have to be minimized to ensure the total Lagrangian is minimum. The Kuhn-Tucker optimality conditions associated to these minimization problems result in:

$$\frac{\partial \mathscr{L}^{d,p}}{\partial \boldsymbol{\sigma}} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathscr{L}^{d,p}}{\partial \dot{\gamma}^{d,p}} = 0 \tag{17}$$

Setting $\boldsymbol{\nu} := \partial\phi/\partial\boldsymbol{\sigma}$, this leads to the following equations of evolution of the internal variables:

$$\dot{\mathbf{D}} : \boldsymbol{\sigma} = \dot{\gamma}^d\,\boldsymbol{\nu} := r\,\dot{\gamma}\,\boldsymbol{\nu} \tag{18}$$

$$\dot{\boldsymbol{\varepsilon}}^p = \dot{\gamma}^p\,\boldsymbol{\nu} := (1-r)\,\dot{\gamma}\,\boldsymbol{\nu} \tag{19}$$

Besides, this minimizing problem also yields the following so-called loading/ unloading conditions:

$$\dot{\gamma}^{d,p} \geq 0\,;\ \phi \leq 0\,;\ \dot{\gamma}^{d,p}\,\phi = 0 \tag{20}$$

2.2.4 Damage and Plasticity Multipliers

In the case $\dot{\gamma}^d > 0$ or $\dot{\gamma}^p > 0$, there is damage or plasticity evolution and, according to (20), $\phi(\boldsymbol{\sigma})$ as to remain null during the process so that the stresses remain admissible. We thus have the consistency condition $\dot{\phi} = 0$ that can be rewritten as:

$$\frac{\partial\phi}{\partial\boldsymbol{\sigma}} : \frac{\partial\boldsymbol{\sigma}}{\partial t} = \boldsymbol{\nu} : \dot{\boldsymbol{\sigma}} = 0 \tag{21}$$

Remarking from (13) that $\dot{\boldsymbol{\varepsilon}}^d = \dot{\mathbf{D}} : \boldsymbol{\sigma} + \mathbf{D} : \dot{\boldsymbol{\sigma}}$ and using Eqs. (9), (18) and (19), we have:

$$\mathbf{D} : \dot{\boldsymbol{\sigma}} = \dot{\boldsymbol{\varepsilon}} - \dot{\gamma}\,\boldsymbol{\nu} \tag{22}$$

Then, assuming $\mathbf{D} \neq \mathbf{0}$, the consistency condition (21) is satisfied when $\dot{\gamma} > 0$ if:

$$\dot{\gamma}\,\boldsymbol{\nu} = \dot{\boldsymbol{\varepsilon}} \tag{23}$$

Or, with the damage and plasticity multipliers $\dot{\gamma}^d > 0$ and $\dot{\gamma}^p > 0$:

$$\dot{\gamma}^d\,\boldsymbol{\nu} = r\,\dot{\boldsymbol{\varepsilon}} \quad \text{and} \quad \dot{\gamma}^p\,\boldsymbol{\nu} = (1-r)\,\dot{\boldsymbol{\varepsilon}} \tag{24}$$

2.2.5 Tangent Modulus

The tangent modulus at mesoscale $\boldsymbol{\lambda}$ is a fourth-order tensor that has been defined in (7) such that $\dot{\boldsymbol{\sigma}} = \boldsymbol{\lambda} : \dot{\boldsymbol{\varepsilon}}$. Assuming that $\mathbf{D}^{-1}$, the inverse of $\mathbf{D}$, exists ($\mathbf{D}^{-1} : \mathbf{D} = \mathbf{I}$, where $\mathbf{I}$ is the identity fourth-order tensor), and reminding Eqs. (22) and (23), we have:

$$\boldsymbol{\lambda} = \begin{cases} \mathbf{D}^{-1} \text{ if } & \dot{\gamma} = 0 \quad (\phi(\boldsymbol{\sigma}) < 0) \\ \mathbf{0} \quad \text{ if } & \dot{\gamma} > 0 \quad (\phi(\boldsymbol{\sigma}) = 0\,;\ \dot{\phi}(\boldsymbol{\sigma}) = 0) \end{cases} \tag{25}$$

To sum up, the proposed material model at E-mesoscale is based on a set of internal variables that consists of the damage compliance tensor $\mathbf{D}$ and the plastic deformation tensor $\boldsymbol{\varepsilon}^p$. Besides, the model is parameterized by the elasticity tensor $\mathbf{C}$, the stress threshold σ_y above which damage or plastic evolution occurs, and the damage-plasticity coupling coefficient r: if $r = 1$, there is no plastic evolution and the material can only damage, while if $r = 0$, there is no damage evolution and the material is perfectly plastic. Figure 4 shows material constitutive behavior at two different material points $\mathbf{x}_1$ and $\mathbf{x}_2$ of the E-mesoscale where the parameters take different values: parameters, and consequently local response, vary over the domain $\mathscr{R}$ due to heterogeneities at E-mesoscale.

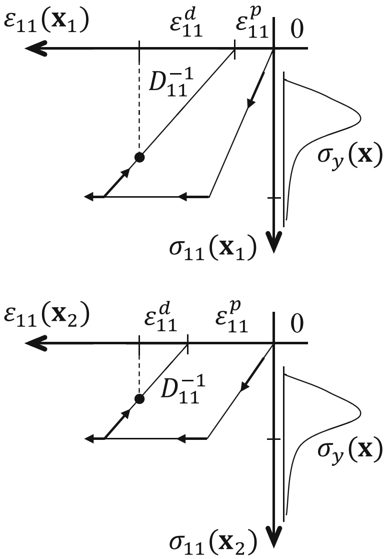


Fig. 4 Example of the model response at two different material points of the E-mesoscale. Spatial variability is explicitly illustrated on the figure. Initial stiffness is determined by a spatially variable elastic modulus ($D_{11}^{-1} = C_{11}$); yielding threshold σ_y fluctuates over $\mathscr{R}$; how fast damage evolves comparing to plasticity is governed by the spatially variable coupling parameter r

2.3 Stochastic Modeling of Heterogeneous E-mesoscale

Spatial variability at E-mesoscale of a set of m parameters $\mathbf{a}$ over $\mathscr{R}$ is conveyed through stochastic modeling: it is assumed that the fluctuations of correlated stochastic fields can describe the actual material heterogeneous meso-structure (A-mesoscale). Thus, we introduce the probability space $(\Theta, \mathfrak{S}, P)$ where Θ is the sample space containing all the possible outcomes θ from the observation of the random phenomenon that is studied; $\mathfrak{S}$ is the σ-algebra associated with Θ; P is a probability measure. A real parameter $a \in \mathbf{a}$ taking values in $\mathscr{V}_a$ is then considered as the realization of a random variable $\mathfrak{a}(\theta)$: $\Theta \rightarrow \mathscr{V}_a$. A random variable can be completely defined by its cumulative distribution function: $\mathscr{F}_{\mathfrak{a}}(a) = \Pr[\mathfrak{a}(\theta) \leq a] = \int_{\{\theta | \mathfrak{a}(\theta) \leq a\}} P(\theta)$ or, when a probability density function (PDF) $p_{\mathfrak{a}}(a)$ exists: $\mathscr{F}_{\mathfrak{a}}(a) = \int_{\{s \in \mathscr{V}_a | s \leq a\}} p_{\mathfrak{a}}(s)\, ds$.

Before we go on with the definition of stochastic fields, we recall some basic definitions for random variables. The mean $\mu_{\mathfrak{a}}$ and the variance $s_{\mathfrak{a}}^2$ of a random variable $\mathfrak{a}$ are defined as:

$$\mu_{\mathfrak{a}} := \mathbb{E}[\mathfrak{a}] = \int_{-\infty}^{+\infty} a\, p_{\mathfrak{a}}(a)\, da \tag{26}$$

$$s_{\mathfrak{a}}^2 := \mathbb{E}[(\mathfrak{a} - \mu_{\mathfrak{a}})^2] \tag{27}$$

where $\mathbb{E}[\cdot]$ is the so-called mathematical expectation and $p_{\mathfrak{a}}(a) = 0, \forall a \in \mathbb{R} \setminus \mathscr{V}_a$. Also, $\mathfrak{a}$ and $\mathfrak{b}$ being two random variables, the covariance is defined as:

$$\begin{aligned} Cov_{\mathfrak{a}\mathfrak{b}} &:= \mathbb{E}[(\mathfrak{a} - \mu_{\mathfrak{a}})(\mathfrak{b} - \mu_{\mathfrak{b}})] \\ &:= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\mathfrak{a} - \mu_{\mathfrak{a}})(\mathfrak{b} - \mu_{\mathfrak{b}})\, p_{\mathfrak{a}\mathfrak{b}}(a, b)\, da\, db \end{aligned} \tag{28}$$

where $p_{\mathfrak{a}\mathfrak{b}}$ is the joint PDF of $\mathfrak{a}$ and $\mathfrak{b}$, with $p_{\mathfrak{a}\mathfrak{b}}(a, b) = 0\ \forall (a, b) \in \mathbb{R}^2 \setminus \mathscr{V}_a \times \mathscr{V}_b$. We also introduce the correlation, which is defined as:

$$R_{\mathfrak{a}\mathfrak{b}} := \mathbb{E}[\mathfrak{a}\, \mathfrak{b}] = Cov_{\mathfrak{a}\mathfrak{b}} + \mu_{\mathfrak{a}}\, \mu_{\mathfrak{b}} \tag{29}$$

And finally the following correlation coefficient will also be used later on:

$$\rho_{\mathfrak{a}\mathfrak{b}} := \frac{Cov_{\mathfrak{a}\mathfrak{b}}}{s_{\mathfrak{a}}\, s_{\mathfrak{b}}} \in [-1,\ 1] \tag{30}$$

2.3.1 Random Vector Fields for Modelind Heterogeneous Meso-Structure

It is assumed that the heterogeneity of the parameters $\mathbf{C}$, σ_y and r of the model developed above for representing material response at E-mesoscale over a concrete elementary domain (ED) can be represented as the realization of a random vector field. A random vector $\boldsymbol{a}(\theta)$ is a vector of random variables. Let $\mathscr{B} \subset \mathbb{R}^d$ be a spatial domain of dimension d; this can be a volume ($d = 3$), an area ($d = 2$) or a length ($d = 1$). A random vector field $\boldsymbol{g}(\mathbf{x}; \theta)$ over $\mathscr{B}$ is a collection of random vectors indexed by the position $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathscr{B}$. For any fixed $\mathbf{x} \in \mathscr{B}$, any component $\mathfrak{g}_j(\mathbf{x})$ of $\boldsymbol{g}$, $j \in [1, \ldots, m]$, is a random variable. In the case of random vector fields, we have the following definitions of the mean, the auto-correlation and cross-correlation functions respectively:

$$\mu_j^{\boldsymbol{g}}(\mathbf{x}) := \mathbb{E}[\mathfrak{g}_j(\mathbf{x})] \qquad j \in [1, \ldots, m] \tag{31}$$

$$R_{jj}^{\boldsymbol{g}}(\mathbf{x}, \boldsymbol{\xi}) := \mathbb{E}[\mathfrak{g}_j(\mathbf{x})\, \mathfrak{g}_j(\mathbf{x} + \boldsymbol{\xi})] \qquad j \in [1, \ldots, m] \tag{32}$$

$$R_{jk}^{\boldsymbol{g}}(\mathbf{x}, \boldsymbol{\xi}) := \mathbb{E}[\mathfrak{g}_j(\mathbf{x})\, \mathfrak{g}_k(\mathbf{x} + \boldsymbol{\xi})] \qquad j \in [1, \ldots, m]\,,\ k \in [1, \ldots, m]\,,\ j \neq k \tag{33}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^T$ is the separation distance vector between two points of $\mathscr{B}$.

To fully characterize random vector fields, we need the marginal and joint PDFs of all possible combinations of random variables $\mathfrak{g}_j(\mathbf{x})$. From a practical point of view, this implies that many concrete ED have to be considered, that for each of them the parameters of interest have to be identified at many points $\mathbf{x}$ all over $\mathscr{B}$, so that these PDFs can be empirically constructed. If gathering such a huge amount of information was needed, the usefulness of using random vector field for modeling heterogeneity in concrete structure at mesoscale would be questionable. Therefore, we will make assumptions on the structure of the random vector field that would justify the efficiency of the proposed approach.

E-mesoscale construction will rely on the two following assumptions:

1. Random fields will be generated as Gaussian. This means that random field $\mathfrak{g}_j(\mathbf{x}; \theta)$ is fully characterized by both its mean function $\mu_j^{\boldsymbol{g}}(\mathbf{x})$ and auto-correlation function $R_{jj}^{\boldsymbol{g}}(\mathbf{x}, \boldsymbol{\xi})$. Nevertheless, non-Gaussian random field can then be obtained through nonlinear translation of Gaussian field, which will be discussed in Sect. 2.3.3.
2. Random fields are jointly homogeneous. This means that their mean function is independent of the position $\mathbf{x}$ and that auto- and cross-correlation functions depend on the separation distance only:

$$\mu_j^{\boldsymbol{g}}(\mathbf{x}) := \mu_j^{\boldsymbol{g}} \qquad j \in [1, \ldots, m] \tag{34}$$

$$R_{jk}^{\boldsymbol{g}}(\mathbf{x}, \boldsymbol{\xi}) := R_{jk}^{\boldsymbol{g}}(\boldsymbol{\xi}) \qquad j \in [1, \ldots, m]\,,\ k \in [1, \ldots, m] \tag{35}$$

Note that efficient techniques can be used to account for heterogeneity in the random field (see e.g. [25]).

Also, we will consider the possible ergodicity of the generated random fields in mean and correlation functions. One realization of such an ergodic random vector field contains all the statistical information needed to retrieve the first two moments: means and correlation functions can be computed as spatial averages.

2.3.2 Spectral Representation of Homogeneous Gaussian Random Vector Fields

We present here the Spectral Representation Method for generating standard homogenous Gaussian fields [6, 7, 26, 33, 34]. Note that considering the Gaussian random fields to be standard, that is with zero mean and unit variance, does not introduce any loss of generality because non-standard Gaussian fields can always be retrieved through linear transformation.

The basic ingredient is the definition of a target correlation matrix, which can be built from experimental observations for instance:

$$\mathbf{R}^0(\boldsymbol{\xi}) = \begin{pmatrix} R^0_{11}(\boldsymbol{\xi}) & \cdots & R^0_{1m}(\boldsymbol{\xi}) \\ \vdots & \ddots & \vdots \\ R^0_{m1}(\boldsymbol{\xi}) & \cdots & R^0_{mm}(\boldsymbol{\xi}) \end{pmatrix} \tag{36}$$

Superscript ^{0} has been added to highlight that these functions are target correlations, which should be retrieved in the statistical analysis of the generated random fields.

According to Wiener-Khinchin theorem, power spectral density functions S^0_{jj}, $j \in [1, \ldots, m]$, and cross-spectral density functions S^0_{jk}, $(j, k) \in [1, \ldots, m]^2$, $j \neq k$, are the Fourier transform of the corresponding correlation functions:

$$S^0_{jk}(\boldsymbol{\kappa}) = \frac{1}{(2\pi)^d} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} R^0_{jk}(\boldsymbol{\xi})\, e^{-i\boldsymbol{\kappa}\cdot\boldsymbol{\xi}}\, d\xi_1 \ldots d\xi_d \tag{37}$$

$$R^0_{jk}(\boldsymbol{\xi}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} S^0_{jk}(\boldsymbol{\kappa})\, e^{i\boldsymbol{\kappa}\cdot\boldsymbol{\xi}}\, d\kappa_1 \ldots d\kappa_d \tag{38}$$

where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_d)^T$ is the wave number vector, $\boldsymbol{\kappa} \cdot \boldsymbol{\xi}$ is the scalar product of the two vectors $\boldsymbol{\kappa}$ and $\boldsymbol{\xi}$, and i is the imaginary unit. Power spectral density functions are by definition real functions of $\boldsymbol{\kappa}$ while cross-spectral functions can be complex functions of $\boldsymbol{\kappa}$. It has been shown in [32] that matrix $\mathbf{S}(\boldsymbol{\kappa})$ is Hermitian and semidefinite positive, which implies that it can be decomposed using Cholesky's method as:

$$\mathbf{S}(\boldsymbol{\kappa}) = \mathbf{H}(\boldsymbol{\kappa})\, \mathbf{H}^{\star T}(\boldsymbol{\kappa}) \tag{39}$$

where $(\cdot)^\star$ denotes the complex conjugate and $\mathbf{H}(\boldsymbol{\kappa})$ is a lower triangular matrix. The diagonal elements H_{jj} are real and non-negative functions of $\boldsymbol{\kappa}$ while the off-diagonal elements can be complex:

$$H_{jk}(\boldsymbol{\kappa}) = |H_{jk}|\, e^{i\,\varphi_{jk}(\boldsymbol{\kappa})}; \quad j \in [1, \ldots, m]\,;\ k \in [1, \ldots, m]\,;\ j > k \tag{40}$$

Then, the jth component of a realization of a mV-dD homogeneous standard Gaussian stochastic vector field $\boldsymbol{g}(\mathbf{x};\theta)$ with cross-spectral density matrix $\mathbf{S}(\boldsymbol{\kappa})$ reads:

$$g_j(\mathbf{x};\theta) = 2\sqrt{\Delta\kappa_1 \dots \Delta\kappa_d} \sum_{l=1}^{m} \sum_{n_1=0}^{N_1-1} \dots \sum_{n_d=0}^{N_d-1} \sum_{\alpha=1}^{2^{d-1}} |H_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d})| \times \cos\left(\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d} \cdot \mathbf{x} - \varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d}) + \Phi^{\alpha}_{l,n_1\dots n_d}(\theta)\right) \tag{41}$$

for $j \in [1,\dots,m]$ and $N_1 \to +\infty$,..., $N_d \to +\infty$. In Eq. (41), the following notation has been introduced:

$$\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d} := \begin{pmatrix} n_1\,\Delta\kappa_1 & I^{\alpha}_2 n_2\,\Delta\kappa_2 & \dots & I^{\alpha}_d n_d\,\Delta\kappa_d \end{pmatrix}^T \tag{42}$$

Wave numbers increments are defined as:

$$\Delta\kappa_i := \frac{\kappa_{u\,i}}{N_i} \quad , \quad i \in [1,\dots,d] \tag{43}$$

where $\kappa_{u\,i}$'s are so-called cut-off wave numbers such that $\mathbf{S}(\boldsymbol{\kappa})$ can be assumed to be null for any $\kappa_i \geq \kappa_{u\,i}$. Also, $(I^{\alpha}_1, I^{\alpha}_2, \dots, I^{\alpha}_d)$ are the α different vectors composed of $+1$'s and -1's where $I^{\alpha}_1 = 1$ for all α. For instance, for $d = 3$, there are the following $\alpha = 2^{3-1} = 4$ different such arrangements: $(1, 1, 1)$, $(1, 1, -1)$, $(1, -1, 1)$ and $(1, -1, -1)$, so that $\boldsymbol{\kappa}^1_{n_1n_2n_3} = (n_1\,\Delta\kappa_1, n_2\,\Delta\kappa_2, n_3\,\Delta\kappa_3)^T$, $\boldsymbol{\kappa}^2_{n_1n_2n_3} = (n_1\,\Delta\kappa_1, n_2\,\Delta\kappa_2, -n_3\,\Delta\kappa_3)^T$, $\boldsymbol{\kappa}^3_{n_1n_2n_3}=(n_1\,\Delta\kappa_1, -n_2\,\Delta\kappa_2, n_3\,\Delta\kappa_3)^T$ and $\boldsymbol{\kappa}^4_{n_1n_2n_3} = (n_1\,\Delta\kappa_1, -n_2\,\Delta\kappa_2, -n_3\,\Delta\kappa_3)^T$. $\boldsymbol{\Phi}^{\alpha}_{l,n_1\dots n_d}(\theta)$ are $m \times 2^{d-1}$ independent sequences of independent random phase angles drawn at any wave number $\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d}$ from a uniform distribution in the range $[0,\ 2\pi]$.

Random fields generated with relation (41) are periodic along the x_i axes, $i \in [1,\dots,d]$, with period:

$$L^0_i := \frac{2\pi}{\Delta\kappa_i} \tag{44}$$

Also, the values of the field are bounded according to:

$$g_j(\mathbf{x};\theta) \leq 2\sqrt{\Delta\kappa_1 \dots \Delta\kappa_d} \sum_{l=1}^{m} \sum_{n_1=0}^{N_1-1} \dots \sum_{n_d=0}^{N_d-1} \sum_{\alpha=1}^{2^{d-1}} |H_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\dots n_d})| \tag{45}$$

It has been shown that the random fields generated according to Eq. (41) have the following properties [6, 7, 26, 33, 34]:

1. They tend to be standard Gaussian as $N_i \to +\infty$, $\forall i \in [1,\dots,d]$; rate of convergence is investigated in [33].
2. They ensemble auto- and cross-correlations are identical to the target functions.

3. Each realization is ergodic in mean and correlation (spatial mean and correlation over domain $\mathscr{R}$ are equal to ensemble mean and correlation) when the size of the spatial domain $|\mathscr{R}|$ tends to be infinite in every directions.
4. Each realization is ergodic in mean as $|\mathscr{R}| = L_1^0 \times \cdots \times L_d^0$ (see Eq. 44).

For properties 3 and 4 to be true, this further condition has to be satisified: $H_{jk}(\kappa_1, \ldots, \kappa_d) = 0$, $(j, k) \in [1, \ldots, m]^2$, as any of the κ_i, $i \in [1, \ldots, d]$, is equal to zero.

5. As discussed in Appendix 1, the random fields generated from Eq. (41) are not ergodic in correlation as $|\mathscr{R}| = L_1^0 \times \cdots \times L_d^0$. However, using properly defined wave-number shifts [7, 26], ergodicity in correlation is recovered on a finite domain as the spatial correlations are calculated over a domain of size $|\mathscr{R}| = m\, L_1^0 \times \cdots \times m\, L_d^0$. In this case, the wave number vector introduced in Eq. (42) is modified so as it also depends on the index l, as follows:

$$\boldsymbol{\kappa}^{\alpha}_{l,n_1\ldots n_d} := \left((n_1 + \frac{l}{m})\Delta\kappa_1 \quad I_2^{\alpha}(n_2 + \frac{l}{m})\Delta\kappa_2 \quad \ldots \quad I_d^{\alpha}(n_d + \frac{l}{m})\Delta\kappa_d \right)^T \tag{46}$$

Besides, as wave-number shifts are introduced, the condition that functions $H_{jk}(\kappa_1, \ldots, \kappa_d)$ be equal to zero as any $\kappa_i = 0$ can be removed for properties 3 and 4 to be valid.

2.3.3 Translation to Non-Gaussian Stochastic Vector Fields

The approach presented above generates m zero-mean unit-variance homogeneous Gaussian stochastic fields $\mathfrak{g}_j(\mathbf{x}; \theta)$, $j \in [1, \ldots m]$, with cross-correlation matrix $\mathbf{R}^g(\boldsymbol{\xi})$. m homogeneous non-Gaussian stochastic translation fields $\mathfrak{f}_j(\mathbf{x}; \theta)$ can be obtained from their Gaussian counterparts $\mathfrak{g}_j(\mathbf{x}; \theta)$. The translation fields are defined by the following memoryless—meaning that the outputs at any point $\mathbf{x}$ do not depend on the inputs at any other point—mapping:

$$\mathfrak{f}_j(\mathbf{x}) = \mathscr{F}^{-1}_{\mathfrak{f}_j}\left(\mathscr{F}_{\mathfrak{g}_j}(\mathfrak{g}_j(\mathbf{x}))\right) = F_j\left(\mathfrak{g}_j(\mathbf{x})\right) \quad , \quad j \in [1, \ldots, m] \tag{47}$$

where $\mathscr{F}_{\mathfrak{g}_j}$ is the standard Gaussian cumulative density function (CDF) of the random variables $\mathfrak{g}_j(\mathbf{x})$, $\mathscr{F}^{-1}_{\mathfrak{f}_j}$ the inverse of the marginal CDF of the non-Gaussian random variables $\mathfrak{f}_j(\mathbf{x})$, and $F_j = \mathscr{F}^{-1}_{\mathfrak{f}_j} \circ \mathscr{F}_{\mathfrak{g}_j}$.

Then, the components of the non-Gaussian correlation matrix can be computed as:

$$\begin{aligned} R^{f}_{jk}(\boldsymbol{\xi}, \boldsymbol{\rho}^g) :=& \mathbb{E}[\mathfrak{f}_j(\mathbf{x})\, \mathfrak{f}_k(\mathbf{x} + \boldsymbol{\xi})] \\ :=& \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F_j\left(\mathfrak{g}_j(\mathbf{x})\right)\, F_k\left(\mathfrak{g}_k(\mathbf{x} + \boldsymbol{\xi})\right) \\ & \times p^{G}_{\mathfrak{g}_j \mathfrak{g}_k}\left(g_j(\mathbf{x}), g_k(\mathbf{x} + \boldsymbol{\xi}); \rho^g_{jk}(\boldsymbol{\xi})\right)\, dg_j(\mathbf{x})\, dg_k(\mathbf{x} + \boldsymbol{\xi}) \end{aligned} \tag{48}$$

where $p^G_{\mathfrak{g}_j\mathfrak{g}_k}$ denotes the standard Gaussian joint PDF of the two random variables $\mathfrak{g}_j(\mathbf{x})$ and $\mathfrak{g}_k(\mathbf{x}+\boldsymbol{\xi})$. Note that in the case of standard Gaussian distribution, we have $\rho^{\boldsymbol{g}}_{jk}(\boldsymbol{\xi}) = R^{\boldsymbol{g}}_{jk}(\boldsymbol{\xi})$ (see Eqs. 29 and 30).

In practice, one is interested in generating realizations of non-Gaussian random fields $\mathfrak{f}_j$ with targeted marginal PDF $\mathscr{F}^0_{\mathfrak{f}_j}$ and targeted cross-correlation matrix $\mathbf{R}^{f\,0}(\boldsymbol{\xi})$. To that purpose, the cross-correlation matrix $\mathbf{R}^g(\boldsymbol{\xi})$ of the underlying standard Gaussian fields $\mathfrak{g}_j(\mathbf{x};\theta)$ has to be determined. We recall (see Eqs. 29 and 30) that:

$$\rho^{\boldsymbol{f}}_{jk}(\boldsymbol{\xi}, \rho^{\boldsymbol{g}}_{jk}) := \frac{R^{\boldsymbol{f}}_{jk}(\boldsymbol{\xi}, \rho^{\boldsymbol{g}}_{jk}) - \mu_{\mathfrak{f}_j}\mu_{\mathfrak{f}_k}}{s_{\mathfrak{f}_j}s_{\mathfrak{f}_k}} \tag{49}$$

Suppose that, $\forall (j,k) \in [1,\ldots,m]^2$, we calculate from relations (48) and (49) the two quantities $\rho^{\boldsymbol{f}\,min}_{jk}(\boldsymbol{\xi}) = \rho^{\boldsymbol{f}}_{jk}(\boldsymbol{\xi}, -1)$ and $\rho^{\boldsymbol{f}\,max}_{jk}(\boldsymbol{\xi}) = \rho^{\boldsymbol{f}}_{jk}(\boldsymbol{\xi}, +1)$. Following [10, 11], if the functions $\rho^{\boldsymbol{f}}_{jk}(\boldsymbol{\xi})$ all fall in the range $[\rho^{\boldsymbol{f}\,min}_{jk}(\boldsymbol{\xi}),\ \rho^{\boldsymbol{f}\,max}_{jk}(\boldsymbol{\xi})]$, $\forall \boldsymbol{\xi}$, then Eq. (48) can be analytically or numerically inverted to calculate a unique $\boldsymbol{\rho}^{\boldsymbol{g}}(\boldsymbol{\xi})$. Besides, it must be verified that the matrix $\boldsymbol{\rho}^{\boldsymbol{g}}(\boldsymbol{\xi})$ really is a correlation matrix, namely that the auto-correlation functions $\rho^{\boldsymbol{g}}_{jj}(\boldsymbol{\xi})$, $j \in [1,\ldots m]$, as well as the correlation matrix $\boldsymbol{\rho}^{\boldsymbol{g}}(\boldsymbol{\xi})$ are positive semi-definite for every separation distance $\boldsymbol{\xi}$.

Inverting relation (48) is not always possible, and when it is not, cross-correlation matrix $\mathbf{R}^f(\boldsymbol{\xi})$ and marginal CDFs $\mathscr{F}_{\mathfrak{f}_j(\mathbf{x})}$, $j \in [1,\ldots,m]$, are said to be "incompatible" [31]. In this case, the iterative method presented in [31] can be implemented (see also [3]). With this method, the non-Gaussian CDFs are taken as $\mathscr{F}_{\mathfrak{f}_j} = \mathscr{F}^0_{\mathfrak{f}_j}$ and the correlation functions $R^{\boldsymbol{g}}_{jk}(\boldsymbol{\xi})$ of the underlying standard Gaussian fields are iteratively modified until the correlation functions of the translated fields are sufficiently close to the targets: $\mathbf{R}^f(\boldsymbol{\xi}) \approx \mathbf{R}^{f\,0}(\boldsymbol{\xi})$.

3 Numerical Implementation

3.1 Random Vector Fields Generation Using FFT

For numerical implementation, Eq. (41) is rewritten as:

$$B^{\alpha}_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\ldots n_d};\theta) := 2\sqrt{\Delta\kappa_1\ldots\Delta\kappa_d}\ |H_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\ldots n_d})|\, e^{-i\,\varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\ldots n_d})}\, e^{i\,\Phi^{\alpha}_{l,n_1\ldots n_d}(\theta)} \tag{50}$$

$$G^{\alpha}_{jl}(\mathbf{x}_{m_1\ldots m_d};\theta) := \sum_{n_1=0}^{N_1-1}\cdots\sum_{n_d=0}^{N_d-1} B^{\alpha}_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1\ldots n_d};\theta)\, e^{i\,\boldsymbol{\kappa}^{\alpha}_{n_1\ldots n_d}\cdot\mathbf{x}_{m_1\ldots m_d}} \tag{51}$$

$$g_j(\mathbf{x}_{m_1\ldots m_d};\theta) = Re\sum_{l=1}^{m}\sum_{\alpha=1}^{2^{d-1}} G^{\alpha}_{jl}(\mathbf{x}_{m_1\ldots m_d};\theta) \tag{52}$$

for $j \in [1, \dots, m]$, with $N_i \in \mathbb{N}^\star$ and $N_i \to +\infty$ for all $i \in [1, \dots, d]$, where $Re(z)$ is the real part of the complex number z, and where we introduced:

$$\mathbf{x}_{m_1 \dots m_d} := (m_1 \Delta x_1 \quad m_2 \Delta x_2 \quad \dots \quad m_d \Delta x_d)^T \quad , \quad m_i \in [0, \dots, M_i - 1] \tag{53}$$

Relation (51) can be numerically computed in an efficient way using fast Fourier transform (FFT) algorithm.

The random fields are generated over a spatial period $L_1^0 \times \cdots \times L_d^0$ setting

$$\Delta x_i := \frac{2\pi}{M_i \, \Delta\kappa_i} \tag{54}$$

with $M_i \geq 2\,N_i$ to avoid aliasing. Introducing definitions (42) and (57), along with (44) and (54) in (51), and reminding that $B_{jl}^{\alpha}(\boldsymbol{\kappa}_{n_1 \dots n_d}^{\alpha}; \theta) = 0$ for any $n_i \geq N_i$, that is $n_i \Delta\kappa_i \geq \kappa_{ui}$, it comes:

$$G_{jl}^{\alpha}(\mathbf{x}_{m_1 \dots m_d}; \theta) := \sum_{n_1=0}^{M_1-1} \dots \left(\sum_{n_d=0}^{M_d-1} B_{jl}^{\alpha}(\boldsymbol{\kappa}_{n_1 \dots n_d}^{\alpha}; \theta)\, e^{2i\pi\, I_d^{\alpha} \frac{m_d n_d}{M_d}} \right) \dots e^{2i\pi\, I_1^{\alpha} \frac{m_1 n_1}{M_1}} \tag{55}$$

where a sequence of d Fourier or inverse Fourier transforms, according to the sign of I_i^{α}, can be recognized.

Note that in the case wave-number shifts are applied, the equations in this Sect. 3.1 can be straightforwardly adapted by introducing Eq. (46) instead of (43) for the wave numbers vector. The side effect is that the periods over which random fields are generated are elongated as:

$$L_i^0 = \frac{2\pi}{\Delta\kappa_i} \quad \to \quad L_i^0 = m \times \frac{2\pi}{\Delta\kappa_i} \quad , \quad i \in [1, \dots, d] \tag{56}$$

and with the random fields generated over the grid (compare to (53)):

$$\mathbf{x}_{m_1 \dots m_d} := (m_1 \Delta x_1 \quad m_2 \Delta x_2 \quad \dots \quad m_d \Delta x_d)^T \quad , \quad m_i \in [0, \dots, m \times M_i - 1] \tag{57}$$

3.2 Material Response at Mesoscale

The components of the elasticity tensor **C**, damage-plasticity ratio r and yield stress σ_y are parameters of the material model introduced in Sect. 2.2. These parameters are realizations of random variables over the ED $\mathscr{R}$, according to the random vector fields generated as presented in the previous section. For the sake of readability, reference to the spatial position ($\mathbf{x}$) and to the random experiment (θ) are dropped in this section.

3.2.1 Discrete Evolution Equations

We introduce the discrete process for the pseudo-time: $\mathscr{T}_0^T = \{t_n, n \in [0, \ldots, N_T]\}$ with $t_0 = 0$, $t_{N_T} = T$, and the pseudo-time increment $t_{n+1} - t_n := \Delta t$.

The numerical integration of the evolution of the internal variables over the process $\mathscr{T}_N$ is performed using the unconditionally stable backward Euler time integration scheme. Accordingly, the evolution of the internal variables (see Eqs. 18 and 19) are implemented as:

$$\mathbf{D}_{n+1} : \boldsymbol{\sigma}_{n+1} = \mathbf{D}_n : \boldsymbol{\sigma}_{n+1} + r\,\gamma_{n+1}\boldsymbol{\nu}_{n+1} \tag{58}$$

$$\boldsymbol{\varepsilon}_{n+1}^p = \boldsymbol{\varepsilon}_n^p + (1-r)\gamma_{n+1}\boldsymbol{\nu}_{n+1} \tag{59}$$

where $\gamma_{n+1} := \gamma(t_{n+1}) := \Delta t\,\dot{\gamma}_{n+1}$.

Besides, considering Eq. (13), we have for the stress tensor:

$$\mathbf{D}_{n+1} : \boldsymbol{\sigma}_{n+1} := \boldsymbol{\varepsilon}_{n+1}^d := \boldsymbol{\varepsilon}_{n+1} - \boldsymbol{\varepsilon}_{n+1}^p \tag{60}$$

$$\Rightarrow \mathbf{D}_n : \boldsymbol{\sigma}_{n+1} = \boldsymbol{\varepsilon}_{n+1} - \boldsymbol{\varepsilon}_n^p - \gamma_{n+1}\boldsymbol{\nu}_{n+1} \tag{61}$$

Finally, the tangent modulus in Eq. (25) is computed as:

$$\boldsymbol{\lambda}_{n+1} = \begin{cases} \mathbf{D}_{n+1}^{-1} & \text{if} \quad \gamma_{n+1} = 0 \quad (\phi(\boldsymbol{\sigma}_{n+1}) < 0) \\ \mathbf{0} & \text{if} \quad \gamma_{n+1} > 0 \quad (\phi(\boldsymbol{\sigma}_{n+1}) = 0\,;\ \dot{\phi}(\boldsymbol{\sigma}_{n+1}) = 0) \end{cases} \tag{62}$$

3.2.2 Solution Procedure

The problem to be solved at any material point $\mathbf{x}$ of the mesoscale reads:
Given $\boldsymbol{\varepsilon}_{n+1} = \boldsymbol{\varepsilon}_n + \Delta\boldsymbol{\varepsilon}_{n+1}$, *find* $\gamma_{n+1}\boldsymbol{\nu}_{n+1}$ *such that* $\phi_{n+1} \leq 0$. This is solved using a so-called return-mapping algorithm (see e.g. [13, 35]) where a trial state is first considered and followed by a corrective step if required:

1. **Trial state:**
   It is assumed that there is no inelastic evolution due to deformation increment $\Delta\boldsymbol{\varepsilon}_{n+1}$, that is $\gamma_{n+1} = 0$. Accordingly, the internal variables remain unchanged: $\mathbf{D}_{n+1}^{trial} = \mathbf{D}_n$ and $\boldsymbol{\varepsilon}_{n+1}^{p,trial} = \boldsymbol{\varepsilon}_n^p$. The trial stress along with the trial criterium function can then be computed as:

$$\boldsymbol{\sigma}_{n+1}^{trial} = \mathbf{D}_n^{-1} : (\boldsymbol{\varepsilon}_{n+1} - \boldsymbol{\varepsilon}_n^p) \tag{63}$$

$$\phi_{n+1}^{trial} = h(\boldsymbol{\sigma}_{n+1}^{trial}) - \sigma_y \tag{64}$$

   The admissibility of this trial state then has to be checked:

   - If $\phi_{n+1}^{trial} \leq 0$, the trial state is admissible and the local variables are updated accordingly: $\boldsymbol{\sigma}_{n+1} = \boldsymbol{\sigma}_{n+1}^{trial}$, $\mathbf{D}_{n+1} = \mathbf{D}_n$, $\boldsymbol{\varepsilon}_{n+1}^p = \boldsymbol{\varepsilon}_n^p$. Besides, the tangent modulus is: $\boldsymbol{\lambda}_{n+1} = \mathbf{D}_n^{-1}$.

- If $\phi^{trial}_{n+1} > 0$, the trial state is not admissible and it has to be corrected as described in the next step 'correction'.

2. **Correction:**
If trial state is not admissible, then $\gamma_{n+1} > 0$ and, according to Eq. (20), the relation $\phi_{n+1} = 0$ has to be satisfied. Solving $\phi_{n+1} = 0$ yields $\gamma_{n+1}\boldsymbol{\nu}_{n+1}$.
Then, the stresses can be calculated following Eq. (61), the internal variables are updated according to Eqs. (58) and (59), and finally tangent modulus reads $\boldsymbol{\lambda}_{n+1} = \mathbf{0}$.

3.3 Material Response at Macroscale

In this section, we derive the equations to be implemented for the numerical computation of the response of the ED at macroscale, that is the macroscopic behavior law $\dot{\boldsymbol{\Sigma}} = \mathbf{L} : \dot{\mathbf{E}}$, where $\boldsymbol{\Sigma}$ and $\mathbf{E}$ are the macroscopic stress and strain tensors while $\mathbf{L}$ denotes the homogenized tangent modulus at macroscale.

3.3.1 Discrete Governing Equations in the ED

The weak form of the boundary value problem in Eq. (1) reads:

$$\begin{aligned} 0 &= \int_{\mathscr{R}} \delta\mathbf{u} \cdot \mathbf{div}\, \boldsymbol{\sigma}\, d\mathscr{R} \\ &= \int_{\mathscr{R}} \boldsymbol{\nabla}^s \delta\mathbf{u} : \boldsymbol{\sigma}(\boldsymbol{\nabla}^s \mathbf{u})\, d\mathscr{R} - \int_{\partial\mathscr{R}} \delta\mathbf{u} \cdot \mathbf{t}\, d\partial\mathscr{R} \end{aligned} \tag{65}$$

where $\delta\mathbf{u}$ is any virtual displacement field that satisfies $\delta\mathbf{u} = \mathbf{0}$ on $\partial_u\mathscr{R}$.

Finite element (FE) method is used to approximate the displacement field over the ED. Accordingly, $\mathscr{R}$ is meshed into N_{el} elements $\mathscr{R}^e$ such that $\bigcup_{e=1}^{N_{el}} \mathscr{R}^e = \mathscr{R}$. Then, in each element, displacement fields is computed as (see e.g. [40]):

$$\mathbf{u}(\mathbf{x}, t)|_{\mathscr{R}^e} = \mathbf{N}^e(\mathbf{x})\, \mathbf{d}^e(t) \tag{66}$$

where $\mathbf{N}^e(\mathbf{x})$ contains the element shape functions and $\mathbf{d}^e(t)$ are the displacements at the nodes of the FE mesh. Equation (65) can then be rewritten as:

$$\begin{aligned} 0 := \mathop{\mathsf{A}}_{e=1}^{N_{el}} \Bigg\{ &\int_{\mathscr{R}^e} \text{sym}\left[\boldsymbol{\nabla}\left(\mathbf{N}^e\, \delta\mathbf{d}^e\right)\right] : \boldsymbol{\sigma}\left(\text{sym}\left[\boldsymbol{\nabla}\left(\mathbf{N}^e\, \mathbf{d}^e\right)\right]\right) d\mathscr{R}^e \\ &- \int_{\partial\mathscr{R}^e} \mathbf{N}^e\, \delta\mathbf{d}^e \cdot \mathbf{t}\, d\partial\mathscr{R}^e \Bigg\} \end{aligned} \tag{67}$$

where $\overset{N_{el}}{\underset{e=1}{\mathsf{A}}}$ denotes the finite element assembly procedure, and $\partial\mathscr{R}^e$ denotes the portion, if any, of the boundary of the element e that is also a part of the boundary of the discretized domain $\partial\mathscr{R}$.

Matrix notations can be conveniently adopted at this stage, so that the elements of the symmetric second-order tensors $\boldsymbol{\sigma}$ and ε are written as vectors:

$$\boldsymbol{\sigma} \;\rightarrow \overline{\boldsymbol{\sigma}} \quad ; \quad \varepsilon := \text{ sym } \left[\nabla\left(\mathbf{N}^e\,\mathbf{d}^e\right)\right] \;\rightarrow\; \overline{\varepsilon} := \mathbf{B}^e\,\mathbf{d}^e \tag{68}$$

The matrix $\mathbf{B}^e$ is composed of derivatives of the element shape functions. With these notations, we have $\varepsilon : \boldsymbol{\sigma} \;\rightarrow \overline{\varepsilon}^T\,\overline{\boldsymbol{\sigma}}$, so that (67) can be rewritten as:

$$0 := \overset{N_{el}}{\underset{e=1}{\mathsf{A}}}\,\delta\mathbf{d}^{eT}\left\{\int_{\mathscr{R}^e}\mathbf{B}^{eT}\overline{\boldsymbol{\sigma}}(\overline{\varepsilon})d\mathscr{R}^e - \int_{\partial\mathscr{R}^e}\mathbf{N}^{eT}\,\mathbf{t}\,d\partial\mathscr{R}^e\right\} \tag{69}$$

Because the equation above has to be satisfied for any virtual nodal displacements vector $\delta\mathbf{d}^e$ that satisfies $\delta\mathbf{d}^e = \mathbf{0}$ at any node on $\partial_u\mathscr{R}$, this is finally the following set of nonlinear equations that has to be solved:

$$\mathbf{0} := \mathbf{r}(\mathbf{d}) := \mathbf{f}^{int}(\mathbf{d}) - \mathbf{f}^{ext} \tag{70}$$

where:

$$\mathbf{f}^{int}(\mathbf{d}_{n+1}) := \overset{N_{el}}{\underset{e=1}{\mathsf{A}}}\int_{\mathscr{R}^e}\mathbf{B}^{eT}\,\overline{\boldsymbol{\sigma}}_{n+1}\,d\mathscr{R}^e; \quad \mathbf{f}^{ext}_{n+1} := \overset{N_{el}}{\underset{e=1}{\mathsf{A}}}\int_{\partial\mathscr{R}^e}\mathbf{N}^{eT}\,\mathbf{t}^e_{n+1}\,d\partial\mathscr{R}^e \tag{71}$$

Here, we added explicit reference to the time discretization to recall that it is a nonlinear evolution problem that has to be solved.

3.3.2 Solution Procedure

First, we separate the degrees of freedom of the N_{bo} nodes that are on the boundary $\partial\mathscr{R}$ of the ED—denoted by the subscript $\bar{u}$—from those pertaining to its interior—denoted by the subscript u—and rearrange them so that:

$$\mathbf{d} = \begin{pmatrix}\mathbf{d}_u\\ \mathbf{d}_{\bar{u}}\end{pmatrix} \quad \text{and} \quad \mathbf{r}(\mathbf{d}) = \begin{pmatrix}\mathbf{r}_u(\mathbf{d})\\ \mathbf{r}_{\bar{u}}(\mathbf{d})\end{pmatrix} := \begin{pmatrix}\mathbf{f}^{int}_u(\mathbf{d})\\ \mathbf{f}^{int}_{\bar{u}}(\mathbf{d})\end{pmatrix} - \begin{pmatrix}\mathbf{0}\\ \mathbf{f}^{ext}_{\bar{u}}\end{pmatrix} \tag{72}$$

$\mathbf{f}^{ext}_u = \mathbf{0}$ because there is no external forces applied on the interior nodes.

As external forces $\mathbf{f}^{ext}_{\bar{u}}$ increase by $\Delta\mathbf{f}^{ext}_{\bar{u}}$ and displacements $\mathbf{d}$ increase by $\Delta\mathbf{d}$, the residual is linearized such that the problem to be solved now reads:

$$\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} := \begin{pmatrix}\mathbf{f}^{int}_u(\mathbf{d})\\ \mathbf{f}^{int}_{\bar{u}}(\mathbf{d})\end{pmatrix} + \begin{pmatrix}\mathbf{K}^{tan}_{uu} & \mathbf{K}^{tan}_{u\bar{u}}\\ \mathbf{K}^{tan}_{\bar{u}u} & \mathbf{K}^{tan}_{\bar{u}\bar{u}}\end{pmatrix}\begin{pmatrix}\Delta\mathbf{d}_u\\ \Delta\mathbf{d}_{\bar{u}}\end{pmatrix} - \begin{pmatrix}\mathbf{0}\\ \mathbf{f}^{ext}_{\bar{u}} + \Delta\mathbf{f}^{ext}_{\bar{u}}\end{pmatrix} \tag{73}$$

where $\mathbf{K}^{tan}$ is the tangent stiffness matrix defined as:

$$\mathbf{K}^{tan} := \frac{\partial \mathbf{f}^{int}(\mathbf{d})}{\partial \mathbf{d}} = \int_{\mathscr{R}} \mathbf{B}^T \, \overline{\boldsymbol{\lambda}} \, \mathbf{B} \, d\mathscr{R} = \mathop{\mathsf{A}}_{e=1}^{N_{el}} \int_{\mathscr{R}^e} \mathbf{B}^{eT} \, \overline{\boldsymbol{\lambda}} \, \mathbf{B}^e \, d\mathscr{R}^e \tag{74}$$

$\overline{\boldsymbol{\lambda}} = \partial \overline{\boldsymbol{\sigma}} / \partial \overline{\boldsymbol{\varepsilon}}$ is the matrix form of the material tangent modulus at mesoscale as introduced in Sect. 2.2.5. Element tangent stiffness and internal forces vector are numerically computed as:

$$\mathbf{K}^{e,tan} := \int_{\mathscr{R}^e} \mathbf{B}^{eT} \, \overline{\boldsymbol{\lambda}} \, \mathbf{B}^e \, d\mathscr{R}^e \approx \sum_{l=1}^{N_{IP}} \mathbf{B}_l^{eT} \, \overline{\boldsymbol{\lambda}}_l \, \mathbf{B}_l^e \, w_l \tag{75}$$

$$\mathbf{f}^{e,int} := \int_{\mathscr{R}^e} \mathbf{B}^{eT} \, \overline{\boldsymbol{\sigma}} \, d\mathscr{R}^e \approx \sum_{l=1}^{N_{IP}} \mathbf{B}_l^{eT} \, \overline{\boldsymbol{\sigma}}_l \, w_l \tag{76}$$

where w_l are the weights associated to the N_{IP} quadrature points.

The macroscopic response of the ED $\mathscr{R}$ is then computed from its description at mesoscale as follows:

1. **Updating of the imposed displacements on $\partial\mathscr{R}$:**
   Impose displacement $\mathbf{d}_{\bar{u}} = \mathbf{d}_{\bar{u}} + \Delta\mathbf{d}_{\bar{u}}$ on the boundary nodes. We recall that we only consider the case of linear displacements imposed all over the boundary of the ED (see Eq. 3). Following the work presented in [22, 30], we can write these imposed displacements at any node q of the N_{bo} nodes of the boundary as:

   $$\Delta\mathbf{d}_q = \mathbf{W}_q^T \, \Delta\overline{\mathbf{E}} \quad \Rightarrow \quad \Delta\mathbf{d}_{\bar{u}} = \left[\mathbf{W}_1 \, \mathbf{W}_2 \, \ldots \, \mathbf{W}_{N_{bo}}\right]^T \Delta\overline{\mathbf{E}} \;=\; \mathbf{W}^T \Delta\overline{\mathbf{E}} \tag{77}$$

   where $\overline{\mathbf{E}}$ is the matrix form of the strain tensor and where the $\mathbf{W}_q$s are geometric matrices built from the coordinates $\mathbf{x}_q$ of the boundary node q.
2. **Iterative updating of $\Delta\mathbf{d}_u$:**
   Because Eq. (73) are nonlinear, we use Newton-Raphson procedure to iteratively seek $\mathbf{d}_u$ as $\mathbf{d}_u^{(k)} = \mathbf{0} + \Delta\mathbf{d}_u^{(1)} + \cdots + \Delta\mathbf{d}_u^{(k)} + \cdots$ until $\mathbf{f}_u^{int}(\mathbf{d}_u^{(l)}) \cdot \Delta\mathbf{d}_u^{(l)} < tol$ ($\mathbf{r}_u = \mathbf{f}_u^{int}$). Displacements $\mathbf{d}_{\bar{u}}$ on the boundary $\partial\mathscr{R}$ are known from step 1 above, which means that at any iteration k, $\Delta\mathbf{d}_{\bar{u}}^{(k)} = \mathbf{0}$. Then, according to Eq. (73), we have at every iteration:

   $$\Delta\mathbf{d}_u^{(k+1)} = -\left(\mathbf{K}_{uu}^{tan,(k)}\right)^{-1} \mathbf{f}_u^{int}(\mathbf{d}^{(k)}) \tag{78}$$

3. **Compute stresses at macroscale:**
   First, the external forces vectors $\mathbf{f}_q^{ext}$ (reactions) at the nodes of the boundary are retrieved as:

   $$\mathbf{r}_{\bar{u}}(\mathbf{d}^{(l)}) := \mathbf{0} \quad \Rightarrow \quad \mathbf{f}_q^{ext} = \mathbf{f}_q^{int}(\mathbf{d}^{(l)}) \quad , \quad q \in [1, \ldots, N_{bo}] \tag{79}$$

Then, the approximation $\mathbf{t}(\mathbf{x}_q)d\partial\mathscr{R} \approx \mathbf{f}_q^{ext}$ is introduced in Eq. (5) and we consider the matrix form $\overline{\boldsymbol{\Sigma}}$ of the stress tensor, which yields [22]:

$$\Delta\overline{\boldsymbol{\Sigma}} = \frac{1}{|\mathscr{R}|}\sum_{q=1}^{N_{bo}} \mathbf{W}_q\,\Delta\mathbf{f}_q^{ext} = \frac{1}{|\mathscr{R}|}\mathbf{W}\,\Delta\mathbf{f}_{\bar{u}}^{ext} \tag{80}$$

4. **Compute tangent modulus at macroscale:**
Considering an equilibrium state, we have $\mathbf{f}_u^{int}(\mathbf{d}) = \mathbf{0}$ and $\mathbf{f}_{\bar{u}}^{int}(\mathbf{d}) = \mathbf{f}_{\bar{u}}^{ext}$. Then, according to Eq. (73), it comes:

$$\Delta\mathbf{d}_u = -\left(\mathbf{K}_{uu}^{tan}\right)^{-1}\mathbf{K}_{u\bar{u}}^{tan}\,\Delta\mathbf{d}_{\bar{u}} \quad \Rightarrow \quad \Delta\mathbf{f}_{\bar{u}}^{ext} = \tilde{\mathbf{K}}_{\bar{u}\bar{u}}^{tan}\,\Delta\mathbf{d}_{\bar{u}} \tag{81}$$

where $\tilde{\mathbf{K}}_{\bar{u}\bar{u}} = \mathbf{K}_{\bar{u}\bar{u}} - \mathbf{K}_{\bar{u}u}\,\mathbf{K}_{uu}^{-1}\,\mathbf{K}_{u\bar{u}}$. Now, combining Eqs. (80), (81) and (77), it comes the following expression for the matrix form of the tangent modulus at macroscale:

$$\overline{\mathbf{L}} := \frac{\Delta\overline{\boldsymbol{\Sigma}}}{\Delta\overline{\mathbf{E}}} = \frac{1}{|\mathscr{R}|}\mathbf{W}\,\tilde{\mathbf{K}}_{\bar{u}\bar{u}}\,\mathbf{W}^T \tag{82}$$

It has to be reminded that the macroscopic stresses and tangent moduli are computed for a given realization θ of the random fields: we have $\overline{\boldsymbol{\Sigma}} = \overline{\boldsymbol{\Sigma}}(\mathbf{X},\theta)$ and $\overline{\mathbf{L}} = \overline{\mathbf{L}}(\mathbf{X},\theta)$. Consequently, there is no guarantee at this point that these quantities are representative of the macroscopic behavior of the material. However, it will be shown in the numerical applications below that for particular structures of the random vector fields that describe an equivalent mesoscale for the material, these macroscopic quantities are almost independent of the realization of the vector fields.

4 Numerical Applications

The purpose of the following numerical applications is twofold. (i) It is first demonstrated in this section that the random vector fields can be parameterized such that a homogeneous material response can be retrieved at macroscale without stochastic homogenization. In this case, macroscopic response does not depend on the realization of the random vector fields that represent variability at an underlying equivalent heterogeneous mesoscale: any realization of the meso-structure yields the same macroscopic response. Consequently, the computational effort is contained at the mesoscale where the nonlinear response of numerous material points has to be computed. (ii) We remind that, because only homogeneous displacement boundary conditions are considered in this work, the homogenous response so retrieved at macroscale is a priori dependent on the boundary conditions. This issue is out of the scope here where we focus on showing that the proposed approach can represent salient features of the concrete macroscopic response in compressive cyclic loading while such features are not explicitly present at the mesoscale (emergence of a macroscopic response).

4.1 1D Homogenized Response at Macroscale

Throughout this section, we only consider uni-dimensional (1D) material behavior in uniaxial loading at any point $\mathbf{X}$ of the macroscale. Consequently, strain and stress vectors $\overline{\mathbf{E}}(\mathbf{X}, t)$ and $\overline{\boldsymbol{\Sigma}}(\mathbf{X}, t)$ degenerate into scalar quantities, respectively $\overline{E}_{33} := \overline{E}$ and $\overline{\Sigma}_{33} := \overline{\Sigma}$.

4.1.1 Spatial Discretization at E-mesoscale

Accordingly, elementary domain (ED) $\mathscr{R}(\mathbf{X})$ is discretized in the framework of the Finite Element (FE) method as a series of N_{el} adjacent two-node bar elements as shown in Fig. 5. The elements do not have common nodes, they are connected through the boundary conditions at $x_3 = 0$ and $x_3 = a_3$. Each node of the FE mesh has one degree of freedom along x_3-axis; besides, each of these nodes belongs to the boundary $\partial\mathscr{R}$ of the ED, that is $\mathbf{d}_{\bar{u}} = \mathbf{d}$ where:

$$\mathbf{d} := \left(d_1^1 \; d_2^1 \; \ldots \; d_1^{N_{el}} \; d_2^{N_{el}}\right)^T \tag{83}$$

Then, homogeneous kinematic boundary conditions are imposed such that:

$$\Delta\mathbf{d} = \mathbf{W}^T \, \Delta\overline{E} \qquad \text{with} \qquad \mathbf{W} = (0 \; a_3 \; \ldots \; 0 \; a_3) \tag{84}$$

At the bar element level: $\mathbf{d}^e = (d_1^e \; d_2^e)^T = (0 \; a_3)^T \, \overline{E}, \forall e \in [1, \ldots, N_{el}]$. Besides, the shape functions are:

$$\mathbf{N}^e = \left(1 - \frac{x_3}{a_3} \;\; \frac{x_3}{a_3}\right) \qquad \Rightarrow \qquad \mathbf{B}^e = \left(-\frac{1}{a_3} \;\; \frac{1}{a_3}\right) \tag{85}$$

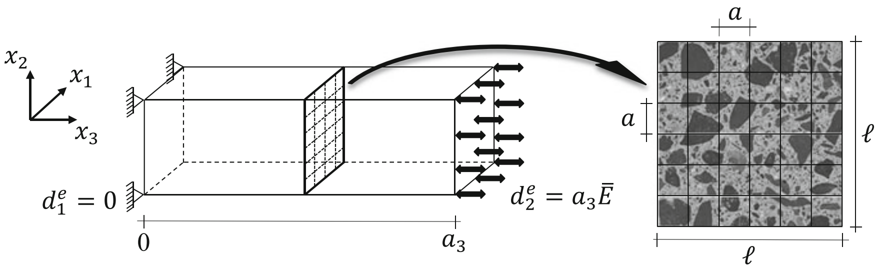


Fig. 5 Concrete ED $\mathscr{R}$ at material point $\mathbf{X}$ of the macroscale. 1D material response only is considered. ED is discretized into $N_{el} = M_f \times M_f$ adjacent bar elements of length a_3 and cross-section $a \times a$. Zero displacement is imposed on the left-hand boundary ($x_3 = 0$) and homogeneous displacement $\bar{u} = a_3\overline{E}$ is imposed all over the right-hand boundary ($x_3 = a_3$) along the x_3-axis

Also, only one numerical integration point is considered along the x_3-axis in each bar element. This implies that the heterogeneity of material properties only has to be accounted for over an ED cross-section and not all over the 3D domain $\mathscr{R}$. The size of $\mathscr{R}$ is $|\mathscr{R}| = \ell \times \ell \times a_3$ and bar elements $\mathscr{R}^e$ are assumed to all have the same size:

$$|\mathscr{R}^e| = a \times a \times a_3 \quad \text{with} \quad a := \frac{\ell}{M_f} \quad , \quad M_f \in \mathbb{N}^\star \tag{86}$$

4.1.2 Homogenized Response

FE approximation introduced above yields, $\forall e \in [1, \ldots, N_{el}]$:

$$\overline{\varepsilon}^e := \mathbf{B}^e \, \mathbf{d}^e = \overline{E} \tag{87}$$

Also, tangent stiffness matrix and internal forces vector in Eqs. (74) and (71) reads:

$$\mathbf{K}^{tan} = \begin{pmatrix} \mathbf{K}^{1,tan} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{K}^{N_{el},tan} \end{pmatrix} \quad \text{and} \quad \mathbf{f}^{e,int} = \begin{pmatrix} \mathbf{f}^{1,int} \\ \vdots \\ \mathbf{f}^{N_{el},int} \end{pmatrix} \tag{88}$$

with, $\forall e \in [1, \ldots, N_{el}]$:

$$\mathbf{K}^{e,tan} = \frac{a^2 \, \overline{\lambda}^e}{a_3} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{f}^{e,int} = a^2 \, \overline{\sigma}^e \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{89}$$

where $\overline{\lambda}^e$ and $\overline{\sigma}^e$ are the tangent modulus and stress computed at the numerical integration point in any bar element e given $\overline{\varepsilon}^e = \overline{E}$ according to the procedure described in Sect. 3.2. Finally, the homogenized quantities at macroscale can be computed as:

$$\overline{\Sigma} = \frac{1}{M_f^2} \sum_{e=1}^{N_{el}} \overline{\sigma}^e \quad \text{and} \quad \overline{L} = \frac{1}{M_f^2} \sum_{e=1}^{N_{el}} \overline{\lambda}^e \tag{90}$$

4.2 *Heterogeneous Structure at E-mesoscale*

In this section, it is described how information is transferred from A-mesoscale to E-mesoscale.

4.2.1 Assumptions About the Structure of the Random Vector Fields

The following assumptions, previously introduced in [26] for modeling material properties, significantly simplify the equations introduced in Sect. 3.1:

- The fields have quadrant symmetry, which implies that the cross-correlation matrix is symmetric and real ($\varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) = 0, \forall (j, l)$ and $\forall \boldsymbol{\kappa}^{\alpha}_{n_1 n_2}$);
- The auto-correlation functions $R^0_{jj} = R^0$ are identical for every components of the vector field;
- The cross-correlation functions R^0_{jk}, $j \neq k$ are expressed as $R^0_{jk} = \rho_{jk} R^0$, where the ρ_{jk}, $j, k = 1, 2, 3$, are so-called correlation coefficients between the components C, σ_y and r of the random vector field. They satisfy $-1 \leq \rho_{jk} \leq 1$.

Accordingly, cross-spectral density matrix of the random vector field reads:

$$\mathbf{S}^0(\kappa_1, \kappa_2) = S^0(\kappa_1, \kappa_2)\, \mathbf{s} \quad \text{with} \quad \mathbf{s} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \tag{91}$$

and Cholesky's decomposition can be applied to $\mathbf{s}$ yielding:

$$\mathbf{s} = \mathbf{h}\,\mathbf{h}^T \tag{92}$$

where $\mathbf{h}$ is a lower triangular matrix. Then, matrix $\mathbf{H}$ (see Eq. 50 for instance), reads:

$$\mathbf{H}(\kappa_1, \kappa_2) = \sqrt{S^0(\kappa_1, \kappa_2)}\, \mathbf{h} \tag{93}$$

For auto-correlation function, we choose the following form:

$$R^0(\overline{\xi}_1, \overline{\xi}_2) = s^2 \exp\left(- \left(\frac{\overline{\xi}_1}{\overline{b}_1} \right)^2 - \left(\frac{\overline{\xi}_2}{\overline{b}_2} \right)^2 \right) \tag{94}$$

where s^2 is the variance of the stochastic fields, $\overline{\xi} = \xi/\ell$, $\overline{b} = b/\ell$ is proportional to $\overline{\ell_c} = \ell_c/\ell$ with ℓ_c denoting the so-called correlation length.

Appealing to the Wiener-Khinchin theorem, we have the power spectral density function that corresponds to the Fourier transform of the correlation function:

$$S^0(\overline{\kappa}_1, \overline{\kappa}_2) = s^2 \frac{\overline{b}_1 \overline{b}_2}{4\pi} \exp\left(- \left(\frac{\overline{b}_1 \overline{\kappa}_1}{2} \right)^2 - \left(\frac{\overline{b}_2 \overline{\kappa}_2}{2} \right)^2 \right) \tag{95}$$

where $\overline{\kappa} = \kappa \times \ell$.

We define in a general way the correlation length $\overline{\ell_c}$ as the distance such that $R(\overline{\ell_c}) = \varepsilon_R\, R(0)$ with $0 < \varepsilon_R \ll 1$. From Eq. (94) comes:

$$\overline{\ell_c} = \overline{b} \sqrt{\ln \frac{1}{\varepsilon_R}} \tag{96}$$

which clearly shows how parameter b is related to the correlation length ℓ_c.

Finally, we analogously define the cut-off wave number as $S(\kappa_u) \leq \varepsilon_S\, S(0)$ with $0 < \varepsilon_S \ll 1$, which from Eq. (95) leads to:

$$|\overline{\kappa}_u| \geq \frac{2}{\overline{b}}\sqrt{\ln\frac{1}{\varepsilon_S}} \tag{97}$$

4.2.2 Parameterization for Random Field Discretization

For the numerical applications, random vector fields are generated according to Eq. (52) with wave-number shifts introduced as in Eq. (46).

Hereafter, random fields parameterization is the same in both directions: $N_1 = N_2 = N$, $M_1 = M_2 = M$, $\kappa_{u\,1} = \kappa_{u\,2} = \kappa_u$, and $\ell_{c\,1} = \ell_{c\,2} = \ell_c$. Then, the random fields are periodic along x_1- and x_2-axes with same period:

$$L^0 := m\frac{2\pi\, N}{\kappa_u} \qquad \text{or} \qquad \overline{L^0} := m\frac{2\pi\, N}{\overline{\kappa}_u} \tag{98}$$

with the dimensionless quantities $\overline{L^0} = L^0/\ell$ and $\overline{\kappa}_u = \kappa \times \ell$. Also, random fields are digitized into $m \cdot M \times m \cdot M$ points regularly distributed over a square grid of size $L^0 \times L^0$. The distance between two adjacent points in both directions of the grid is $\Delta x = L^0/(m\,M)$, or $\overline{\Delta x} = \overline{L^0}/(m\,M)$ where $\overline{\Delta x} = \Delta x/\ell$.

To define a straightforward mapping of the random field grid onto the FE mesh over $\mathscr{R}$, we set:

$$\Delta x = a \qquad \Rightarrow \qquad \overline{\Delta x} = \frac{1}{M_f} \qquad \Rightarrow \qquad \overline{L^0} = \frac{m\,M}{M_f} \tag{99}$$

Also, we enforce the following condition to avoid any situation where the random material meso-structure would show some periodicity:

$$\overline{L^0} \geq 1 \qquad \Rightarrow \qquad m\,M \geq M_f \tag{100}$$

Then, combining Eqs. (98) and (99), we have:

$$\overline{\kappa}_u = 2\pi\frac{N}{m\,M}M_f \tag{101}$$

which introduced in relation (97) yields:

$$\frac{M_f}{m\,M} \geq \frac{1}{\pi\,N\overline{b}}\sqrt{\ln\frac{1}{\varepsilon_S}} \tag{102}$$

Finally, recalling Eq. (96), we have the following relations that the parameterization has to satisfy:

$$1 \geq \frac{M_f}{m\,M} \geq \frac{1}{\pi\,N\overline{\ell_c}} \ln \frac{1}{\varepsilon_{RS}} \tag{103}$$

with $0 < \varepsilon_{RS} = \varepsilon_R = \varepsilon_S \ll 1$.

For all the numerical applications presented hereafter, we choose $M_f = 96$, $N = 16$ and $\varepsilon_{RS} = 0.01$. With this parameterization, $\overline{\Delta x} = 1/M_f = 0.010$. Then, to avoid aliasing in the computation of the FFTs (see Sect. 3.1), we take $M \geq 2\,N \geq 32$. With this choice, $m\,M \geq 96 \geq M_f$ so that the left-hand part in (103) is satisfied. The right-hand part in (103) can be rewritten as:

$$\overline{\ell_c} \geq \frac{m\,M}{\pi\,N\,M_f} \ln \frac{1}{\varepsilon_{RS}} = \overline{\ell_c}_{min} \tag{104}$$

4.2.3 Parameterization of the 1D Material Response at E-mesoscale

The material law $\Delta\overline{\sigma}^e = \overline{\lambda}^e\,\Delta\overline{\varepsilon}^e$ considered in these numerical applications corresponds to the 1D version of the equations developed in Sect. 2.2 completed by the set of equations in Appendix 2. Figure 4 shows cyclic compressive response obtained from this model at two material points of E-mesoscale, that is in two different elements of the FE mesh over the elementary domain $\mathscr{R}$.

In each element e of the FE mesh over $\mathscr{R}$, material parameters C^e, σ_y^e and r^e take different values due to material heterogeneities. The spatial variability of these three parameters ($m = 3$) over any cross-section of $\mathscr{R}$ (d=2) is represented by a 3-variate 2-dimensional random vector field that is generated following Sect. 3.1 with wave-number shifts introduced.

Correlation coefficients in Eq. (91) are set to $\rho_{12} = \rho_{13} = \rho_{23} = 0.9$. This corresponds to strongly correlated random fields, which comes from considering that the three parameters all depend on the geometrical structure of concrete at A-mesoscale: aggregates in a hardened cement paste, as illustrated in Fig. 3.

In the absence of experimental evidence about A-mesoscale, we choose uniform distributions for the parameters, except for the elastic modulus. The reason why a log-normal distribution has been retained for C will be apparent in Sect. 4.3.2. Specifically, Table 1 presents the distributions used hereafter to build an E-mesoscale that would yield a macroscopic response exhibiting salient features of concrete 1D response in uniaxial compressive cyclic loading. How to translate Gaussian fields to uniform fields is described in Appendix 3.

Table 1 Distribution laws for the set of heterogeneous material parameters

Parameter	Distribution law	Mean	Std. deviation	COV (%)
C	$log\mathscr{N}(30e^3, 15e^3)$	$\mu_C = 30.0\,\text{GPa}$	$s_C = 15.0\,\text{GPa}$	50.0
σ_y	$\mathscr{U}(0, 70)$	$\mu_{\sigma_y} = 35.0\,\text{MPa}$	$s_{\sigma_y} = 20.2\,\text{MPa}$	57.7
r	$\mathscr{U}(0, 0.6)$	$\mu_r = 0.3$	$s_r = 0.17$	56.7

4.3 Concrete Response in Uniaxial Compressive Cyclic Loading

Based on the preceding assumptions and equations, 1D macroscopic response of concrete in uniaxial compressive cyclic loading is now numerically computed. The general-purpose Finite Element Analysis Program FEAP [36] is used for the finite element solution procedure; Python [27] has been used for the implementation of the equations to generate random vector fields; a Python interface has been developed to both generate the random fields and run the FE analyses in an automatic procedure.

The purpose here is not to present a parametric analysis but to show that with the proposed approach, homogeneous response can be retrieved at macroscale without stochastic homogenization and to show that characteristic features of the concrete uniaxial response in cyclic compressive loading at macroscale can emerge from numerous simpler correlated nonlinear and uncertain mechanisms at E-mesoscale. More details about the potential influence of random field properties on the stochastic finite element method, albeit not exactly in the same context as the work presented here, can be found for instance in [4].

Also, in the absence of detailed information about the correlations at E-mesoscale, the potential problem of incompatible correlation matrix and marginal CDFs presented in Sect. 2.3.3 has not been treated in these numerical applications.

4.3.1 Modeling Concrete Representative Elementary Domain

We first investigate whether a representative response of the concrete elementary domain (ED) $\mathscr{R}$ can be retrieved at macroscale by the proposed modeling. It is reminded that only one type of boundary conditions is considered in this work, namely homogeneous displacements. Consequently, the results shown hereafter could be different for other boundary conditions and the terms "representative response" have to be interpreted accordingly.

Five different combinations of parameter $\overline{L^0}$ and correlation length $\overline{\ell_c}$ are considered (see Table 2). 500 realizations of meso-structures are generated for each of these 5 cases. Figure 6 shows samples of such meso-structures in cases #1 and #3. The 500 corresponding 1D macroscopic responses in uniaxial monotonic compressive loading are computed for each of the 5 cases. Figure 7 presents the mean and standard deviation of these macroscopic responses (Σ-E law) throughout loading evolution.

Table 2 500 meso-structures are generated for 5 parameterizations. The mean μ, standard deviation s and coefficient of variation COV of the macroscopic responses of the concrete ED $\mathscr{R}$ are computed at the end of the monotonic compressive loading ($E = -3.5e^{-3}$)

Case #	M	$\overline{L^0}$	$\overline{\ell_c}$	$\overline{\ell_{cmin}}$	$\mu_{3.5}$ (MPa)	$s_{3.5}$ (MPa)	$COV_{3.5}$ (%)
1	32	**1**	0.4	0.09	−36.6	0.89	2.4
2	32	**1**	0.2	0.09	−38.3	0.44	1.2
3	**32**	**1**	**0.1**	**0.09**	**−39.1**	**0.22**	**0.6**
4	64	2	0.2	0.18	**−39.1**	0.61	1.6
5	128	4	0.4	0.37	**−39.0**	0.22	3.3

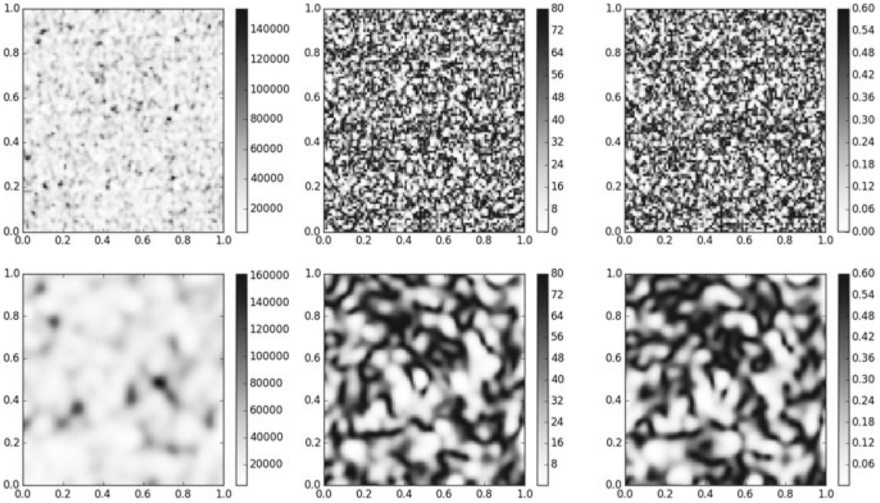
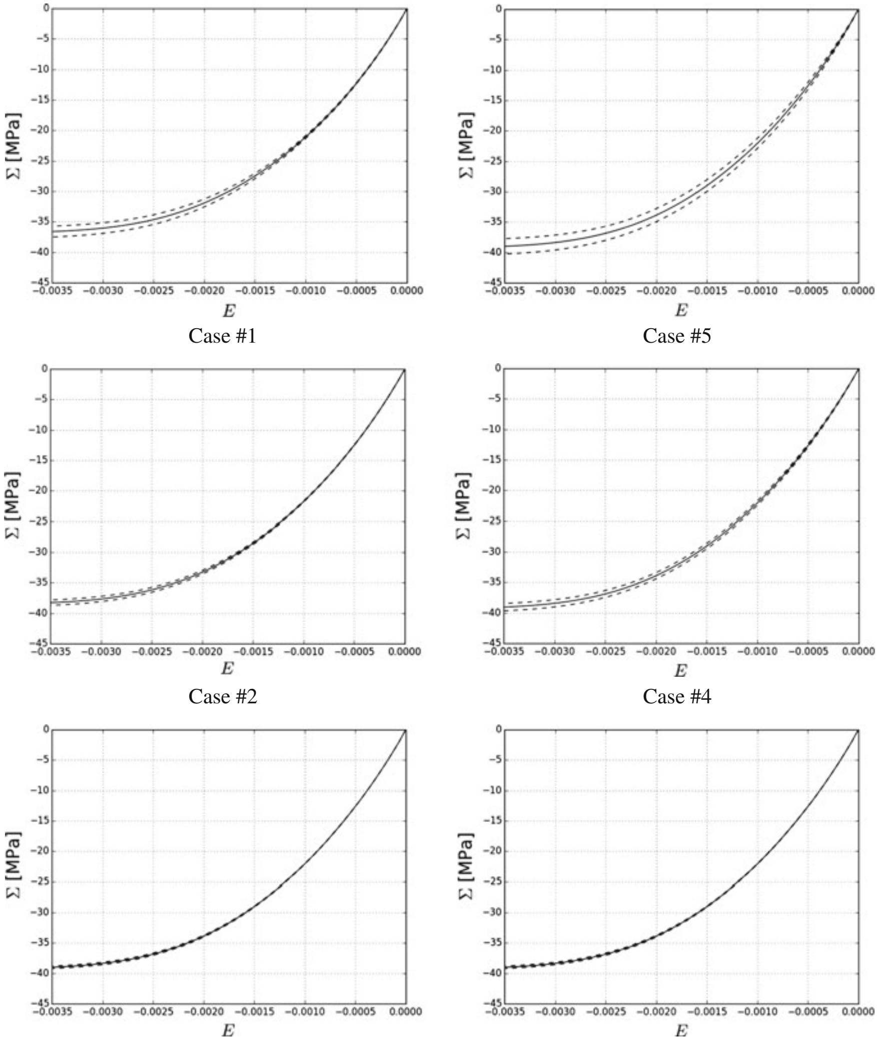


Fig. 6 Samples of heterogeneous meso-structures generated over a normalized area $\overline{\mathscr{R}} = \{(\overline{x}_1, \overline{x}_2) \in [0, 1]^2\}$ meshed into $M_f \times M_f = 96 \times 96$ squares and with $M = 32$. (*top*) $\overline{\ell_c} = 0.1$ (case #3); (*bottom*) $\overline{\ell_c} = 0.4$ (case #1); (*left*) Elastic modulus C [MPa]; (*middle*) Yield stress σ_y [MPa]; (*right*) Damage-plasticity coupling ratio r [-]

The mean $\mu_{3.5}$, standard deviation $s_{3.5}$ and coefficient of variation $COV = s/|\mu|$ are computed at the end of the loading as the imposed displacement reaches $E = -3.5e^{-3}$. These values are reported in Table 2. Some noteworthy conclusions can be drawn from these results:

- As correlation length $\overline{\ell_c}$ decreases, so does the variability ($COV_{3.5}$) of the macroscopic response.
- Case #3 shows that it is possible to find a set of parameters that satisfies $\overline{\ell_c} \geq \overline{\ell_{cmin}}$ and for which the variability of the macroscopic material response is very small ($COV_{3.5} = 0.6\,\%$). This means that any E-mesoscale generated in case #3 yields almost the same material response at macroscale, which can be qualified as a representative response for the boundary conditions considered.



Case #1

Case #5

Case #2

Case #4

Case #3

Case #3

Fig. 7 Mean (*solid line*) along with mean plus or minus standard deviation (*dashed lines*) of the 500 1D macroscopic responses of the ED $\mathscr{R}$ in uniaxial monotonic compression for the 5 meso-structures considered (cases #1 to #5)

- There is a strong reduction of the variability that drops from $COV \geq 50\,\%$ for the material parameters at E-mesoscale to $COV_{3.5} \leq 3.3\,\%$ for the macroscopic material response at maximum compression.
- As $\overline{L^0}$ is kept constant and equal to 1 while $\overline{\ell_c}$ decreases (scenario #1), that are cases #1, #2 and #3 (left column in Fig. 7), mean response changes. On the contrary, as

the $\overline{\ell_c}/\overline{L^0}$ ratio is kept constant while $\overline{\ell_c}$ decreases (scenario #2), that are cases #5, #4 and #3 (right column in Fig. 7), mean response remains almost unchanged.

- Also, for scenario #1, variability ($COV_{3.5}$) is less than for scenario #2 for a same correlation length.

Considering same correlation length in scenarios #1 and #2, there are still two major differences between both scenarios. Firstly, the discretization of the power spectral density function (Eq. 95) is not the same because $\Delta\overline{\kappa} = \overline{\kappa}_u/N$ depends on M (see Eq. 101). Secondly, meso-structures have (asymptotically) ergodic properties in mean and correlation for scenario #1 ($\overline{L^0} = 1$), while this is no more the case in scenario #2 ($\overline{L^0} \geq 1$).

4.3.2 Emergence of a Macroscopic Response

1D macroscopic compressive cyclic response of a concrete elementary area generated with parameters $M = 32$ and $\overline{\ell_c} = 0.1$ (case #3) is shown in Fig. 8. Two different distributions for elastic modulus C are considered: (i) log-normal distribution as introduced in Table 1 and (ii) uniform distribution $C \sim \mathscr{U}(10e^3, 50e^3)$. Because concrete macroscopic response is more realistic for the log-normal distribution (compare with Fig. 1), this distribution was adopted for the numerical applications previously shown in Sect. 4.3.1.

Figure 8 shows that salient features of the experimentally observed concrete behavior (Fig. 1) are represented by the multi-scale stochastic approach presented in this chapter. An initial elastic phase ($E \leq 0.2e^{-3}$) is followed by nonlinear strain hardening; stiffness degradation is observed when unloading (damage); residual deformation remains after complete unloading (plasticity). Besides, in unloading-reloading cycles, hysteresic behavior is produced. It is interesting to observe that nonlinear hardening along with hysteresis in unloading-reloading cycles at macroscale are not explicitly modeled at E-mesoscale (see Fig. 4): they emerge from numer-

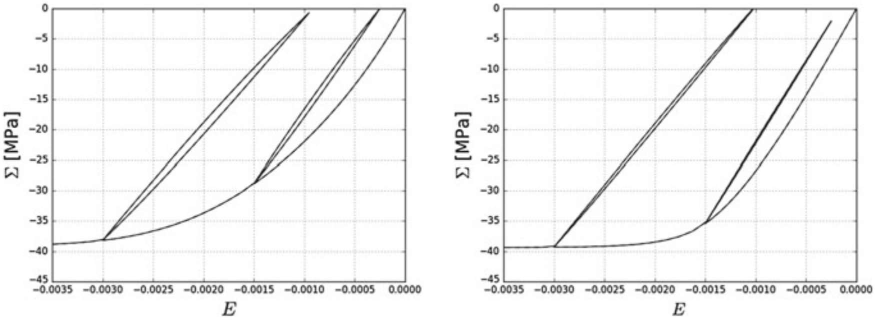


Fig. 8 1D macroscopic compressive cyclic response of a concrete elementary area for two different distributions for elastic modulus C: (*left*) log-normal distribution and (*right*) uniform distribution

ous nonlinear and uncertain responses at E-mesoscale consequently to both spatial variability and averaging of the responses at E-mesoscale over $\mathscr{R}$.

One interesting feature in the macroscopic 1D response of concrete in uniaxial compressive cyclic loading is the hysteresis observed in unloading-reloading cycles: while reloading, the $\varSigma$-E curve follows another path than while unloading, which generates energy dissipation at the structural level. As it is a source of damping in reinforced concrete structures in seismic loading, which modeling is a challenging issue, modeling this hysteresis has been the focus of research work (see e.g. [17, 28]). In [16], a simplified version of the stochastic multi-scale material model presented in this chapter has been developed with only the yield stress σ_y being heterogenous and without damage-plasticity coupling. The material model has been implemented in a beam element and the capacity of the concrete behavior law to generate structural damping has been shown in the numerical testing of reinforced concrete columns in free vibration. Besides, this has shown that the proposed material model can be used in solving numerical nonlinear dynamic analyses of structural frame elements.

5 Conclusion

A stochastic multi-scale approach has been presented in this chapter for numerical modeling of complex materials, that are materials for which macroscopic response results from the interaction of numerous intertwined nonlinear and uncertain mechanisms at lower scales. This approach is based on the construction of an equivalent mesoscale (E-mesoscale) where material properties are heterogenous and where local behavior is nonlinear, coupling mechanisms such as plasticity and damage. Macroscopic response is then computed using averaging formula over an elementary domain (ED). The approach is used to model the uni-dimensional response of concrete material in uniaxial compressive cyclic loading. It is shown that a random E-mesoscale can be generated by spectral representation in such a way that the macroscopic response does not depend on the realization of the random meso-structure. The ED, equipped with such an E-mesoscale, can then be considered as a representative material domain because homogeneous macroscopic properties are retrieved. Besides, this also means that this homogeneous macroscopic behavior is obtained without stochastic homogenization. Because only homogeneous displacements are considered for the boundary conditions for the ED, note that the term "representative" does not imply here independence of the boundary conditions. Moreover, the macroscopic concrete response modeled by this approach exhibits most of the salient features observed in experimental uniaxial cyclic compressive tests on concrete specimens, and particularly the hysteresis loops observed in unloading-reloading cycles. Considering that some of these features are not explicitly represented at the E-mesoscale, this shows the capacity of the approach for letting macroscopic behaviors emerge from simpler mechanisms at lower scales.

In this chapter, the E-mesoscale for concrete material is built on a conjectural basis. Nevertheless, the assumptions that are made both about the mechanical behavior at this scale and the description of the heterogeneity in the properties yield a macroscopic response that reproduces salient features that can be observed experimentally testing concrete specimen. Consequently, although the proposed approach needs to be fed by experimental evidence, it certainly can also trigger experimental research because it provides a rational explanation of macroscopic mechanisms from lower-scale information.

Acknowledgments The author thanks Prof. George Deodatis for fruitful discussions about the content of this book chapter and for hosting him during its preparation. This research is supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (proposal No. 275928).

Appendix 1: On the Ergodicity in Correlation of the Random Fields Simulated with Eq. (41)

From Eq. (41), $j \in [1, \ldots, m]$:

$$g_j(\mathbf{x};\theta) = 2\sqrt{\Delta\kappa_1 \Delta\kappa_2} \sum_{l=1}^{m} \sum_{\alpha=1}^{2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} |H_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2})| \times \cos(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2} \cdot \mathbf{x} - \varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) + \Phi^{\alpha}_{l,n_1 n_2}(\theta)) \tag{105}$$

On the one hand, because the random phases $\Phi(\theta)$ are independent and uniformly distributed over $[0\,,\,2\pi]$, the ensemble correlation function of two sample functions $g_j(\boldsymbol{\xi};\theta)$ and $g_k(\boldsymbol{\xi};\theta)$ reads:

$$\begin{aligned} R_{jk}(\boldsymbol{\xi}) &:= \mathbb{E}\left[g_j(\mathbf{x};\theta)\; g_k(\mathbf{x}+\boldsymbol{\xi};\theta)\right] \\ &= \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} g_j(\mathbf{x};\theta)\; g_k(\mathbf{x}+\boldsymbol{\xi};\theta)\; d\Phi^a\, d\Phi^b \end{aligned} \tag{106}$$

On the other hand, the spatial correlation of the two sample functions $g_j(\boldsymbol{\xi};\theta)$ and $g_k(\boldsymbol{\xi};\theta)$ over a 2D area of size $L_1^0 \times L_2^0$ reads:

$$\begin{aligned} \tilde{R}_{jk}(\boldsymbol{\xi}) &:= \langle g_j(\mathbf{x};\theta)\; g_k(\mathbf{x}+\boldsymbol{\xi};\theta)\rangle_{L_1^0 \times L_2^0} \\ &:= \frac{1}{L_1^0\, L_2^0} \int_0^{L_1^0} \int_0^{L_2^0} g_j(\mathbf{x};\theta)\; g_k(\mathbf{x}+\boldsymbol{\xi};\theta)\; dx_1\, dx_2 \end{aligned} \tag{107}$$

Then, we have from Eq. (105):

$$
\begin{aligned}
g_j(\mathbf{x};\theta)\, g_k(\mathbf{x}+\boldsymbol{\xi};\theta) = 4\Delta\kappa_1\Delta\kappa_2 \\
\times \sum_{l^a=1}^{m}\sum_{\alpha^a=1}^{2}\sum_{n_1^a=0}^{N_1-1}\sum_{n_2^a=0}^{N_2-1}\sum_{l^b=1}^{m}\sum_{\alpha^b=1}^{2}\sum_{n_1^b=0}^{N_1-1}\sum_{n_2^b=0}^{N_2-1} |H_{jl^a}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a})| \\
\times |H_{kl^b}(\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b})|\, A_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}; \theta)
\end{aligned}
\tag{108}
$$

where we introduced

$$
\begin{aligned}
A_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}; \theta) = \cos(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}\cdot\mathbf{x} - \varphi_{jl^a}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}) + \mathit{\Phi}^{\alpha^a}_{l^a, n_1^a n_2^a}(\theta)) \\
\times \cos(\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}\cdot(\mathbf{x}+\boldsymbol{\xi}) - \varphi_{kl^b}(\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}) + \mathit{\Phi}^{\alpha^b}_{l^b, n_1^b n_2^b}(\theta))
\end{aligned}
\tag{109}
$$

Using now the relation $\cos\beta\,\cos\gamma = \frac{1}{2}\,\{\cos(\beta+\gamma)+\cos(\beta-\gamma)\}$, it comes:

$$
\begin{aligned}
A_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}; \theta) = \frac{1}{2}\Big\{ \cos\Big((\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b} + \boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a})\cdot\mathbf{x} + \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}\cdot\boldsymbol{\xi} - \varphi_{jl^b}(\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}) \\
- \varphi_{kl^a}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}) + \mathit{\Phi}^{\alpha^b}_{l^b, n_1^b n_2^b}(\theta) + \mathit{\Phi}^{\alpha^a}_{l^a, n_1^a n_2^a}(\theta)\Big) + \cos\Big((\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b} - \boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a})\cdot\mathbf{x} \\
+ \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}\cdot\boldsymbol{\xi} - \varphi_{jl^b}(\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}) + \varphi_{kl^a}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}) + \mathit{\Phi}^{\alpha^b}_{l^b, n_1^b n_2^b}(\theta) - \mathit{\Phi}^{\alpha^a}_{l^a, n_1^a n_2^a}(\theta)\Big)\Big\}
\end{aligned}
\tag{110}
$$

To calculate the ensemble correlations $R_{jk}(\boldsymbol{\xi})$, we have to calculate:

$$
B_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}) = \int_0^{2\pi}\int_0^{2\pi} A_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}; \theta)\, d\mathit{\Phi}^a\, d\mathit{\Phi}^b
\tag{111}
$$

Because functions $A_{jkl^a l^b}$ are periodic of period 2π, functions $B_{jkl^a l^b} = 0$ except in the case where $\mathit{\Phi}^{\alpha^b}_{l^b, n_1^b n_2^b}(\theta) = \mathit{\Phi}^{\alpha^a}_{l^a, n_1^a n_2^a}(\theta)$, that is as $n_1^a = n_1^b = n_1$ and $n_2^a = n_2^b = n_2$ and $\alpha^a = \alpha^b = \alpha$ and $l^a = l^b = l$. This yields:

$$
B_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}) = 2\pi^2 \cos\left(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}\cdot\boldsymbol{\xi} - \varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) + \varphi_{kl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2})\right)
\tag{112}
$$

and, finally:

$$
R_{jk}(\boldsymbol{\xi}) = 2\Delta\kappa_1\Delta\kappa_2 \sum_{l=1}^{m}\sum_{\alpha=1}^{2}\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} \cos\left(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}\cdot\boldsymbol{\xi} - \varphi_{jl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) + \varphi_{kl}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2})\right)
\tag{113}
$$

Then, to calculate the spatial correlations $\tilde{R}_{jk}(\boldsymbol{\xi})$, we have to calculate:

$$
\tilde{B}_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \boldsymbol{\xi}; \theta) = \int_0^{L_1^0}\int_0^{L_2^0} A_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \mathbf{x}, \boldsymbol{\xi}; \theta)\, dx_1\, dx_2
\tag{114}
$$

Because functions $A_{jkl^a l^b}$ are periodic of period $L_1^0 \times L_2^0$, and with the condition that $H_{jk} = 0$ as any $\kappa_i = 0$, for any $(j, k) \in [1, \ldots, m]^2$ and for any $i \in [1, \ldots, d]$, functions $\tilde{B}_{jkl^a l^b}$ are equal to zero, except if $\boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b} = \boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}$, that is $n_1^a = n_1^b = n_1$ and $n_2^a = n_2^b = n_2$ and $\alpha^a = \alpha^b = \alpha$, in which case:

$$\tilde{B}_{jkl^a l^b}(\boldsymbol{\kappa}^{\alpha^a}_{n_1^a n_2^a}, \boldsymbol{\kappa}^{\alpha^b}_{n_1^b n_2^b}; \boldsymbol{\xi}; \theta) = \frac{L_1^0\, L_2^0}{2} \cos\big(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2} \cdot \boldsymbol{\xi} - \varphi_{jl^b}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) + \varphi_{kl^a}(\boldsymbol{\kappa}^{\alpha}_{n_1 n_2}) + \Phi^{\alpha}_{l^b, n_1 n_2}(\theta) - \Phi^{\alpha}_{l^a, n_1 n_2}(\theta)\big) \tag{115}$$

With this expression of the functions $\tilde{B}_{jkl^a l^b}$, we do not have $\tilde{R}(\boldsymbol{\xi}) = R(\boldsymbol{\xi})$. However, when wave-number shifts are introduced so that wave numbers $\boldsymbol{\kappa}$ become dependent on the index l (as in [7, 26]), the condition $l^a = l^b = l$ has to be added for $\tilde{B}_{jkl^a l^b}$ not to be equal to zero. Consequently, $\Phi^{\alpha}_{l^b, n_1 n_2}(\theta) = \Phi^{\alpha}_{l^a, n_1 n_2}(\theta)$ in Eq. (115) and we finally recover $R(\boldsymbol{\xi}) = \tilde{R}(\boldsymbol{\xi})$, meaning that sample fields $g_j(\boldsymbol{\xi}; \theta)$ are ergodic in correlation.

Appendix 2: Material Model at Mesoscale for the Numerical Applications

For the one-dimensional material model at mesoscale used in the numerical applications shown in Sect. 4, we use $h(\sigma) = |\sigma|$ in the definition of the criterium function (see Eq. 11 in Sect. 2.2):

$$\phi_{n+1} = |\sigma_{n+1}| - \sigma_y \quad \Rightarrow \quad \nu_{n+1} := \frac{\partial \phi_{n+1}}{\partial \sigma_{n+1}} = sign(\sigma_{n+1}) \tag{116}$$

Then, Eq. (61) can be written as:

$$\sigma_{n+1} = \sigma^{trial}_{n+1} - D_n^{-1} \gamma_{n+1} sign(\sigma_{n+1}) \tag{117}$$

Multiplying both sides of Eq. (117) by $sign(\sigma_{n+1})$, it comes:

$$|\sigma_{n+1}| = \sigma^{trial}_{n+1} sign(\sigma_{n+1}) - D_n^{-1} \gamma_{n+1} \tag{118}$$

Multiplying now both sides of Eq. (118) by $sign(\sigma^{trial}_{n+1})$, it comes:

$$\left(|\sigma_{n+1}| + D_n^{-1} \gamma_{n+1}\right) sign(\sigma^{trial}_{n+1}) = |\sigma^{trial}_{n+1}| sign(\sigma_{n+1}) \tag{119}$$

Setting $\gamma_0 = 0$ and $D_0 > 0$, $\left(|\sigma_{n+1}| + D_n^{-1} \gamma_{n+1}\right)$ necessarily is non-negative because $\dot{\gamma} \geq 0$ and $\dot{D}|\sigma| = r\,\dot{\gamma} \geq 0$. Consequently:

$$sign(\sigma_{n+1}) = sign(\sigma^{trial}_{n+1}) \tag{120}$$

Then, we have from Eqs. (117) and (116):

$$|\sigma_{n+1}| = |\sigma_{n+1}^{trial}| - D_n^{-1}\gamma_{n+1} \tag{121}$$

$$\phi_{n+1} = \phi_{n+1}^{trial} - D_n^{-1}\gamma_{n+1} \tag{122}$$

with $\phi_{n+1}^{trial} = |\sigma_{n+1}^{trial}| - \sigma_y$, from which we can calculate γ_{n+1} in case of inelastic evolution:

$$\phi_{n+1} = 0 \quad \Rightarrow \quad \gamma_{n+1} = D_n\, \phi_{n+1}^{trial} \tag{123}$$

Appendix 3: Translation from Gaussian to Uniform Distributions

Let $\mathfrak{a}_1$ and $\mathfrak{a}_2$ be two independent normal Gaussian variables: $(\mathfrak{a}_1, \mathfrak{a}_2) \sim \mathcal{N}(0, 1)$. Then $\mathfrak{b} = \exp(-(\mathfrak{a}_1^2 + \mathfrak{a}_2^2)/2)$ is a random variable with uniform distribution in $[0,\ 1]$: $\mathfrak{b} \sim \mathcal{U}(0, 1)$. Indeed:

$$\Pr[\mathfrak{b} \leq b] = \frac{1}{2\pi} \int_{\{(a_1,a_2)|e^{-\frac{1}{2}(a_1^2+a_2^2)} \leq b\}} e^{-\frac{1}{2}(a_1^2+a_2^2)} da_1 da_2 \tag{124}$$

Then:

- if $b > 1$, $\Pr[\mathfrak{b} \leq b] = 1$ because $e^{-\frac{1}{2}(a_1^2+a_2^2)} \leq 1, \forall (a_1, a_2) \in \mathbb{R}^2$;
- if $b < 0$, $\Pr[\mathfrak{b} \leq b] = 0$ because $e^{-\frac{1}{2}(a_1^2+a_2^2)} > 1, \forall (a_1, a_2) \in \mathbb{R}^2$;
- and, if $0 \leq b \leq 1$, we can rewrite relation (124) with polar coordinates as:

$$\Pr[\mathfrak{b} \leq b] = \frac{1}{2\pi} \int_0^{2\pi} \int_{\sqrt{-2\ln b}}^{+\infty} e^{-\frac{1}{2}r^2}\, r\, dr d\theta = \left[-e^{-\frac{1}{2}r^2}\right]_{\sqrt{-2\ln b}}^{+\infty} = b$$

References

1. Ben-Dhia H (1998) Multiscale mechanical problems: the Arlequin method. C R Acad Sci Paris, Série IIb, 326:899–904
2. Benkemoun N, Hautefeuille M, Colliat J-B, Ibrahimbegovic A (2010) Failure of heterogeneous materials: 3D meso-scale FE models with embedded discontinuities. Int J Num Meth Eng, doi:10.1002/nme.2816
3. Bocchini P, Deodatis G (2008) Critical review and latest developments of a class of simulation algorithms for strongly non-Gaussian random fields. Probab Eng Mech, doi:10.1016/j.probengmech.2007.09.001
4. Charmpis DC, Schuëller GI, Pellissetti MF (2007) The need for linking micromechanics of materials with stochastic finite elements: A challenge for materials science. Comput Mat Sci, doi:10.1016/j.commatsci.2007.02.014
5. Cottereau R (2013) Numerical strategy for unbiased homogenization of random materials. Int J Num Meth Eng, doi:10.1002/nme.4502

6. Deodatis G (1996) Non-stationary stochastic vector processes: seismic ground motion applications. J Eng Mech 11:149–168
7. Deodatis G (1996) Simulation of ergodic multivariate stochastic processes. ASCE J Eng Mech 122(8):778–787
8. Feyel F (2003) A multilevel finite element method (FE^2) to describe the response of highly non-linear structures using generalized continua. Comput Meth Appl Mech Eng, doi:10.1016/S0045-7825(03)00348-7
9. Germain P, Nguyen QS, Suquet P (1983) Continuum thermodynamics. ASME J Appl Mech 50:1010–1020
10. Gioffre M, Gusella V, Grigoriu M (2000) Simulation of non-Gaussian field applied to wind pressure fluctuations. Probab Eng Mech 15:339–45
11. Grigoriu M (1995) Applied non-Gaussian processes: Examples, theory, simulation, linear random vibration and MATLAB solutions. Prentice-Hall, Englewood Cliffs
12. Hill R (1950) The mathematical theory of plasticity. Oxford University Press, London
13. Ibrahimbegovic A (2009) Nonlinear solid mechanics: Theoretical formulations and finite element solution methods. Springer Netherlands
14. Ibrahimbegovic A, Jehel P (2008) Coupled damage-plasticity constitutive model and direct stress interpolation. Comput Mech, doi:10.1007/s00466-007-0230-6
15. Ibrahimbegovic A, Markovic D (2003) Strong coupling methods in multi-phase and multi-scale modeling of inelastic behavior of heterogeneous structures. Comput Meth Appl Mech Eng, doi:10.1016/S0045-7825(03)00342-6
16. Jehel P, Cottereau R (2015) On damping created by heterogeneous yielding in the numerical analysis of nonlinear reinforced concrete frame elements. Comput Struct, doi:10.1016/j.compstruc.2015.03.001
17. Jehel P, Davenne L, Ibrahimbegovic A, Léger P (2010) Towards robust viscoelastic-plastic-damage material model with different hardenings / softenings capable of representing salient phenomena in seismic loading applications. Comp Concr 7(4):365–386
18. Ladevèze P, Loiseau O, Dureisseix D (2001) A micro-macro and parallel computational strategy for highly heterogeneous structures. Int J Num Meth Eng, doi:10.1002/nme.274
19. Lubliner J (1984) A maximum-dissipation principle in generalized plasticity. Acta Mech 52:225–237
20. Markovic D, Ibrahimbegovic A (2006) Complementary energy based FE modelling of coupled elasto-plastic and damage behavior for continuum microstructure computations. Comput Meth Appl Mech Eng, doi:10.1016/j.cma.2005.05.058
21. Maugin G (1999) The thermodynamics of nonlinear irreversible behaviors: An introduction. World Scientific, Singapore
22. Miehe C, Koch A (2002) Computational micro-to-macro transitions of discretized microstructures undergoing small strains. Arch Appl Mech, doi:10.1007/s00419-002-0212-2
23. Nemat-Nasser S, Hori M (1993) Micromechanics: Overall properties of heterogenous materials. Elsevier Science Publishers B.V., Amsterdam, The Netherlands
24. Ostoja-Starzewski M (2006) Material spatial randomness: From statistical to representative volume element. Probab Eng Mech, doi:10.1016/j.probengmech.2005.07.007
25. Papadopoulos V, Soimiris G, Papadrakakis M (2013) Buckling analysis of I-section portal frames with stochastic imperfections. Eng Struct, doi:10.1016/j.engstruct.2012.09.009
26. Popescu R, Deodatis G, Prevost JH (1998) Simulation of homogeneous nonGaussian stochastic vector fields. Probab Eng Mech 13(1):1–13
27. Python Software Foundation (2015) Python langage documentation, version 3.4.3. http://docs.python.org/3.4 (last consulted Nov. 6, 2015)
28. Ragueneau F, La Borderie C, Mazars J (2000) Damage model for concrete-like materials coupling cracking and friction, contribution towards structural damping: first uniaxial applications. Mech Cohes-Frict Mater 5:607–625
29. Ramtani S (1990) Contribution to the modeling of the multi-axial behavior of damaged concrete with description of the unilateral characteristics. PhD Thesis (in French), Paris 6 University

30. Savvas D, Stefanou G, Papadrakakis M, Deodatis G (2014) Homogenization of random heterogeneous media with inclusions of arbitrary shape modeled by XFEM. Comput Mech, doi:10.1007/s00466-014-1053-x
31. Shields MD, Deodatis G (2013) A simple and efficient methodology to approximate a general non-Gaussian stationary stochastic vector process by a translation process with applications in wind velocity simulation. Probab Eng Mech, doi:10.1016/j.probengmech.2012.10.003
32. Shinozuka M (1987) Stochastic fields and their digital simulation. In: Schuëller et al. (eds) Stochastic Methods in Structural Dynamics, Martinus Nijhoff Publishers, Dordrecht
33. Shinozuka M, Deodatis G (1991) Simulation of of stochastic processes by spectral representation. Appl Mech Rev 44(4):191–204
34. Shinozuka M, Deodatis G (1996) Simulation of multi-dimensional Gaussian stochastic fields by spectral representation. Appl Mech Rev 49(1):29–53
35. Simo JC, Hughes TJR (1998) Computational Inelasticity. Springer, Berlin
36. Taylor R (2005) FEAP: A finite element analysis program, User manual & Programmer manual. University of California Berkeley, California, Version 7.4
37. Torrenti J-M, Pijaudier-Cabot G, Reynouard J-M (2012) Mechanical behavior of concrete. ISTE, London & Wiley, Hoboken, NJ
38. Trigo APM, Liborio JBL (2014) Doping technique in the interfacial transition zone between paste and lateritic aggregate for the production of structural concretes. Mat Res, doi:10.1590/S1516-14392013005000169
39. Zaoui A (2002) Continuum micromechanics: Survey. J Eng Mech 128(8):808–816
40. Zienkiewicz OC, Taylor RL (2000) The finite element method, 5th ed Butterworth-Heinemann, Boston

Relating Structure and Model

Ivica Kožar

Abstract In order to gain additional insight into large structure a model is usually built which leaves us with a problem of transfer of parameters between the model and the structure. Problem is addressed on the general level but after discretization and is formulated as a relationship between relevant parameters of the structure and its model. Scaling matrices in parameter and in measurement space are determined.

Keywords Scaling matrix · Parameter space · Measurement space · Force reconstruction · Dynamics

1 Introduction

Large structures are usually important in which case they could be under monitoring for relevant parameters. Sometimes relevant parameters could not be monitored/measured directly so they are determined using some inverse procedure. In some cases, relationship between parameters is highly non-*linear* or includes some stochastic properties. This is the case of the relationship between strains and loading in wind power plants. In order to gain additional insight into the structure a model is usually built. This approach leaves us with a problem of transfer of parameters between the model and the structure, i.e. with determination of an appropriate scaling of parameters. Practical motivation for this work is given in Fig. 1 where the large structure is 2.5 MW wind turbine produced and owned by Končar—Croatia and the small structure is a model wind turbine in a wind tunnel of the Faculty of Electrical Engineering and Computing University of Zagreb.

Our final goal is to relate measurements on those two different structures.

Problem is addressed on the general level but after discretization and is formulated as a relationship between relevant parameters of a structure and its model. Approach to inverse problem for beam before discretization can be found e.g. in [1]. It is quite

I. Kožar (✉)
Faculty of Civil Engineering, University of Rijeka, Rijeka, Croatia
e-mail: ivica.kozar@gradri.hr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_7

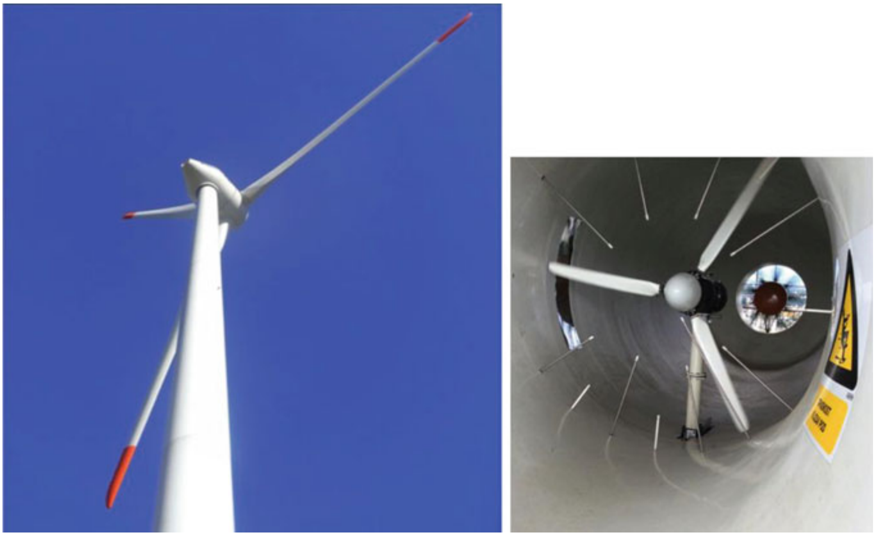


Fig. 1 Large structure and its model: how to relate measurements

a different approach from a review of scaling laws presented in [2]. In the case of wind power plant that is the problem of scaling of forces induced by wind acting on the structure. Not only is the relationship non-linear, it is highly stochastic (both in intensity and direction).

Primary intention of the model is to serve for testing of control procedures for control of blade speed rotation ω, which requires for relation between angles of blade inclination φ, e.g. see [3]. We assume the relationship between force on structure and blade rotational velocity and inclination as $F_1 = W_1(\omega,\varphi)$ and $F_2 = W_2(\omega,\varphi)$ where indexes '1' and '2' are structure and model respectively.

F_1 and F_2 are forces induced by the wind and W_1 and W_2 are unknown functions relating force on the structure and wind strength and direction expressed through blade parameters (rotational velocity and blade inclination). Function W_2 (belonging to model) can be determined for certain wind conditions by measurements in wind tunnel. Function W_1 (belonging to structure) cannot be determined due to uncontrolled wind conditions in reality. This paper suggests how to establish their relationship so that W_1 can be determined from W_2.

After discretization in structural equations forces F are represented as vectors $\mathbf{F}$ and we can establish the relationship between structural forces as $\mathbf{F}_1 = \mathbf{S}\mathbf{F}_2$ where $\mathbf{S}$ is the novel 'scaling matrix'. $\mathbf{F}_1$ and $\mathbf{F}_2$ could be determined from measured strains by applying inverse procedure $\mathbf{F}_1 = \mathbf{H}_1\mathbf{d}_1$ and $\mathbf{F}_2 = \mathbf{H}_2\mathbf{d}_2$ where $\mathbf{H}$ is a generalized inverse of a measurement mapping matrix and $\mathbf{d}$ is some measured value, e.g. strain, displacement, velocity, acceleration etc. This inverse procedure could be deterministic or stochastic or a combination of both.

Knowing $\mathbf{F}_1$ and $\mathbf{F}_2$, $\mathbf{S}$ could be determined. There is a review of force reconstruction methods in [4]. If $\mathbf{F}$ would be a simple force vector, the problem would be undetermined. However, $\mathbf{F}$ is a function of time $\mathbf{F} = \mathbf{F}(\mathrm{t})$ measured in discrete time intervals and can be represented as a rectangular matrix of dimension [rows = number of forces, columns = number of measurements]. In the case of constant monitoring number of columns should correspond to the size of time window for which we conduct analysis (optimal size of this window could depend on many parameters and is not subject of this work). The above equation for $\mathbf{S}$ could be transformed into quadratic form and solved as an optimization problem. Elements of matrix $\mathbf{S}$ are then coefficients of the quadratic form matrix.

Terminology
We are connecting two linear systems $\mathbf{y}_1 = \mathbf{H}_1\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{H}_2\mathbf{x}_2$ where '**y**' stands for 'measured values', '**x**' are model parameters and '**H**' is a 'measurement matrix' mapping from the parameter space R^n into the measurement space R^m.

2 Scaling

In order to relate (already discretized) structure and its model they have to be of the same scale, i.e. one of them has to be scaled to match the other. Not all parameters of the structure can be related at the same time. Assuming that the structure is described with matrix equation $\mathbf{y} = \mathbf{A}(\mathbf{k})\mathbf{x}$ we could relate parameters $\mathbf{y}$ with $\mathbf{y}_1 = \mathbf{S}_\mathrm{y}\mathbf{y}_2$, parameters $\mathbf{x}$ with $\mathbf{x}_1 = \mathbf{S}_\mathrm{x}\mathbf{x}_2$ or parameters $\mathbf{k}$ with $\mathbf{k}_1 = \mathbf{S}_\mathrm{k}\mathbf{k}_2$ where $\mathbf{S}$ is some scaling matrix and indices '1' and '2' could for e.g. represent structure and model respectively. In this paper $\mathbf{x}$ is related in parameter space, $\mathbf{y}$ is related in measurement space and relating of implicit parameters $\mathbf{k}$ is not dealt with.

The principle and equations are general and applicable to all structures where a relation is to be established. In addition, scalar can be introduced anywhere in the relations so that values are not equal but a multiple of each other.

2.1 Scaling in Parameter Space

In this case we assume $\mathbf{y}_1 = \mathbf{y}_2$ where indices '1' and '2' stand for structure and model respectively. That means that we want the measured values to be the same on some points on the structure and on the model and we need to scale the parameters to achieve that goal (number of points has to be the same [m×1]).

We have:

$$\begin{aligned} \mathbf{x}_2 &= \mathbf{S}_x\mathbf{x}_1; \\ \mathbf{H}_1\mathbf{x}_1 &= \mathbf{H}_2\mathbf{x}_2 \end{aligned} \tag{1}$$

where $\mathbf{S}_x$ is the new ‘scaling matrix’ connecting the two vectors in parameter space R^n.

Generally $\mathbf{A}$ is not square and

$$\mathbf{S}_x = \mathbf{H}^{-g}\mathbf{H}_1 \tag{2}$$

where $\mathbf{H}^{-g}$ is the generalized (Moon-Penrose) inverse and the scaling matrix $\mathbf{S}_x$ is not deterministic but a result of an optimization procedure. In that case $\mathbf{x}_2 = \mathbf{S}_x\mathbf{x}_1$ is valid even if $\mathbf{x}_1$ and $\mathbf{x}_2$ are of different dimension (e.g. $x_1 \rightarrow [p\times 1]$ and $x_2 \rightarrow [q\times 1]$). In that case $\mathbf{S}_x$ is of dimension [q×p] and exists if Rank[$\mathbf{H}_2$] $\Rightarrow$ q. It is a function of only the measurement matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. Scaling in the parameter space R^n is the least squares problem and $\mathbf{S}_x$ corresponds to the model resolution matrix, i.e. describes how well are model parameters resolved from structural parameters (or vice versa, describes how are structural parameters related to model parameters).

2.2 Scaling in Measurement Space

In this type of analysis we would like to relate measured values and assume that some measurement-scaling matrix relates structure and model measurements

$$\mathbf{y}_1 = \mathbf{S}_y\mathbf{y}_2 \tag{3}$$

where $\mathbf{S}_y$ is the measurement space R^m. In the case when $\mathbf{y}$ are single column vectors $\mathbf{S}_y$ is undetermined unless $\mathbf{y}$ all have the same size and are independent, in which case $\mathbf{S}_y$ is diagonal. In practice, we have multiple realizations of measurements so $\mathbf{y}$ have dimension [m×t] where ‘t’ stands for time instances (not necessarily the physical time).

In the case of multiple realizations of measurements, vectors $\mathbf{y}$ become matrices $\mathbf{Y}$ [m×t] and the scaling is

$$\mathbf{Y}_1 = \mathbf{S}_y\mathbf{Y}_2 \quad ; \quad \mathbf{S}_y = \mathbf{Y}_2^{-g}\mathbf{Y}_1 \tag{4}$$

where $\mathbf{Y}^{-g}$ is again the generalized (Moon-Penrose) inverse. Scaling matrix $\mathbf{S}_y$ in the measurement space R^m is a function of only series of measurements $\mathbf{y}$ and corresponds to the data resolution matrix, i.e. describes how well are structural data (structural measurements) resolved from model data (model measurements) or how are model data related to structural data.

2.3 Scaling of Dynamically Loaded Structures

D’Alembert’s equilibrium equation of dynamically loaded structures includes two types of forces: inertial and elastic. We will ignore damping and the scaling equation now involves two scaling matrices $\mathbf{S}_\mathrm{s}$—static scaling matrix and $\mathbf{S}_\mathrm{d}$—dynamic scaling matrix

$$\mathbf{M}_1\ddot{\mathbf{y}}_1(t) + \mathbf{K}_1\mathbf{y}_1(t) = \mathbf{S}_d\mathbf{M}_2\ddot{\mathbf{y}}_2(t) + \mathbf{S}_s\mathbf{K}_2\mathbf{y}_2(t) \tag{5}$$

where $\mathbf{M}$ and $\mathbf{K}$ are mass and stiffness matrices of structures '1' and '2' respectively. First, some relation between measured values is assumed, similar to Eq. (3)

$$\mathbf{y}_1 = \mathbf{I}_S\mathbf{y}_2 \tag{6}$$

and the scaling matrix $\mathbf{I}_\mathrm{S}$ is diagonal, i.e. $\mathbf{y}_1$ and $\mathbf{y}_2$ have the same size and are independent. We can relate static and dynamic part separately, i.e. even if the problem is dynamical, we want the same static scaling matrix $\mathbf{S}_\mathrm{s}$ to relate the static problem as well, so we write

$$\mathbf{K}_1\mathbf{I}_S = \mathbf{S}_s\mathbf{K}_2 \tag{7}$$

and follows $\mathbf{S}_s = \mathbf{K}_1\mathbf{I}_S\mathbf{K}_2^{-1}$. Now, parts belonging to the elastic force are equal and for inertial forces remains the relation

$$\mathbf{M}_1\mathbf{I}_S^2 = \mathbf{S}_d\mathbf{M}_2 \tag{8}$$

because $\ddot{\mathbf{y}}_1 = \mathbf{I}_S^2\ddot{\mathbf{y}}_2$. The dynamic scaling matrix is $\mathbf{S}_d = \mathbf{M}_1\mathbf{I}_S^2\mathbf{M}_2^{-1}$.

Scaling matrices $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ can be determined and are well conditioned if structural mass and stiffness matrices are well conditioned. In addition, mass matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are in many cases lumped, i.e. diagonal in which case $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ are easily calculated. Structures related with $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ scaling matrices have the same eigenvalues and eigenvectors, i.e. resonant frequencies and the same response to dynamic loading.

3 Loading Reconstruction

Force reconstruction belongs to a class of source identification problems; problem is explicitly formulated, i.e. there is no need for special inverse formulation of the problem. Instead we introduce the measurement matrix equation $\mathbf{y} = \mathbf{Hx}$ where we try to determine the parameter vector $\mathbf{x}$ from measurements vector $\mathbf{y}$. In the implicit formulation, some parameter of matrix $\mathbf{H}$ would have to be determined requiring special inverse formulation of the problem. Matrix $\mathbf{H}$ relates parameters (in our case loading) and measurements. The main difficulty is that matrix $\mathbf{H}$ is not square (we usually have more measurements then parameters) and that the measurement vector $\mathbf{y}$ contains some noise so that the real measurement equation is $\mathbf{y} = \mathbf{Hx} + \mathbf{w}$ where $\mathbf{w}$ is some measurement noise. Some of necessary assumptions about $\mathbf{w}$ can be found in e.g. [5].

3.1 Treatment of Measurement Noise

All measurement contains some noise that can be described as a stochastic process with some probability distribution function. Based on the level of our knowledge of that stochastic process it is possible to reduce to certain extent the effect of the measurement noise. It is beneficiary to have more measurement points then parameters that are looked for, in which case we apply least squares (LS) procedure. Here we apply it through generalized inverse matrix what is equivalent to the LS procedure [6]. Additional knowledge about noise distribution allows us to implement weighted least squares (WLS) procedure. Assuming measurement noise in different in each position then weighting matrix is equal to the reciprocal of the measurement residual. Knowing probability distribution function of the measurement noise permits us to apply the maximum likelihood (ML) method to reduce the influence of noise on parameter reconstruction. Finally, knowledge of the pdf of the model parameters allows formulation of the simulation procedure, e.g. Monte Carlo procedure. Adopted approach is based on [6, 7] and LS and WLS approach will be demonstrated in the examples.

Besides measurement noise, some other measuring properties are important for loading (and any other) reconstruction, especially in the case of dynamic loading. So, it is better to measure acceleration then displacement although both are needed in equations. Obtaining acceleration from displacements requires differentiation, which is mostly bad conditioned and introduces additional errors. On the other hand, obtaining displacements from accelerations requires integration. There is a comparison of errors introduced by differentiation and integration of measured data in [8].

3.2 Static Loading

In the case of static loading, we are reconstructing loading forces from displacement or strain measurements. Better results are obtained when matrix $\mathbf{H}$ has more rows then columns since this results with the real least square problem. That means that we need more measurement points then recovering parameters what is in most cases easily accomplished. Measurement matrix $\mathbf{H}$ is composed from parts of the structure flexibility matrix when the measuring values are displacements. For strain measurement, we have to introduce constants that describe the material and cross section properties.

The measurement equation $\mathbf{y} = \mathbf{Hx}$ is solved using the generalized (Moon-Penrose) inverse which is equivalent to the LS method. In the case of the full column-rank $\mathbf{H}$ we could use $\mathbf{H}^{-g} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$. Better elimination of noise can be obtained with WLS where weighing matrix $\mathbf{W}$ is obtained from the measurement residuals. The measurement residual matrix $\mathbf{R}$ is constructed from the estimated measurement mean and variance

$$\mu_{meas.} = \frac{1}{N_{meas.}} \sum_{meas.} y_{meas.}$$

$$\sigma^2_{meas.} = \frac{1}{N_{meas.}} \sum_{meas.} (y_{meas.} - \mu_{meas.})^2 \tag{9}$$

and

$$\mathbf{R} = diag\left[\sigma^2_{meas.}\right]$$

$$\mathbf{W} = \mathbf{R}^{-1} \tag{10}$$

The inverse equation is now

$$\mathbf{P} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}$$

$$\mathbf{y}_W = \mathbf{H}^T\mathbf{W}\mathbf{y}$$

$$\mathbf{x} = \mathbf{P}\mathbf{y}_W \tag{11}$$

and **P** contains a-posterior variance of reconstructed parameters

$$\sigma_i = \sqrt{\mathbf{P}_{i,i}} \tag{12}$$

$\boldsymbol{\sigma}$ give good error bound for reconstructed parameters. Off-diagonal elements of **P** contain parameter cross-correlation coefficients.

3.3 Dynamic Loading

Dynamically loaded structures are much more sensitive to the measurement noise because forces change in time and general mean cannot be established. Consequently different noise-canceling procedures have to be applied, e.g. Kalman filter [7]. Another approach is to use modal analysis like in [5]. The benefit is that the modal shapes are assumed to be known from the structure model and can be used to reduce the measurement noise. In addition, analysis time is reduced (if the reduced modal space is used which is usual). On the down side, the result has to be transformed from the modal space back into the global structure space, which is an additional source of error. Dynamic structure equation in modal space has the well-known form

$$\mathbf{M}_{\mathrm{mod}}\ddot{\mathbf{y}}_{\mathrm{mod}}(t) + \mathbf{K}_{\mathrm{mod}}\mathbf{y}_{\mathrm{mod}}(t) = \mathbf{F}_{\mathrm{mod}}(t) \tag{13}$$

with $\mathbf{M}_{\mathrm{mod}} = \boldsymbol{\Phi}^T\mathbf{M}\boldsymbol{\Phi}$ and $\mathbf{K}_{\mathrm{mod}} = \boldsymbol{\Phi}^T\mathbf{K}\boldsymbol{\Phi}$ and $\mathbf{F}_{\mathrm{mod}} = \boldsymbol{\Phi}^T\mathbf{F}$ with $\boldsymbol{\Phi}$ being the eigenmatrix. In forward simulation, Eq. (13) is solved using modified Newmark method from [9]. In inverse formulation, we first recover modal components (displacement, velocities or accelerations)

$$\mathbf{y}_{\mathrm{mod}} = \mathbf{H}_{\mathrm{mod}}^{-g}\mathbf{y} \tag{14}$$

where $\mathbf{H}_{\mathrm{mod}}^{-g}$ is the generalized inverse of the modal measurement matrix

$$\mathbf{H}_{\mathrm{mod}} = \mathbf{H}_0\mathbf{\Phi}_r \tag{15}$$

Eigenmatrix $\mathbf{\Phi}_r$ does not have to cover the whole modal space; only 'r' components could be used. Measurement matrix $\mathbf{H}_0$ has to be adapted to the measured values: displacement, velocities or accelerations or any combination.

After modal components have been recovered from measured values using Eq. (14), additional, not measured, modal components are determined using numerical derivatives or numerical integrals or both (in the case when only velocities are measured). With all the modal components available, modal loading is recovered using Eq. (13). Modal loading is transferred into global loading using

$$\mathbf{F} = \mathbf{\Phi}_{\mathrm{mod}}^{-g}\mathbf{F}_{\mathrm{mod}} \tag{16}$$

Number of modes in the eigenmatrix $\mathbf{\Phi}_r$ has to be such that the matrix is regular, i.e. 'r' has to be equal or greater then the number of measuring points; in the contrary, special recovery procedures have to be used.

4 Examples

In all the examples, measured data has been synthetically generated using random number generators in Wolfram Mathematica and MathCad.

Two completely different structures, a truss and a beam, are scaled in parameter and in measurement space. Structural properties are:
Truss: 33 nodes, 72 bars, EA = 1000., L = 10.0, h = 2.0
Beam: 21 node, 20 beams, EI = 1000.0, L = 1.0
Truss, its supports and loaded nodes are presented in Fig. 2

Structures have been scaled in spaces of the same size, i.e. structural matrices have been statically condensed $\mathbf{K}_{\mathrm{cond}} = \mathbf{K}_{\mathrm{pp}} - \mathbf{K}_{\mathrm{ps}}\mathbf{K}_{\mathrm{ss}}^{-1}\mathbf{K}_{\mathrm{sp}}$ where $\mathbf{K}_{\mathrm{ss}}$ is the matrix part to be condensed and $\mathbf{K}_{\mathrm{pp}}$ is the part to be kept. Two examples were calculated, one with 5 and the other with 7 nodes kept; smaller number of nodes kept gives better results. Besides the nodes kept, in the full (non-condensed) structure, the other nodes scale very well, too. Condensed nodes, i.e. points where the scaling is performed are presented in Fig. 3.

Scaling with 5 nodes kept is not presented (comparison is even better). Index '1' stands for beam and index '2' for truss.

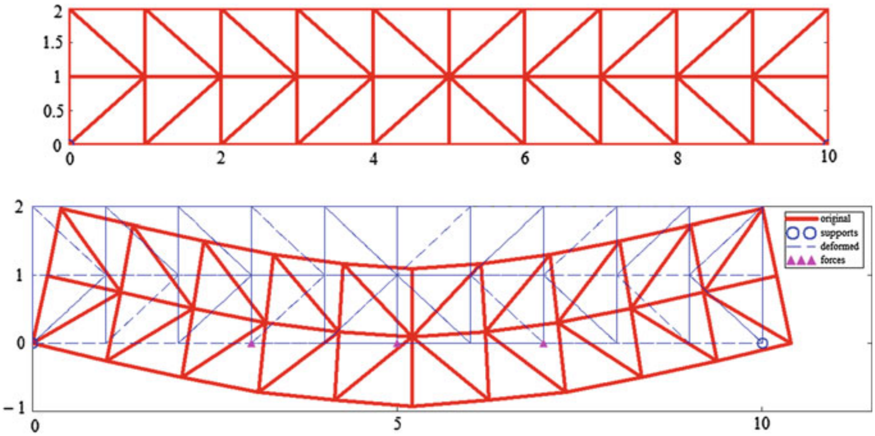


Fig. 2 Truss: supports and loaded nodes

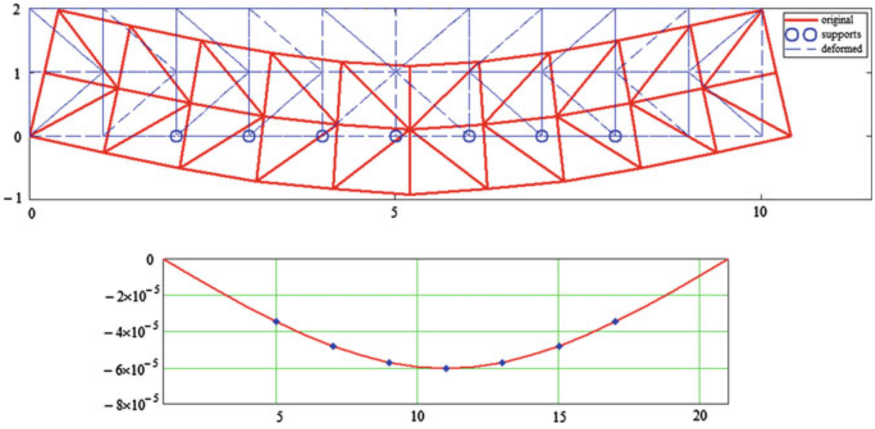


Fig. 3 Nodes for comparison of truss and beam

4.1 Scaling in Parameter Space

In this example, we are scaling truss forces so that displacements in the beam and in the truss have the same value at selected nodes $\mathbf{x}_2 = \mathbf{S}_x \mathbf{x}_1$. Scaling matrix is given as $\mathbf{S}_x = \left(\mathbf{A}_2^T \mathbf{A}_2\right)^{-1} \mathbf{A}_2^T \mathbf{A}_1$ where $\mathbf{A}_1$ and $\mathbf{A}_2$ are condensed stiffness matrices of beam and truss, respectively.

The beam is loaded only in the middle with a unit force and the truss is loaded with unit forces in three points visible in Fig. 3. Comparison of the results is in Fig. 4 and scaling is excellent.

This example demonstrates how can we scale the loading of two completely different structures (i.e. large structure and a small model) and obtain the same displace-

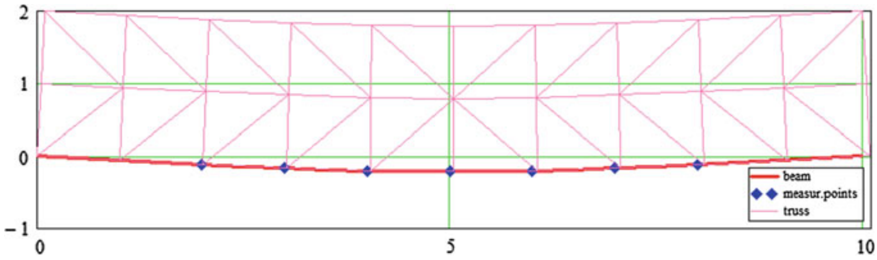


Fig. 4 Comparison of displacements of truss and beam with scaled loading

ments. Notice that the scaling matrix $\mathbf{S}_{\mathrm{x}}$ is a result of the least squares optimization procedure and not a deterministic function. This means that it has to be determined for every structure pair from the beginning, using the presented procedure. Also, matrix $\mathbf{S}_{\mathrm{x}}$ is dense so all measurement points on the truss are loaded, which is opposite to the beam where only central point is loaded.

4.2 Scaling in Measurement Space

In this example, we are scaling truss displacement so that they are equal to beam's displacements under the same loading, $\mathbf{y}_1 = \mathbf{S}_y\mathbf{y}_2$. Two single displacement vectors can be related in an arbitrary way. One possibility is to use generalized inverse matrix in the form of $\mathbf{S}_y = \mathbf{y}_2^T \left(\mathbf{y}_2\mathbf{y}_2^T\right)^{-1} \mathbf{y}_1$that corresponds to the minimum length solution. However, scaling results are poor under the general loading (for single loading case $\mathbf{S}_{\mathrm{y}}$ degenerates into a scalar). Under the assumption of independence of displacements of the truss and the beam, scaling matrix $\mathbf{S}_{\mathrm{y}}$ is diagonal with elements $(\mathbf{S}_y)_{i,i} = {}^{(\mathbf{y}_1)_i}/_{(\mathbf{y}_2)_i}$.

In Fig. 5 there is a comparison of scaled displacements for the two methods and diagonal scaling matrix $\mathbf{S}_{\mathrm{y}}$ gives excellent results.

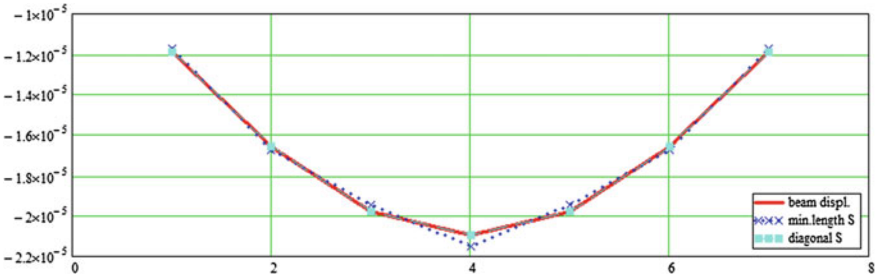


Fig. 5 Comparison of displacements of truss and beam with scaled loading

$$\Omega 2 = \begin{pmatrix} 0.141 \\ 0.011 \\ 4.57 \times 10^{-3} \\ 2.474 \times 10^{-3} \\ 2.28 \times 10^{-3} \\ 1.42 \times 10^{-3} \\ 1.42 \times 10^{-3} \end{pmatrix}$$

$$\Omega 1 = \begin{pmatrix} 2.325 \times 10^{3} \\ 107.792 \\ 20.342 \\ 5.662 \\ 2.41 \\ 1.146 \\ 0.868 \end{pmatrix}$$

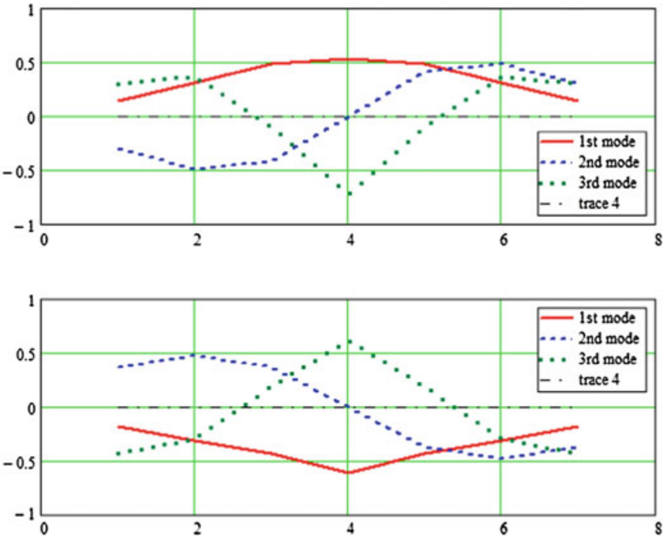


Fig. 6 Comparison of eigenmodes of truss and beam

4.3 Scaling of Dynamic Properties

We would like to compare eigenfrequencies and modal shapes (eigenvalues) of a model and a structure, so they have to be scaled. In our example structure is the truss and model is the beam from above and both have lumped mass matrices. Truss and beam eigenmodes are presented in Fig. 6.

Eigenmodes in Fig. 6 cannot be scaled ‘per se’; mass and stiffness matrices have to be scaled to produce the same eigenvalus, instead. Matrix $\mathbf{I}_S$ from Eq. (6) is made equal to $\mathbf{S}_y$ from example 4.2 producing scaling matrices $\mathbf{S}_s = \mathbf{K}_1\mathbf{I}_S\mathbf{K}_2^{-1}$ and $\mathbf{S}_d = \mathbf{M}_1\left(\mathbf{I}_S^2\right)^{-1}\mathbf{M}_2^{-1}$. The scaled truss dynamic matrix $\mathbf{D}_s = \left(\mathbf{S}_d\mathbf{M}_2\mathbf{I}_S^2\right)\left(\mathbf{S}_s\mathbf{K}_2\mathbf{I}_S^{-1}\right)^{-1}$ has the same eigenvalues and eigenvectors as the beam dynamic matrix $\mathbf{D}_1 = \mathbf{M}_1\mathbf{K}_{con}^{-1}$. Now, relating the appropriate dynamic matrices can scale eigenvalues using the eigenvalue decomposition theorem: from ‘r’ measured eigenvalues Ω and eigenvectors ϕ dynamic matrices are formed $\mathbf{D} = \sum_r \Omega_r \Phi_r \left(\Phi_r^{-g}\right)^T$ and eigenmodes related.

4.4 Force Reconstruction—Static Loading

Force reconstruction is demonstrated in the example where a console is loaded with a force and a moment at the top. It is assumed that a number of measurements are performed at 3 places near the top; actually, measurements are synthetically generated according to the Gauss distribution with mean equal to the exact displacements for force and moment $F = 10.0$ and $M = 5.0$ and $\sigma = 5\,\%$. The Wolfram Mathematica [10] is used for modeling the example. Applied reconstruction procedures are least squares (LS) and weighed least squares (WLS).

Measurement equation is established $\mathbf{y} = \mathbf{HF}$ and measuring matrix $\mathbf{H}$ formulated, where $\mathbf{y}$ are measured values and $\mathbf{F}$ is the loading vector whose components are to be reconstructed from measurements. Since this is a simulation, we know the exact value of the loading vector $\mathbf{F} = \left\{ \begin{array}{c} F = 10.0 \\ M = 5.0 \end{array} \right\}$. It is inevitable that measurements contain some noise and to reduce its influence to some extent, as many as possible measurement points are needed; the result is more rows then columns in matrix $\mathbf{H}$. Finding the generalized or Moon-Penrose inverse gives the solution in the LS sense $\mathbf{F} = \left\{ \begin{array}{c} F = 10.031 \\ M = 4.778 \end{array} \right\}$ for measured values with small differences from the 'exact' ones $\Delta = \left\{ \begin{array}{c} -0.022 \\ -0.007 \\ +0.006 \end{array} \right\} [\%]$. Also, a-posterior residual is small $\Delta = \left\{ \begin{array}{c} -0.013 \\ +0.009 \\ -0.007 \end{array} \right\} [\%]$ but parameter variance is very large giving very large error bounds $\left\{ \begin{array}{c} \sigma_F = 7.5 \\ \sigma_M = 52.8 \end{array} \right\}$.

The loading reconstruction result can be further improved if there are more measurements in each measuring point, as illustrated in Fig. 7. This problem can be solved as a LS problem as above, by replacing each series of measurements with its mean value but better results are obtained with WLS method.

Weighing matrix $\mathbf{W}$ is constructed from variances of the series of data measurements so that worse measurements are less weighed. A-posterior variance matrix is then $\mathbf{P} = \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1}$ and the reconstructed solution is

$$\left\{ \begin{array}{c} F \\ M \end{array} \right\}_m = \mathbf{P}\mathbf{y}_W, \quad \text{with } \mathbf{y}_W = \mathbf{H}^T \mathbf{W} \delta_S \tag{17}$$

The solution gives $\mathbf{F} = \left\{ \begin{array}{c} F = 9.99 \\ M = 5.02 \end{array} \right\}$, which is only marginally better then LS but parameter variance is much smaller giving tighter error bounds $\left\{ \begin{array}{c} \sigma_F = 0.836 \\ \sigma_M = 5.832 \end{array} \right\}$. It can be seen in both methods that F is better resolved then M although there is a high correlation between them -0.999.

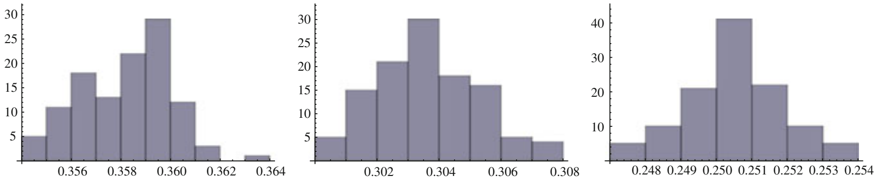


Fig. 7 Histograms of 114 generated measurements at 3 points on the console

In the above examples the exact position of the loading has been known but that is not always the case, e.g. in wind turbines besides the turbine pressure there is also a wind force along the tower.

4.5 Force Reconstruction—Dynamic Loading

Force reconstruction under the dynamic loading is significantly more difficult because there is no well-defined mean value; force is constantly changing in time and an inverse procedure different from above is required. As stated before, inverse procedure for dynamic loading recovery is formulated in modal space. In order to assess quality of the procedure, both, forward and inverse problems are solved and results compared.

Our model console has the following properties: L = 5.0 m, EI = 500.0 kNm2, ρ = 2.0 t/m.

Discretization: No. space intervals n = 50, No. time intervals m = 1000, time increment ΔT = 0.005 s, total analysis time = 5.0 s.

Structure eigenfrequencies: $f_1 = 0.35$ Hz, $f_2 = 2.17$ Hz, $f_3 = 6.08$ Hz, $f_4 = 11.88$ Hz, etc.

Forward problem
Dynamic loading acts in three points: (1) at the top, (2) 10% below the top, (3) at the middle. Force amplitude and frequency is according to equations (Fig. 8).

$$\begin{array}{lll}
P_1(t) = \frac{P_1}{\Delta x} sin\left[\frac{2\pi t}{T_{period}}\right] & P_1 = 0.1kN & T_{period} = 2.0\,sec \\
P_2(t) = \frac{P_2}{\Delta x} sin\left[\frac{2\pi(t+0.885T_{period})}{T_{period}}\right] & P_2 = 0.05kN & T_{period} = 2.0\,sec \\
P_3(t) = \frac{P_3}{\Delta x} sin\left[\frac{2\pi(t+1.865T_{3period})}{T_{3period}}\right] & P_3 = 0.01kN & T_{3period} = 3.0\,sec
\end{array}$$

Note that all three forces are out of phase. Force 3 has been given the period close to the first eigenfrequency of the structure but with much smaller amplitude; the result is visible in accelerations in Fig. 11.

Loading is discretized in advance to speed up the calculation and discretizion in space (50 intervals) and time (1000 intervals) looks as in Fig. 9 (50000 points in total).

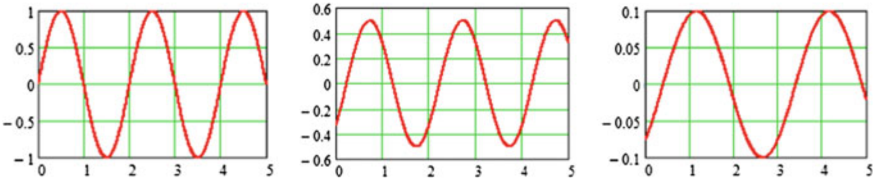


Fig. 8 Dynamic loading in points 1, 2 and 3 on the console

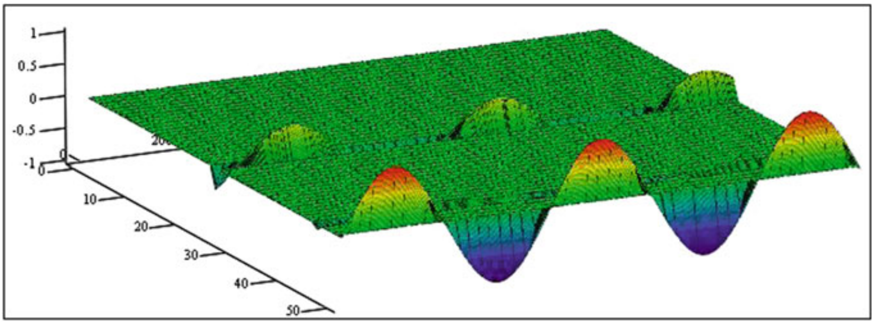


Fig. 9 Dynamic loading in points 1, 2 and 3 on the console, discretized in space and time

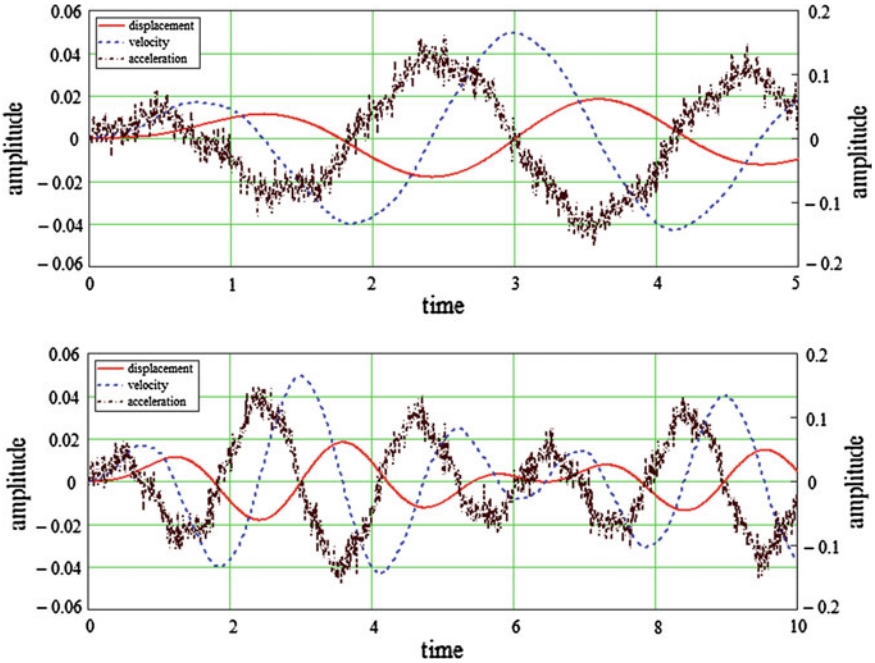


Fig. 10 Comparison of response in time for the *top* point of the console, ΔT = 0.005 s and ΔT = 0.01 s

After calculation with modified Newmark method (see [9] and [11]), displacements, velocities and accelerations are obtained at each point at each time increment. Results in time for the top point of the console are visible in Fig. 10 and results in space and time for the whole structure are in Fig. 11 (total of 50000 points for each result, units are [m], [m/s] and [m/s^2] respectively). We see that velocities and accelerations approximately follow the displacements with some shift in phase. Also, accelerations are not smooth but one can still recognize the shape of displacements in it; it is so only when there is some loading with the period close to one of the

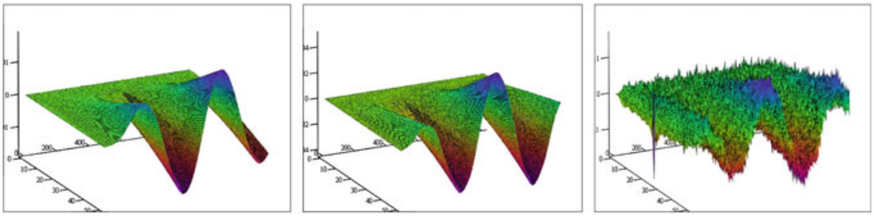


Fig. 11 Total response in space and time of the whole structure ΔT = 0.005 s

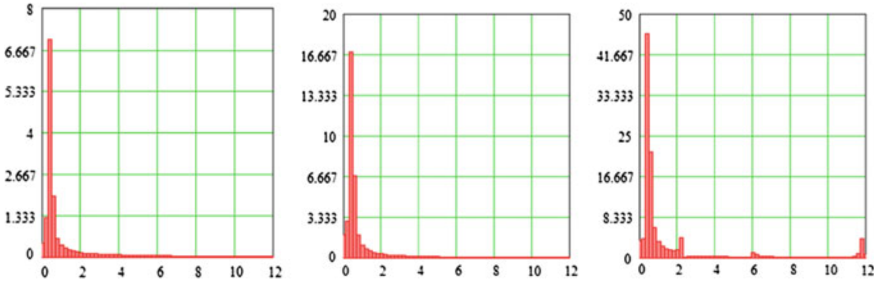


Fig. 12 DFT of the structure displacement, velocity and acceleration ΔT = 0.005 s

eigenfrequencies, otherwise, accelerations look quite messy. Out of phase forces at the initial time increment (t = 0) are well visible in the accelerations plot in Fig. 11.

After results have been obtained in space and time, discrete Fourier transform (DFT) of the results is performed. DFT has been performed for the top point of the structure for each group of the results separately: displacements, velocities and accelerations.

DFT is an important step in understanding how measurements are to be performed and what information can be extracted from the data. Space discretization gives us measurement points and time discretization represents sampling frequency; by varying them we can simulate and optimize the measurement process.

DFT in Fig. 12 gives information about structure eigenfrequencies. Higher frequencies are less pronounced and less precise due to spectral leakage, i.e. there are DFT results around real frequencies that carry power through the structure. Spectral leakage is a consequence of DFT performed over a finite time interval, a window in the time domain of which the exact FFT result is a function of infinite width ($\mathrm{Sinc}(x) = \frac{\sin(x)}{x}$), see [12]. Also, from Fig. 12 it is evident that measurement instruments have to be chosen carefully, accelerations carry more information about frequencies and it is easier to extract eigenfrequencies from them. However, frequency information depends on the frequency resolution, which is function of sampling interval $\Delta_{frequency} = \frac{1}{\Delta T}$. The same analysis is performed with $\Delta T = 0.01$ so that the total analysis time is 10 s. While space and time results do not differ much (see Fig. 13), DFT has better frequency resolution and longer time interval, resulting in less spectral leakage (Figs. 14 and 15). By careful numerical simulation we could deduce about necessary number of samples and sampling time of the instrument.

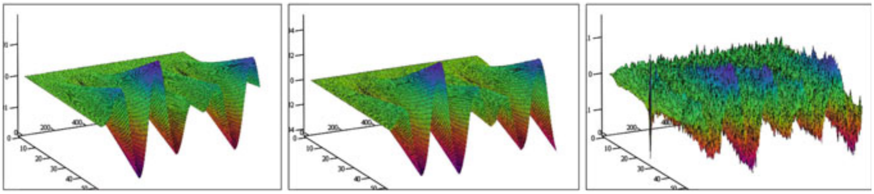


Fig. 13 Total response in space and time of the whole structure $\Delta T = 0.01$ s

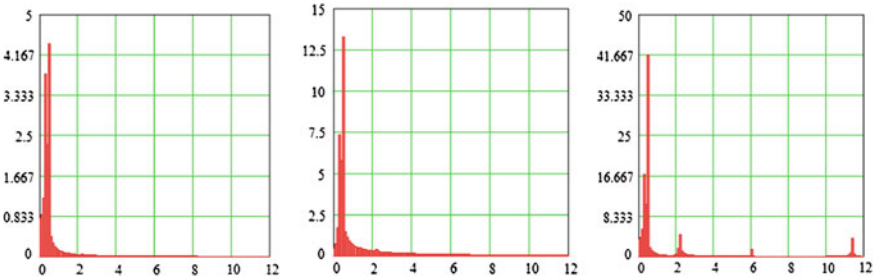


Fig. 14 DFT of the structure displacement, velocity and acceleration $\Delta T = 0.01$ s

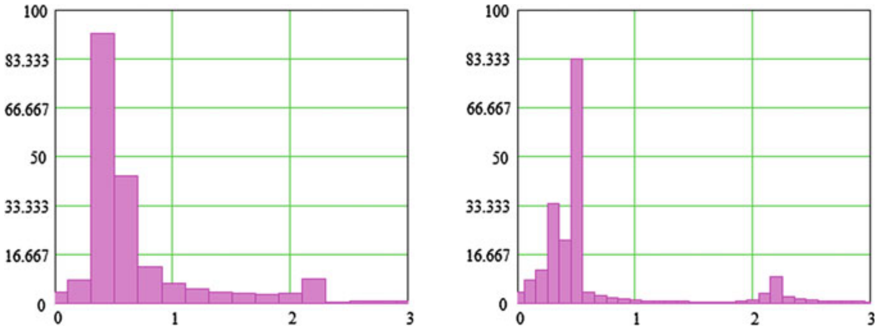


Fig. 15 Comparison of DFT from acceleration for $\Delta T = 0.005$ s and $\Delta T = 0.01$ s

Influence of frequency resolution (a function of time increment in the model or sampling interval during measurement) is best visible in Fig. 15. In both figures the second eigenfrequency ($f_2 = 2.17$ Hz) is clearly recognizable but in the first picture, the first eigenfrequency ($f_1 = 0.35$ Hz) is hidden among loading frequencies ($p_1 = 0.5$ Hz and $p_2 = 0.333$ Hz). Second picture has better frequency resolution and two loading frequencies are separated with the first structure eigenfrequency visible among them.

Inverse problem—displacement measurement
We are trying to recover loading for the console loaded with two dynamic forces acting in the same points as in the example above. Lower positioned force has half the amplitude and is shifted in phase

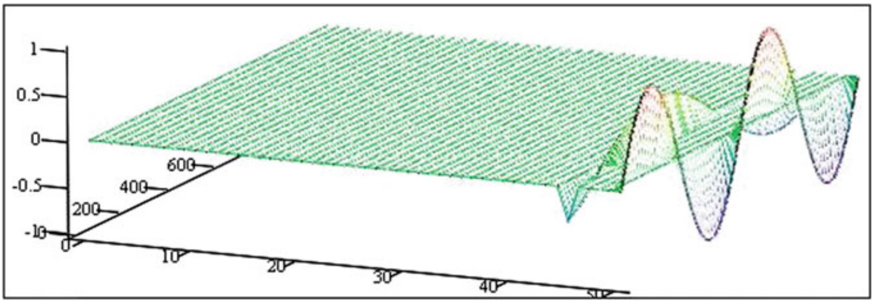


Fig. 16 Dynamic loading in points 1 and 3 on the console, discretized in space and time

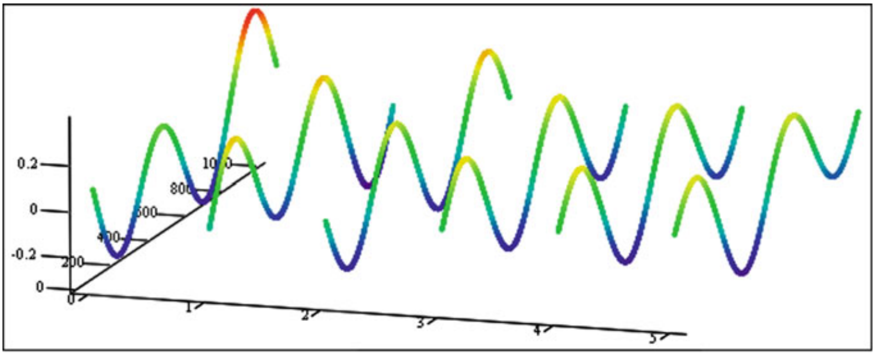


Fig. 17 Dynamic loading from Fig. 16 transformed into modal space

$$P_1(t) = \frac{P_1}{\Delta x} sin\left[\frac{2\pi t}{T_{period}}\right] \qquad P_1 = 0.1kN \qquad T_{period} = 5.0\,sec$$

$$P_3(t) = \frac{P_3}{\Delta x} sin\left[\frac{2\pi(t+1.865T_{3period})}{T_{3period}}\right] \qquad P_3 = 0.05kN \qquad T_{3period} = 10.0\,sec$$

Loading is discretized in advance to speed up the calculation and discretizion in space (50 intervals) and time (1000 intervals) looks as in Fig. 16 (50000 points in total).

Calculation is performed in modal space using the first five eigenmodes ($r = 5$) so loading is transformed into modal space using $\mathbf{\Phi}_{\text{mod}} = \mathbf{\Phi}_r\mathbf{P}$ and displayed in Fig. 17.

Modal dynamic equation is formed according to Eq. (13); the two dynamic systems have the same eigenmodes. For the sake of accuracy assessment, forward analysis is performed so that we can compare recovered values of displacements, velocities and accelerations with the exact ones. Of course, all the calculated values are in modal space.

Measurement vector is defined as $\text{t}_\text{m} = (50\ 45\ 40\ 30\ 25\ 10)$ where the point 50 is at the top of the console. State and measurement equations depend on the measured values: displacements, velocities or accelerations where the parameter of the state equation is $\mathbf{x}(\text{t}) = \{\text{d(t) v(t)}\}^\text{T}$ (i.e., displacement and velocity) (Table 1).

Table 1 State and measurement equations depending on the measured variable

State equation	$\dot{\mathbf{x}}(t) = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & \mathbf{0} \end{bmatrix} \mathbf{x}(t) + \begin{Bmatrix} \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{H}_0 \end{Bmatrix} \mathbf{p}(t)$
Measurement equation for displacements	$\mathbf{y}(t) = \begin{Bmatrix} \mathbf{H}_0 \\ \mathbf{0} \end{Bmatrix} \mathbf{x}(t)$
Measurement equation for velocities	$\mathbf{y}(t) = \begin{Bmatrix} \mathbf{0} \\ \mathbf{H}_0 \end{Bmatrix} \mathbf{x}(t)$
Measurement equation for accelerations	$\mathbf{y}(t) = \begin{Bmatrix} -\mathbf{H}_a\mathbf{M}^{-1}\mathbf{K} \\ \mathbf{0} \end{Bmatrix} \mathbf{x}(t) + \begin{Bmatrix} \mathbf{H}_a\mathbf{M}^{-1}\mathbf{H}_0 \\ \mathbf{0} \end{Bmatrix} \mathbf{p}(t)$

Modal measuring matrix is formed from measuring matrix by using Eq. (15). Although is the measurement matrix $\mathbf{H}_0$ very sparse, the modal measurement matrix $\mathbf{H}_{\text{mod}}$ is dense; this is important for obtaining the well posed generalized inverse of the modal measurement matrix $\mathbf{H}_{\text{mod}}^{-g}$.

Calculated modal displacements are polluted with artificial Gaussian noise to serve as a measured input value. The noise is zero-mean small standard deviation of 0.1 % as presented in Fig. 18.

Modal displacements are recovered from 'noisy' measurements using Eq. (16) and are compared with exact values in Fig. 19.

The first mode displacements in Fig. 19 are almost exact but in the second mode errors are visible; this trend continues for higher modes.

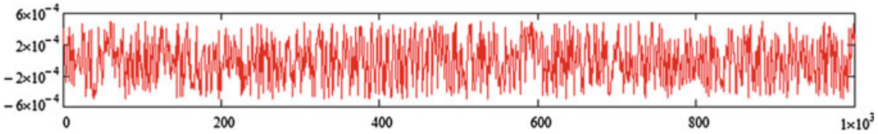


Fig. 18 Gaussian noise introduced to simulate measurements

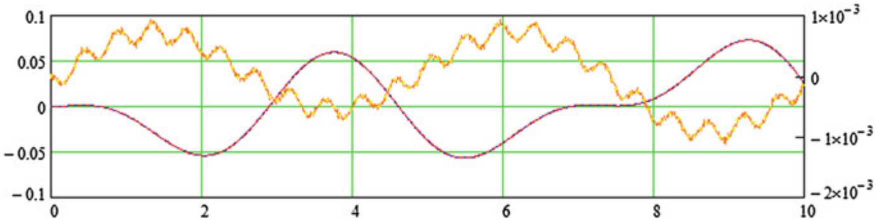


Fig. 19 Comparison of recovered and exact modal displacements

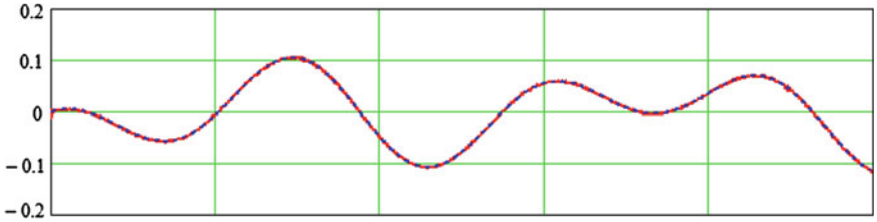


Fig. 20 Comparison of recovered and exact modal velocities

Modal velocities are recovered from modal displacements by derivation. Actually, this is not advisable since derivatives significantly amplify the noise; e.g. if one performs the derivation on a noisy harmonic function $\sin(\omega y + w)$, the result is function amplified ω times, i.e. $\omega \cos(\omega y + w)$. This problem is addressed in [8]; here we use Pade derivative, which is much more exact then finite differences of any form since it uses all the points to produce a derivative. Note that we do not need to know the function to calculate its derivative; discrete points are sufficient. For a discrete function represented with 1000 points that signifies solving of a system of 1000 linear equations. Pade derivative is formulated using matrix differentiation operator similar to [11], $\mathbf{D}_{mat}\mathbf{d}_u = d_P(\mathbf{u}, x_a, x_b, n)$ where $\mathbf{d}_\mathrm{u}$ is vector of derivatives, **u** vector of function points, $\mathrm{x_a}$, $\mathrm{x_b}$ limits of data points interval, n number of points, $\mathbf{D}_\mathrm{mat}$ is (sparse) matrix of Pade coefficients and $\mathbf{d}_\mathrm{P}$ is derivative vector produced from function's points. In our example, where $\mathbf{d}_\mathrm{u}$ is vector of modal velocities ϕ_v, **u** vector of modal displacements ϕ_d, $\mathrm{x_a} = 0$, $\mathrm{x_b} = \mathrm{T_{total}}$ time interval of the analysis, n=m number of time increments. After solving the equation $\boldsymbol{\phi}_v = \mathbf{D}_{mat}^{-1} d_P\left(\boldsymbol{\phi}_d, 0, T_{total}, m\right)$, the result is points representing the velocities. In Fig. 20 there is a comparison of exact velocities and those calculated through Pade derivative.

Noise in reconstructed modal velocity vector is barely visible; coincidence with exact values is very good. One of tests the inverse formulation has to pass is its ability to completely and without any error recovers the modeled function in the absence of noise. This formulation fulfills this task completely for all recovering quantities.

Further derivation of velocities produces accelerations needed for our state equation. After solving the equation $\boldsymbol{\phi}_a = \mathbf{D}_{mat}^{-1} d_P\left(\boldsymbol{\phi}_v, 0, T_{total}, m\right)$, the result is points

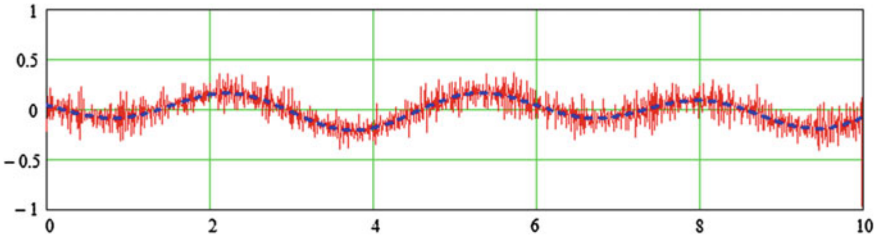


Fig. 21 Comparison of recovered and exact modal accelerations

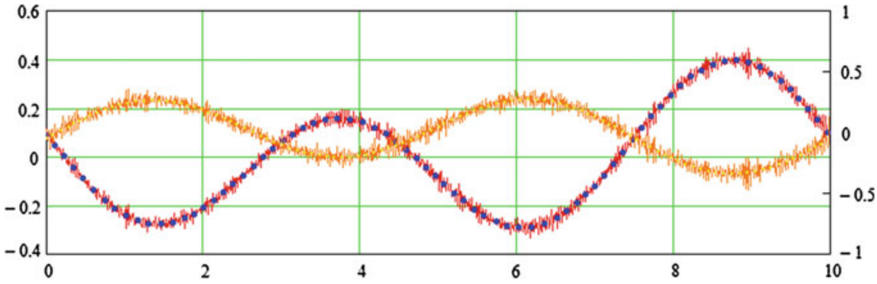


Fig. 22 Comparison of recovered and exact modal loading for modes 1 and 2

representing the accelerations. Comparison of exact accelerations and those calculated by solving this equation is in Fig. 21.

Noise in reconstructed modal accelerations is more then noticeable. It would be advisable to eliminate the noise as much as possible in every step of the force reconstruction. Nevertheless, in this example we are proceeding without noise elimination in every step of the force recovery. With recovered modal displacements and modal accelerations at hand, we can recover modal loading using Eq. (13). The result in Fig. 22 is a comparison of the exact and recovered modal loading for modes 1 and 2. Noise in the recovered modal loading seems to be at the level of modal accelerations.

In final stage of the loading recovery process global load is recovered from modal load using Eq. (16). There are some restrictions in the application of Eq. (16); number of modes in the eigenmatrix $\mathbf{\Phi}_r$ has to be such that the matrix is regular, i.e. 'r' has to be equal or greater then the number of measuring points; in the contrary, $\mathbf{\Phi}_r$ has to be regularized and special recovery procedures have to be used. Also, load can be recovered only in points where measurements have been made (this too, can be avoided using special formulations but they will not be discussed here). With all the restrictions applied, the result of force recovery for loaded points 50 and 45 on the console is compared with exact values in Fig. 23.

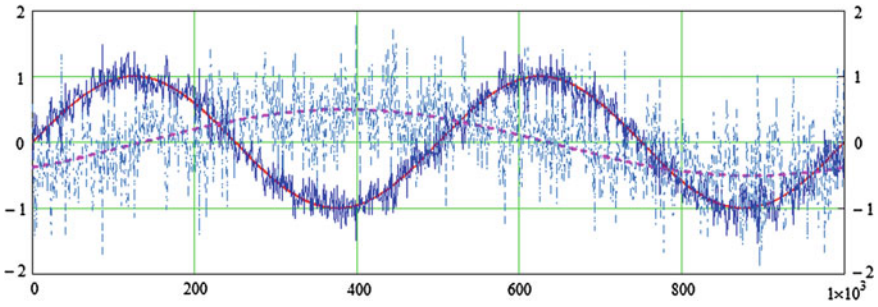


Fig. 23 Comparison of recovered and exact global loading in points 50 and 45 of the console

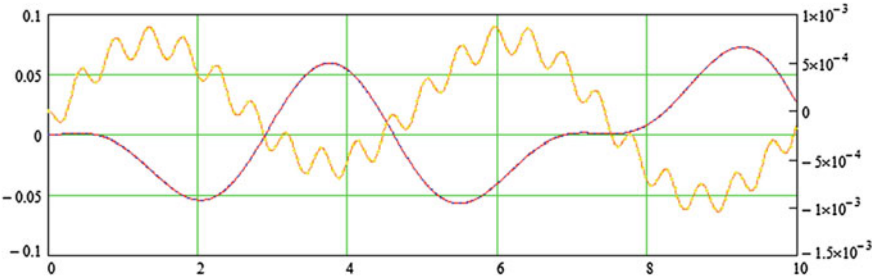


Fig. 24 Comparison of recovered and exact modal displacements obtained from measured velocities

Recovered load in point 50 looks satisfactory but load in point 45 is rather noisy. However, the mean value of recovered forces seems to follow the required trend and the result can be further enhanced using various techniques. Other nodes are not loaded and their recovered forces should be zero; in reality small values are obtained. Instead of improving on this result a posterior, we will try to use a priory approach by obtain better measurements.

Inverse problem—velocity measurement

We are trying recovery process by measuring velocities instead of displacements; there will be no need to differentiate twice to obtain accelerations and differentiation is considered one of great sources of error. Recovery procedure is the same up to Fig. 18, after which we have to integrate velocities to obtain displacements. This time Pade matrix differentiation operator is modified so that it can be used for integration as well; it is modified so that vector function $\mathbf{d}_{\mathrm{P}}$ is not longer needed. The derivation equation is now $\mathbf{d}_u = \mathbf{D}_d\mathbf{u}$ where $\mathbf{D}_d = \mathbf{D}_{mat}^{-1} d_d\ (x_a, x_b, n)$ and $\mathbf{d}_{\mathrm{d}}$ is modified vector function $\mathbf{d}_{\mathrm{P}}$ so that it does not include $\mathbf{u}$. Now, the integration is simply the inversion of matrix $\mathbf{D}_{\mathrm{d}}$. Comparison of modal displacementsin modes 1 and 2 in Fig. 24 resembles modal displacements from Fig. 19.

The modal accelerations in mode 1 and mode 2 are obtained by derivation of velocities as before; the result is in Fig. 25.

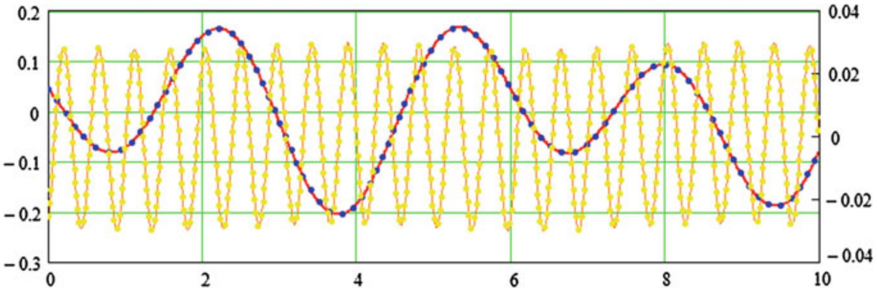


Fig. 25 Comparison of recovered and exact modal accelerations

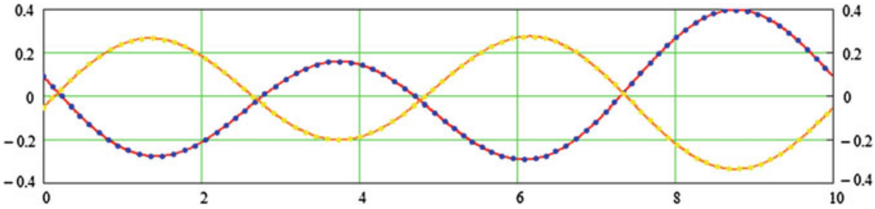


Fig. 26 Comparison of recovered and exact modal loading for modes 1 and 2

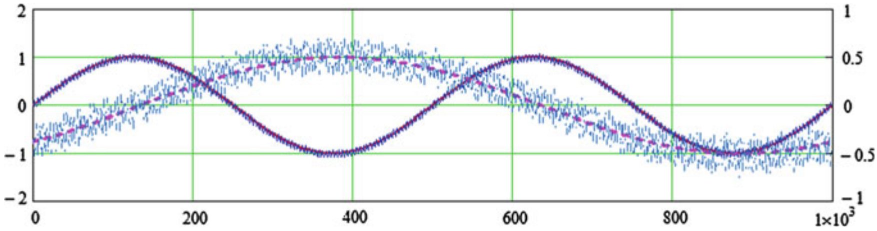


Fig. 27 Comparison of recovered and exact global loading in points 50 and 45 of the console

Noise in the modal accelerations reconstructed from velocities is barely visible.

Reconstruction of modal loading follows; comparison of reconstructed and exact values is presented in Fig. 26 (compare with Fig. 22—recovery from displacements).

Finally, the result of force recovery for loaded points 50 and 45 on the console is compared with exact values in Fig. 27 (compare with Fig. 23—recovery from disp lacements).

Recovered load in point 50 looks very good and it is satisfactory in point 45. In this example, too, nodes that are not loaded show some small recovered forces instead of being zero. The mean value of recovered forces follow the required trend even more then previous results in Fig. 23 and further enhancing of the result should be straightforward.

5 Discussion and Conclusion

In the paper we have demonstrated how two structures, usually a real structure and its model, could be related in order to gain better insight into behavior of large structures. The relation is established after discretization, through scaling of one of the structures to the scale of the other one. Structures are considered undamaged, so we are dealing with linear systems of equations. Author is especially interested in relating measurements and structure parameters. As a consequence, linear systems can be over or underdetermined, depending on the number of parameters and measuring points on the structure. Structure parameter under consideration is loading and, since load mostly cannot be directly measured, its recovery from indirect measurements.

Paper distinguishes between scaling in the parameter space and in measurement space, between static and dynamic loading. Recovery of static loading is performed using least squares and weighed least squares and recovery of dynamic loading is based on inverse modal analysis. Inverse modal analysis applied for force reconstruction involves several stages: recovery of modal displacements, modal velocities and modal accelerations from some measurements. In the nest stage modal load is recovered. Finally, global load is recovered from modal load. Elimination of the noise, as much as possible, in every step of the force reconstruction is planned in future development.

Discrete Fourier transform is applied to gain better insight into various stages of load recovery process. Its main use is to help us to reduce the noise in measurement through selection of the measuring values and optimization of the measurement process.

Numerous examples in the paper explain the whole process of structure scaling and load reconstruction [13].

Acknowledgments This paper has been partial founded through Croatian Science Foundation project 9068 Multi-scale concrete model with parameter identification, which is fully appreciated.

References

1. A. Maciag, A. Pawinska. *Solution of the direct and inverse problems for beam*, Comp.Appl.Math. Springer. doi:10.1007/s40314-014-0189-9
2. A. Ghosh, *Mechanics Over Micro and Nano Scales*, Editor S. Chakraborty, Springer 2011.
3. F.D. Bianchi, H. de Batista, R.J. Mantz, *Wind Turbine Control Systems*, Springer 2007.
4. J. Sanchez, H. Benaroya. *Review of force reconstruction techniques*, Journal of Sound and Vibration, vol 333, page 2999–3018, 2014.
5. X. Liu, P.J. Escamilla-Ambrosio, N.A.J. Lieven. *Extended Kalman filtering for the detection of damage in linear mechanical structures*, Journal of Sound and Vibration, vol 325, page 1023–1046, 2009.
6. R.C. Aster, B. Borchers, C.H. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, 2013.
7. B.P. Gibbs, *Advanced Kalman Filtering, Least-Squares and Modeling*, John Wiley & Sons, 2011.
8. I. Kožar, D.L. Kožar, Ž. Jeričević, *A note on the reservoir routing problem*, European Journal of Mechanics B/ Fluids, vol 29, page 522–533, 2010.
9. I. Kožar, *Security aspects of vertical actions on bridge structure: Comparison of earthquake and vehicle induced dynamical forces*, Engineering Computations, vol 26, page 145–165, 2009.
10. S. Mangano, *Mathematica Cookbook*, O'Reilly 2010.
11. I. Kožar, N. Torić Malić, *Spectral method in realistic modeling of bridges under moving vehicles*, Engineering Structures, vol 50, page 149–157, 2013.
12. M.J. Roberts, *Signals and systems*, McGraw-Hill, 2012.
13. W.E. Sabine, *Discrete-signal analysis and design*, John Wiley & Sons, 2008.

Fat Latin Hypercube Sampling and Efficient Sparse Polynomial Chaos Expansion for Uncertainty Propagation on Finite Precision Models: Application to 2D Deep Drawing Process

Jérémy Lebon, Guénhaël Le Quilliec, Piotr Breitkopf, Rajan Filomeno Coelho and Pierre Villon

Abstract In the context of uncertainty propagation, the variation range of random variables may be many oder of magnitude smaller than their nominal values. When evaluating the non-linear Finite Element Model (FEM), simulations involving contact/friction and material non linearity on such small perturbations of the input data, a numerical noise alters the output data and distorts the statistical quantities and potentially inhibit the training of Uncertainty Quantification (UQ) models. In this paper, a particular attention is given to the definition of adapted Design of Experiment (DoE) taking into account the model sensitivity with respect to infinitesimal numerical perturbations. The samples are chosen using an adaptation of the Latin Hypercube Sampling (Fat-LHS) and are required to be sufficiently spaced away to filter the discretization and other numerical errors limiting the number of possible numerical experiments. In order to build an acceptable Polynomial Chaos Expansion with such sparse data, we implement a hybrid LARS+Q-norm approach. We showcase the proposed approach with UQ of springback effect for deep drawing process of metal sheet, considering up to 8 random variables.

Keywords Uncertainty quantification · Model sensitivity · Springback variability assessment · Sensitivity-constrained design of experiment · Sparse polynomial chaos expansion

J. Lebon · G. Le Quilliec · P. Breitkopf (✉) · P. Villon
Laboratoire Roberval, UMR 7337, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne cedex, France
e-mail: piotr.breitkopf@utc.fr

J. Lebon · R. Filomeno Coelho
BATir Department, Université Libre de Bruxelles (ULB),
CP 194/2, 50, Avenue Franklin Roosevelt, 1050 Brussels, Belgium

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_8

1 Introduction

The hierarchical combination of a "high-fidelity" (expensive and accurate) model with a "lower-fidelity" model (less accurate but also less expensive) may lead to significant decrease of the computational cost involved in the search for optimal design [1–3], in the characterization of system variability [4, 5] or in sensitivity studies [6–8]. In an optimization context, rigorous convergence proofs may be established as long as the "lower-fidelity" model is consistent (generally to the first [9] or second order [10]) with the "higher-fidelity" model. "Multi-fidelity" or "variable-fidelity" refer to models built from simplified physics [11], coarse discretization [12–14] or partial convergence [15]. Multi-fidelity surrogate-based approaches [16–18] mostly refer to interpolating or regression based non-physics metamodels such as polynomial response surface [19], moving least squares [20, 21], kriging, [22], etc.

Due to strong mathematical basis and functional representations of stochastic variabilities, Polynomial Chaos Expansions (PCE) are attractive for uncertainty quantification. Notably, [23] PCE combined with FE in an intrusive manner forms the basis of the stochastic finite element method, [24] provides a non-intrusive combination scheme. [25–27] focuses on PCE computational costs by non intrusive adaptive schemes to perform robust, reliability, and sensitivity analysis.

In metal forming applications, Monte Carlo simulation on a polynomial response surface permits to quantify statistical quantities (mean and standard deviation) [28, 29] uses Monte Carlo simulations and linear response surfaces to identify the most significant variable and to build an approximation of the probabilistic response. A second order Polynomial Chaos Expansion (PCE) is used in [30] to assess the variability of the tolerance prediction of the formed metal sheet submitted to random process and material data and [31] uses the moving least squares instead of the classical quadratic order response surface to perform reliability analysis of the sheet metal forming process.

However, nonlinear Finite Elements may suffer from numerical instabilities resulting from round-off errors, convergence errors and discretization errors [32]. In the general case, nonlinear FEM does not converge in a monotonous way to the equilibrium state. Taking into account nonlinear phenomena such as large strain, material non linearities, contact-friction, etc., requires convergence criteria implying limited precision.

Contact problems are also likely to provide numerical instabilities and sensitivity problems. Solving a contact problem requires to determine the closest distance between surfaces. [33] identifies in Sect. 2.2.4, Fig. 2.8 the projection zone of the slave points with regards to their paths, and concludes that "a small change in position of the slave point does not always result in a small change in the closest point projection". This difficulty is particularly likely to occur when the slave point is relatively far from the master surface. Then if a small change in the geometry results in a jump in the projection point, the corresponding dissipated energy will also experience a jump: the virtual work of frictional forces becomes discontinuous.

Focusing now on the FEM of deep drawing process, sources of instability come from the discretization errors [34]. The through-thickness stress profile defines the internal bending moment which governs the springback phenomenon when removing tools. When the material undergoes plastic deformation, the stress profile becomes non-smooth due to the presence of elastic-plastic transitions.

The observed instabilities of the FE model do not provide a clearly established reference from which we may evaluate the bias or the convergence rate of a surrogate model. In this context a “good” surrogate model:

- intrinsically contains convergence properties to the actual response, when the number of training points increases,
- allows to retrieve statistical moment analytically,
- is economical in terms number of high fidelity simulations.

In this paper we propose a custom PCE extrapolating the physical model towards “very small” perturbations.

We propose to filter the numerical instabilities and to build a custom polynomial surrogate of the response. The construction of surrogates in presence of epistemic and random uncertainties is a key component of experimental design [35, 36]. Assessing the parameters of the response surface requires to define these parameters as random variables. Various statistical criteria (D-, A-, T-, E-optimality) allow to assess the quality of the sampling scheme by decreasing the bias and the variance of the model parameters to evaluate. In such cases, the parameters have to be sampled either from their prior or posterior distribution, both requiring an extensive number of calls to the “high-fidelity” model.

The proposed methodology provides:

- the maximal numbers of samples that may be generated taking into account the finite precision of the FE simulations softwares, and an associated Latin Hypercube Sampling [37],
- the “best” PCE fit using an adaptive algorithm combining Q-norm and Least Angle Regression Stagewise algorithm.
- the control of the quality of the PCE by Leave One Out error.

The exhaustive identification of the numerical artefacts and their deepened analysis is beyond the scope of this paper and is still an open research area. Here we propose a pragmatic methodology which is not aimed at removing the bias induced by the noisy behavior of the FEM high-fidelity model.

The paper is organized as follows. Section 2 proposes a general insight into errors produced by FEM modeling and a global error assessment based on the finite difference scheme. In Sect. 3 an introduction to the basics of PCE is provided. Section 4 provides a sampling scheme based on the Latin Hypercube Sampling (LHS), taking the model resolution into account. In Sect. 5 we implement three different economical (sparse) PCE schemes to efficiently and accurately propagate the uncertainty. Finally, in Sect. 6 we showcase the efficiency of the proposed approach on the deep drawing process of a 2D, U-shaped metal sheet considering up to 8 random variables. Conclusions and prospectives are provided in Sect. 7.

2 FEM Error Assessment Using Finite Difference Scheme

We consider random perturbations on the input parameters $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_M]$ around a nominal value $\boldsymbol{\xi}_{\text{nom}} = [\xi_1^{\text{nom}}, \xi_2^{\text{nom}}, \ldots, \xi_M^{\text{nom}}]$ and their influence on the output function $y(\boldsymbol{\xi})$. To characterize the stability of the model output, we evaluate the relative adimensional sensitivity of the computational model separately for each variable ξ_i, $i \in \{1, \ldots, M\}$ using the finite difference scheme:

$$\mu_i = \frac{\Delta y(\xi_i)}{\Delta \xi_i} \times \frac{\xi_i^{\text{nom}}}{y(\xi_i^{\text{nom}})} \tag{1}$$

where

$$\Delta y(\xi_i) = y\left(\xi_i^{\text{nom}} + \frac{\Delta \xi_i}{2}\right) - y\left(\xi_i^{\text{nom}} - \frac{\Delta \xi_i}{2}\right). \tag{2}$$

From actual computation of non linear Finite Element (see Sect. 6.2.4), we observe that when decreasing the order of magnitude of the perturbation ($-\log(\Delta\xi)$ increasing), one may identify different behaviors of μ_i (Fig. 1):

1. Firstly, for “large” $\Delta\xi_i$, the variation of μ_i reveals a non-linear behavior of the model and the model may be considered as trustworthy.
2. Secondly, μ_i stabilizes around a constant value $\bar{\mu_i}$ where the model may be considered as linear. This is the zone used for finite difference gradients (e.g. in optimization).
3. Thirdly, on reaching the threshold Ⓑ, μ_i becomes unstable. We call the corresponding $\Delta\xi_i^*$ the *resolution* threshold of the model. Below, unstable data is generated and may not be used when training a metamodel.
4. Finally, Ⓐ shows the model sensitivity limit: $\Delta y = 0$.

The point Ⓑ may be considered as the limit resolution of the model (in the following we simply refer to the *resolution* of the model). The approach developed in this paper aims at the cases when random variation of input parameters encompasses the instability zone.

The question for an automatic identification of the resolution threshold Ⓑ is left open. For practical implementation, we consider the model output unstable when one simultaneously observes a significant change in the magnitude of μ_i followed by algebraic sign inversions.

3 Introduction to Polynomial Chaos Expansion

The PCE [38] is a stochastic metamodel, that is intended to give an approximation $\tilde{y}$ of the stochastic behavior of a functional y (*scalar random process*) that is defined as a function of an input *M-dimensional* random vector $\boldsymbol{\xi}$. We assume that y is a second order random variable ($\mathbb{E}(y^2) < \infty$), that the M coordinates are independent, and thus that the probability density function $f_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ may be decomposed on a product of the marginal probability density functions $f_{\xi^{(i)}}$ (Eq. 3):

$$f(\boldsymbol{\xi}) = \prod_{i=1}^{M} f_{\xi_i}(\xi_i) \tag{3}$$

Given the natural inner product for arbitrary function ϕ with respect to each of the marginal probability function $f_{\xi}(\xi)$, we define an infinite set of mono-variate orthogonal polynomials $\boldsymbol{\varphi}(\xi) = \{\varphi_k(\xi), k \in \mathbb{N}\}$. Hermitian polynomials respect this condition for Gaussian random variables.

For other types of random variables, different orthogonal polynomials may be used leading to the generalized PCE or Wiener-Askey scheme [39].

Using the tensor product one may obtain an infinite set of multi-variate polynomials (with a preserved orthogonality property) $\boldsymbol{\psi} = \{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{N}^M\}$ where $\boldsymbol{\alpha} \in \mathbb{N}^M$ is a multi-index set.

According to the Cameron-Martin's theorem [40], the exact polynomial expansion of the functional y is

$$y(\boldsymbol{\xi}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^{\text{std}}). \tag{4}$$

where $\{\gamma_{\boldsymbol{\alpha}}\}$, $\boldsymbol{\alpha} \in \mathbb{N}^M$ are the coefficients of the PCE to be identified and $\boldsymbol{\xi}^{\text{std}}$ the Gaussian standardized counterpart of the physical random input $\boldsymbol{\xi}$. When solved by least square (Eq. 7) the transformation from physical to standardized random variable avoids system matrix to be ill-determined specially when there are model parameters of very different orders of magnitude. In the general case, where the components of the random vector $\boldsymbol{\xi}$ are independent and non Gaussian, the transformation to the

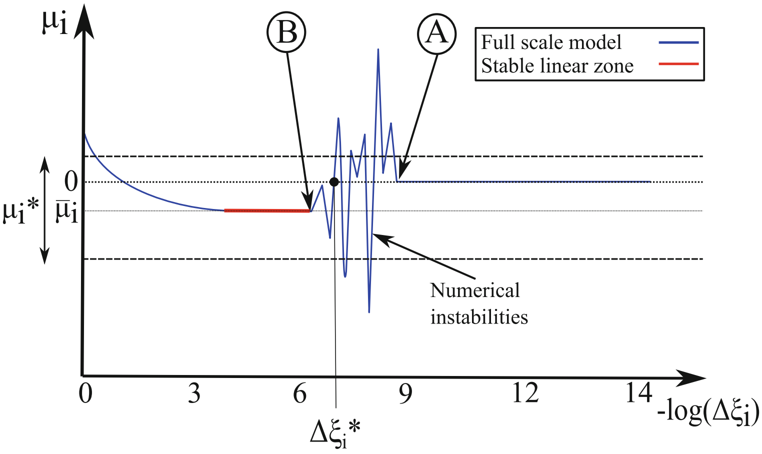


Fig. 1 Qualitative sensitivity results issued from actual computation

standardized, centered and independent variables requires the use of a non linear, e.g. Nataf [41] or Rosenblatt [42] transformation. In the general non Gaussian case, the transformations have to be constructed numerically.

In practice we have to truncate the full basis in order to only retain a *finite set* $\mathcal{A}$ of P polynomial terms

$$\tilde{y}^{\mathcal{A}} = \sum_{\alpha \in \mathcal{A}} \gamma_\alpha \Psi_\alpha(\xi^{\mathrm{std}}). \tag{5}$$

To compute the coefficients of the PCE, intrusive Galerkin type approach has been proposed [38]. Non intrusive projection based methods take advantage of the orthogonal properties of the multivariate polynomials of the expansion. Stochastic collocation is based on a Lagrangian interpolation in the stochastic space. It may be proved that this method is equivalent to the former [43].

In the regression based approach (on which we focus in this paper), the set of coefficients may be computed as

$$\boldsymbol{\gamma} = \operatorname{argmin}(\|\mathbf{y}(\boldsymbol{\xi}) - \boldsymbol{\Psi}(\boldsymbol{\xi}^{\mathrm{std}})\boldsymbol{\gamma}^\top\|^2) \tag{6}$$

solved by least squares:

$$\boldsymbol{\gamma} = (\boldsymbol{\Psi}(\boldsymbol{\xi}^{\mathrm{std}})\boldsymbol{\Psi}(\boldsymbol{\xi}^{\mathrm{std}})^\top)^{-1}\boldsymbol{\Psi}(\boldsymbol{\xi}^{\mathrm{std}})\mathbf{y}(\boldsymbol{\xi}), \tag{7}$$

where $\boldsymbol{y}(\boldsymbol{\xi})$ is the S-sized column vector containing the "high-fidelity" evaluations of the S samples and $\boldsymbol{\Psi}$ the $S \times P$ matrix containing the evaluation of the P polynomial terms for the S samples.

The optimal number of realizations needed to assess the coefficients with a good accuracy is still an open research area, but [24] proposes an empirical rule

$$S_{lb} = (M - 1) \times P. \tag{8}$$

We consider S_{lb} as a lower bound on the number of simulations ($S \geq S_{lb}$) to build a PCE.

Once the set of P coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ has been determined, one may compute the statistical moments of y analytically avoiding Monte Carlo simulations. The first two statistical moments are given by:

$$\mathbb{E}(y) = \gamma_0 \tag{9}$$

$$\sigma^2(y) = \sum_{\boldsymbol{\alpha} \in \mathcal{A} - \{0\}} \mathbb{E}(\Psi_{\boldsymbol{\alpha}}^2)\gamma_{\boldsymbol{\alpha}}^2 \tag{10}$$

4 Sampling Scheme Taking into Account Model Resolution

We propose to modify the classical Latin Hypercube Sampling (LHS) [37] scheme in order to take into account model resolution thresholds for each of the M variable separately. LHS randomly distributes samples in equiprobable bins in such a way that there is exactly one sample point per row in each of the M directions, eventually verifying contraints such as space-filling [44, 45], correlation [46], nested configurations [47], etc. The LHS advantages are [48]:

- as long as the number of samples S is large compared to the number of variables M, LHS eventually provides estimators with lower variances for any function with finite variance,
- in any case S-sized LHS does not perform worse than $(S-1)$-sized crude Monte Carlo.

However, LHS shows also the following limitations:

- The error estimates may not be improved by iteratively increasing the number of samples: the resulting sampling is not a LHS anymore (see [47, 49, 50] for Nested LHS).
- There is a risk that some of the random samples form a cluster to the detriment of some unexplored part of the design space. To circumvent these issues, re-sampling strategy [46], optimization algorithm [51], or geometrical consideration [44, 52, 53] has been proposed. The two first are straightforwardly compatible with our approach, but not investigated in the present paper.

When performing stochastic studies, small variations of the random input parameters may result in unstable output responses. To alleviate this limitation, we propose to build a restricted area (free of other sampling points) around each sampling point. The shape of this restricted area is parameterized by δ_i^* (Eq. 11) and may be defined as follows:

$$\delta_i^* = \underset{\Delta\xi^{(i)}}{\operatorname{argmin}}\, \mu_i(\Delta\xi^{(i)}) > \mu_i^*, \quad i \in \{1, \ldots, M\}. \tag{11}$$

Depending on the chosen norm ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_\infty, \ldots$), different shapes are obtained for the restricted area. The influence of the choice of the norm on the performance of the sampling scheme is still an open issue. However, due to the curse of dimensionality occuring in high dimension we preferentially use $\mathcal{L}_\infty$ norm. This approach coupled with LHS requirements permits us to filter the spacial sensitivity but limits the maximum number of samples available for PCE determination. In the remainder of the paper we denote this upper bound on the number of samples at hand by S_{ub}.

4.1 Implementation of the Fat-LHS

The goal of the **Algorithm** 1 is to identify the maximum number of sufficiently spaced sample points while preserving LHS criteria. In practice a high number of candidate of LHS are generated and the DoE with the best properties is kept.

The inputs of the algorithm are:

- the range of variation of M parameters,
- their associated probability density function, $p_i(\xi_i), i = 1, \dots, M$
- the resolution of the model associated to each parameter, $\Delta\xi_i^*, i = 1, \dots, M$,
- the number S of initial points. We recommend to choose S such as in each dimension the size of the narrowest LHS bin is at least one order of magnitude smaller than the resolution threshold for the considered parameter.

We denote the DoE by

$$\Xi = \left\{ \boldsymbol{\xi}_1 \ \boldsymbol{\xi}_2 \cdots \boldsymbol{\xi}_S \right\}$$

where $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S$ are M-dimensional column vectors. If the sampling is uniform, we denote $\Xi^{\mathcal{U}}$ and $\Xi^{\mathcal{P}}$ otherwise.

The pairwise distance matrix $\boldsymbol{D}$ is given by:

$$\boldsymbol{D}^k = \begin{pmatrix} 0 & \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| & \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_3\| & \cdots & \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_S\| \\ & \ddots & \|\boldsymbol{\xi}_2 - \boldsymbol{\xi}_3\| & \cdots & \|\boldsymbol{\xi}_2 - \boldsymbol{\xi}_S\| \\ & SYM & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots \\ & & & \|\boldsymbol{\xi}_S - \boldsymbol{\xi}_{S-1}\| & 0 \end{pmatrix} = \begin{pmatrix} 0 & d_{1,2} & d_{1,3} & \cdots & d_{1,S} \\ & \ddots & d_{2,3} & \cdots & d_{2,S} \\ & SYM & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots \\ & & & d_{S,S-1} & 0 \end{pmatrix} \tag{12}$$

where $\| \ . \ \|$ refers to the chosen norm. To remove the illegal neighbors, one has to identify for each sampling point $\boldsymbol{\xi}_{\text{current}}$ the number of illegal neighbors inside its restricted area. To do so we define:

$$\boldsymbol{D}_m^\top = \{\text{sign}(|\boldsymbol{\xi}_{\text{current}} - \boldsymbol{\xi}_1| - \boldsymbol{\mu}), \dots, \text{sign}(|\boldsymbol{\xi}_{\text{current}} - \boldsymbol{\xi}_S| - \boldsymbol{\mu})\} \tag{13}$$

If one observes that a particular line l^* of $\boldsymbol{D}_m(l^*, :) = \underbrace{[1, 1, \dots, 1]}_{\text{M times}}$ then the sampling point $\boldsymbol{\xi}_{l^*}$ is located inside the restricted area of $\boldsymbol{\xi}_{\text{current}}$. For each sampling point it is then possible to identify the number of illegal neighbors. We store this information into an $S \times S$ matrix denoted $\boldsymbol{N}$. Each line i and column j is set to 1 if $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are illegal neighbors and 0 otherwise. Then we sequentially remove the points with the greatest number of illegal neighbors and update $\boldsymbol{N}$ after each removal operation until $\boldsymbol{N}$ only contains 0 values.

To discard an illegal neighbor $\boldsymbol{\xi}_i$, we remove it from Ξ and denote the this operation by:

$$\Xi_{\{-i\}} = \left\{ \boldsymbol{\xi}_1 \; \boldsymbol{\xi}_2 \; \cdots \; \boldsymbol{\xi}_{i-1} \; \boldsymbol{\xi}_{i+1} \; \cdots \; \boldsymbol{\xi}_S \right\}.$$

Following steps are then performed within a loop:

- Step 1 is dedicated to the generation of a large number IT of different S-sized candidate LHS.
- Step 2 consists in the sequential removal of the points with the greatest number of illegal neighbors for each of the candidate DoE.

The algorithm may be easily implemented into a parallel environment and converges to an LHS containing the maximum number of points sufficiently spaced away. It provides two outputs:

- the maximum number S of "high-fidelity simulation" on which the PCE construction relies,
- and the associated S-sized design of experiment (Fig. 2).

Algorithm 1 Fat-LHS algorithm

```
Inputs
• $\boldsymbol{\Delta\xi}^{*}_{\boldsymbol{i}\;i}$, $i = 1, \ldots, M$ the given sensitivity for each parameter.
• $p_i(\xi_i)$, $i = 1, \ldots, M$ be their respective probability density functions.

• Set IT, the maximum number of LHS to be generated
• Set $M$, the number of stochastic variables
• Set $S$, the initial number of sampling points
• Choose a norm $\| \,.\, \|$ among $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_\infty$, ...

for $it$=$1$ to $IT$ do
  - Set $S^{it} = S$
  - Generate an $S^{it}$-sized and $M-$dimensional LHS denoted $\Xi^{\mathcal{U}}$ uniformly distributed over
  [0;1]
  - $\Xi^{\mathcal{U},it}$ is an $M \times S$ matrix containing the $M$-dimensional trial points as column vectors
  - Transform the $S$ uniform random vectors into the desired probabilistic space
  - $\Xi^{\mathcal{P},it} = T^{-1}(\Xi^{\mathcal{U},it})$
  - Compute $\boldsymbol{D}^{it}$ the $S^{it} \times S^{it}$ matrix containing the pairwise distance
  - Compute $\boldsymbol{N}^{it}$
  - Set $IN^{it}$ equals to the number of non zero lines of $\boldsymbol{N}$
  - Set $K^{it} = 0$
  while $IN^{it} > 0$ do
    - Identify using $\boldsymbol{N}^{it}$ the sampling points with the highest number of illegal neighbors
    - Remove one of them and update $\boldsymbol{N}^{it}$
    - Update the sampling size $S^{it} = S - 1$
  end while
end for
- $it^* = \underset{it}{\text{argmax}}\, S^{it}$

Outputs:
• Return $S^{it^*}$
• Save the LHS $\Xi^{\mathcal{P},\boldsymbol{it^*}}$
```

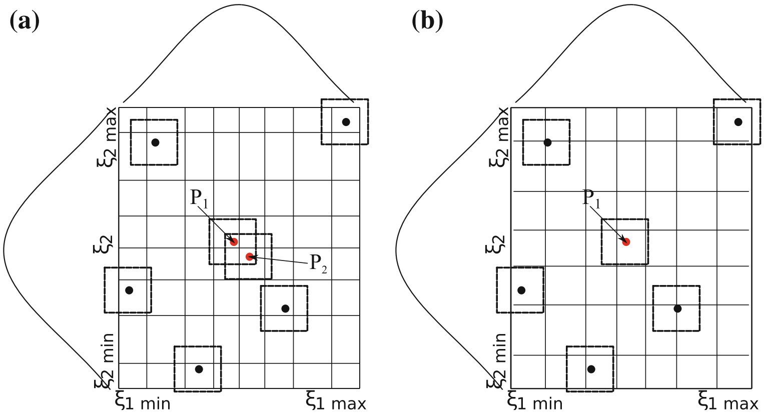


Fig. 2 Illustration of the fat latin hypercube sampling points removal procedure, **a** Two illegal neighbors with regards to the $\mathcal{L}_\infty$ norm, **b** Deleting the illegal neighbors and rearranging the bins size

5 Sparse PCE Models for Restricted Training Sets

The Fat-LHS (**Algorithm** 1) provides a set of spaced away sampling points consistent simultaneously with the distribution of the random parameters and the model resolution. The latter feature decreases the number of available simulations. Among the infinite set of PCEs terms one has to select the most relevant (correlated) with regards to the sampling at hand.

5.1 Truncating Multi-variate Polynomials Expansion

For Fat-LHS sampling scheme, we need an economical PCE scheme requiring less than $S \leq S_{ub}$ samples. The truncation schemes result in different lower bounds on sample sizes S_{lb}. In the following we implement a scheme satisfying $S_{lb} \leq S \leq S_{ub}$.

Among all $\{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{N}^M\}$ the truncation scheme [24, 38, 54] retains only the multi-variate polynomial terms whose degree does not exceed an a priori fixed N leading to the following multi-index set:

$$\mathcal{A}_q^M = \{\boldsymbol{\alpha} \in \mathbb{N}^M, ||\boldsymbol{\alpha}||_q \leq p\}, \tag{14}$$

where $||\boldsymbol{\alpha}||_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{1/q}$.

The truncated model may then be written as:

$$y^{\mathcal{A}_q^M}(\boldsymbol{\xi}) \approx \sum_{\boldsymbol{\alpha} \in \mathcal{A}_q^M} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^{\text{st}}). \tag{15}$$

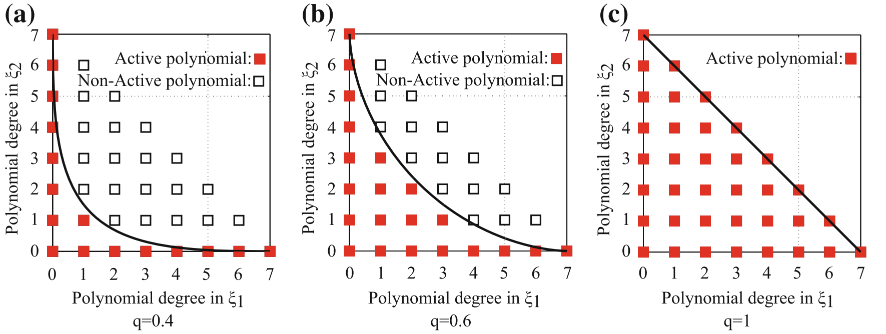


Fig. 3 Illustrations of Q-norm truncation with different values of the truncation parameter q for a 7th order PCE

For $q = 1$, the number P of coefficients in the PCE is given by

$$P = \sum_{k=0}^{N} C_{M+k+1}^{k} = \frac{(N+M)!}{N!M!} \tag{16}$$

and increases exponentially. So does the number of "high-fidelity" function evaluations needed to compute the number of PCE coefficients: whatever the method used, at least $S = P + 1$ samples are necessary.

Q-norm generalizes the classical truncation scheme by varying $0 \leq q \leq 1$ [55]. Figure 3 illustrates typical truncated index sets for different values of q on a 7th order bi-variate PCE.

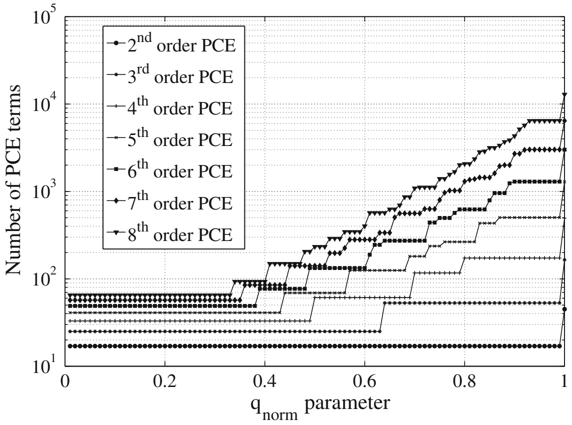


Fig. 4 Number of polynomials terms in truncated PCE with regards to the q truncation parameter for 8 variables

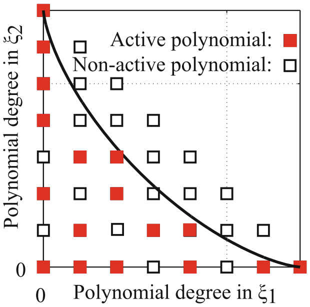


Fig. 5 An example of combined Q-norm + LARS truncation for a 7th PCE order, k = 17, q = 0.7

The set of active polynomials in the PCE decomposition is decreased when q decreases. Figure 4 illustrates the evolution of the number of 8-variate polynomial terms in log scale against q values. The rightmost value $q = 1$ corresponds to the classical truncation Eq. 16.

LARS truncation scheme. The Least Angle Regression Stagewise algorithm [55–57] issued from the variable selection community iteratively adds to the current model the polynomial terms which are the most correlated with the residual response.

Figure 5 illustrates truncated index obtained after $k = 17$ iterations applied to a 2-variate 7th order PCE previously .

At step k, k predictors have been added to the approximated model. We denote by $\mathcal{A}^M_{\text{LARS}}(k)$ the corresponding multi-index set whose cardinal is k.

5.2 Combining Q-norm and LARS

Considering the limited Fat-LHS sampling, we need to find the optimal sparse index set $\mathcal{A}^*$ such that the error produced on the resulting approximation model $y_{\mathcal{A}^*}$ is as low as possible. To perform this optimization task, we combine in an iterative manner the Q-norm and LARS truncations. We index by $\mathcal{A}^M_{\text{q}}$ the set of polynomials $\boldsymbol{\Psi}_{\text{q-truncated}}$ obtained by a Q-norm. From this set, one may apply LARS algorithm to the most correlated polynomial. We index by $\mathcal{A}^M_{\text{q+LARS}}$ the sparser set of polynomials $\boldsymbol{\Psi}_{\text{q+k}}$ obtained after a Q-norm and k steps using LARS.

Figure 5 illustrates this method by showing a sparse set of active polynomials obtained for $q = 0.6$ and $k = 8$.We note that $\text{card}(\mathcal{A}^M_{\text{q+k}}) \leq \text{card}(\mathcal{A}^M_{\text{q}})$.

We thus solve a combinatorial optimization problem [26]

$$\begin{cases} \underset{N,q,k}{\text{argmin}}\ Error(\mathcal{A}^M_{\text{q+k}}) \\ s.t.\ S_{\text{lb}} \leq S \leq S_{\text{ub}} \end{cases} \tag{17}$$

where *Error* is an estimator of the PCE quality that we describe in the Sect. 5.3, and S is the available number of samples (**Algorithm** 2).

Algorithm 2 Optimization of Q-norm + LARS parameters [26]

```
• Arbitrarily choose a set $\mathbf{N} = \{N_1, \dots, N_n\}$ of PCE orders, ($N_n$ possibly high).
• Store the response obtained using an Fat-LHS sampling in a vector $\boldsymbol{y} = [y_1, \dots, y_N]$
for $idx_n = 1 : n$ do
  • Choose a set of significant $q$ values
  $q \in \{q_{(i)} | i \in 1, \dots, Q\}$
  for $idx_q = 1 : Q$ do
    - Compute the $N^{\text{th}}_{idx_n}$ order full polynomial basis $\boldsymbol{\Psi}_{\text{full}}$.
    - Truncate the full polynomial basis using the $q_{(idx_q)}$ norm giving $\boldsymbol{\Psi}_{\text{q-truncated}}$
    - Let $P_{\text{remain}} = card(\mathcal{A}^M_{\text{q}})$ be the number of remaining polynomials after truncation
    - Perform a V-fold cross validation as follows, with K=2.
    for $idx_P = 1 : \min(P_{\text{remain}}, N)$ do
      - Divide the sampling in 2 populations of equal size $\boldsymbol{\xi}_{\text{test}}$ and $\boldsymbol{\xi}_{\text{verif}}$
      for $v = 1 : V$ do
        - Compute the LARS Algorithm on the $P_1$ population
        - Verify the results on the $P_2$ population by computing the chosen error estimate.
      end for
      - Retain the best LARS step $k^*$ according to the selected error criterion
    end for
  end for
end for
• Retain the best model with the best $N^*$, $q^*$, $k^*$ according to the selected error criterion
```

5.3 *Error Evaluation of the Polynomial Expansion*

According to Cameron Martin's theorem [40], when truncating the multi-index set, one may not reach the convergence to the exact solution in the $\mathcal{L}_2$ sense. We assess the results for the corrected Leave-One-Out (LOO) error estimate taking into account the overfitting phenomenon [55].

An estimation of the $\mathcal{L}_2$ error is given by the empirical formula:

$$\text{Err}_{\text{emp}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{y}(\boldsymbol{\xi}_i) - \mathbf{y}^{\mathcal{A}^M_{\text{q+k}}}(\boldsymbol{\xi}_i))^2. \tag{18}$$

However, Err_{emp} is known to under-estimate the $\mathcal{L}_2$ error. When increasing the complexity of the PCE, Err_{emp} systematically decreases, while the actual $\mathcal{L}_2$ error may eventually increase (overfitting phenomenon). By construction, the LOO error [26] is less sensitive to the overfitting. It relies on the computation of the predicted residual

$$\Delta^{(i)} = \mathbf{y}(\boldsymbol{\xi}_i) - \mathbf{y}_{-i}^{\mathcal{A}^M_{\text{q+k}}}(\boldsymbol{\xi}_i), \tag{19}$$

for each evaluation $\boldsymbol{\xi}_i$, $i \in \{1, \dots, N\}$, where $\mathbf{y}_{-i}^{\mathcal{A}^M_{\text{q+k}}}(\boldsymbol{\xi}_i)$ denotes the approximated model evaluated in $\boldsymbol{\xi}_i$ trained in $\Xi / \{\boldsymbol{\xi}_i\}$. The LOO error is then computed as

$$\mathrm{Err}_{\mathrm{LOO}} = \frac{1}{N}\sum_{i=1}^{N} \Delta^{(i)2}. \tag{20}$$

In the general case, the computation of the predicted residuals is a greedy process. In our case, these may be analytically computed

$$\Delta^{(i)} = \frac{\mathbf{y}(\boldsymbol{\xi}_i) - \mathbf{y}^{\mathcal{A}^M_{\mathrm{q+k}}}(\boldsymbol{\xi}_i)}{1 - h_i} \tag{21}$$

where h_i is the ith diagonal term of the $\boldsymbol{\Psi}^{\mathcal{A}^M_{\mathrm{q+k}}}\left({\boldsymbol{\Psi}^{\mathcal{A}^M_{\mathrm{q+k}}}}^{\top}\boldsymbol{\Psi}^{\mathcal{A}^M_{\mathrm{q+k}}}\right)^{-1}{\boldsymbol{\Psi}^{\mathcal{A}^M_{\mathrm{q+k}}}}^{\top}$ matrix.

Finally, we compute the absolute LOO as

$$\mathrm{Err}_{\mathrm{LOO}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbf{y}^{\mathcal{A}^M_{\mathrm{q+k}}}(\boldsymbol{\xi}_i) - \mathbf{y}^{\mathcal{A}^M_{\mathrm{q+k}}}(\boldsymbol{\xi}_i)}{1 - h_i}\right)^2. \tag{22}$$

and its relative counterpart

$$\epsilon_{\mathrm{LOO}} = \frac{\mathrm{Err}_{\mathrm{LOO}}}{\sigma(\mathbf{y})^2}. \tag{23}$$

We use ϵ_{LOO} to control the “goodness of fit” of PCEs.

6 Results and Discussions

6.1 Analytical Example

Let us consider f a $4th$ order Hermite-based polynomial function

$$f(\xi) = \xi^4 + \xi^3 - 5\,\xi^2 - 2\,\xi + 3 + \mathscr{N}. \tag{24}$$

We add to f a noise $\mathscr{N}$ built from three independent uniform random variables n_1, n_2, n_3 such that $\mathscr{N}$ has non zero mean, non unitary variance and has a random magnitude to retrieve a behavior as in Fig. 1

$$\mathscr{N} = n_1 * n_2 - n_3. \tag{25}$$

Illustrations of the smooth function, the additive noise and the resulting noisy function are respectively given in Fig. 6a–c.

Let us assume $\xi \in \mathcal{N}(0.75, 1)$. We limit ξ variation to the interval $[-3; 3]$ by truncating its Gaussian distribution. The computation of the sensitivity around the

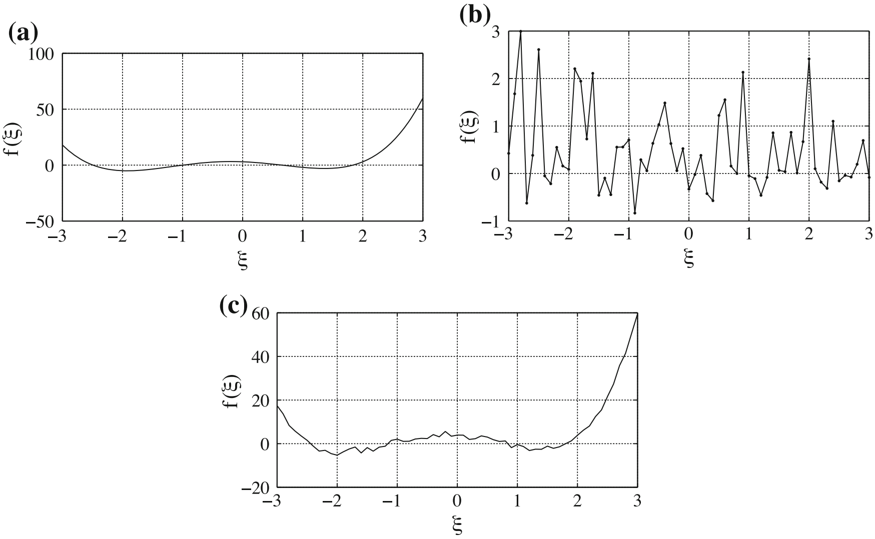


Fig. 6 Smooth function (**a**) noise (**b**) noisy function (**c**)

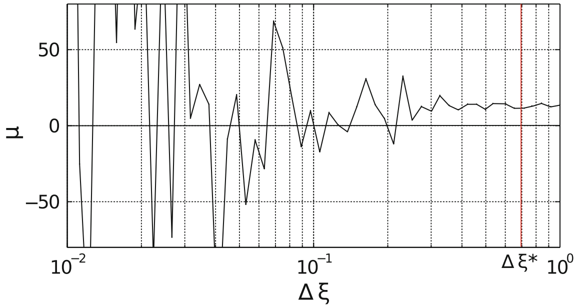


Fig. 7 Admissible resolution threshold

mean value has been done using a set of 151 points. Figure 7 provides the obtained variation of μ. We consider the variation limit threshold $\Delta\xi^* = 0.7$.

We realize two random samplings each containing 5 trials and both consistent with the Gaussian ξ distribution (Fig. 8). One of this set of random variables is obtained by repeated random sample generations until the distance between each pair of points is equal or greater than the model resolution threshold. This sampling is refereed as "Spaced random trials" in Fig. 8. The second set of sampling trials ("Closer random trials") is chosen such as the minimal distance between trial points is lower than $\Delta\xi^*$. PCE approximations are computed using the regression approach described in Sect. 3.

Figure 8 shows that the proposed Fat-LHS-PCE provides closer results to the original function than the LHS-PCE scheme.

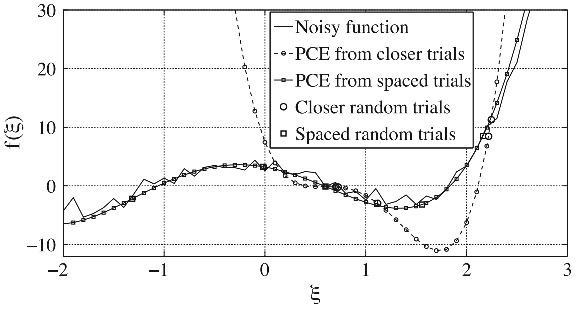


Fig. 8 Approximation results considering two different set of spaced and closer trial points

However, the PCE approximation does not allow to correct the approximation bias. Figure 9a–e plot γ_0, γ_1, γ_2, γ_3, γ_4 with respect to the number of sampling points. All figures show that the coefficients computed using the "spaced (random) trials " converge with a higher rate than those obtained using the "closer (random) trials" to a stable value.

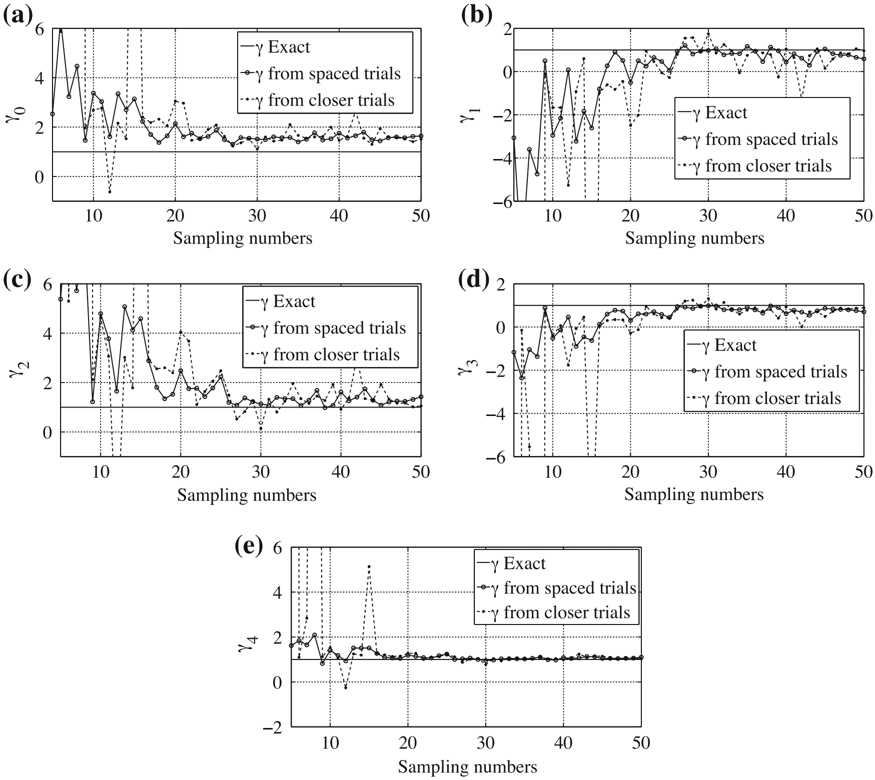


Fig. 9 Convergence results of the PCE coefficients for LHS-PCE and Fat-LHS-PCE scheme

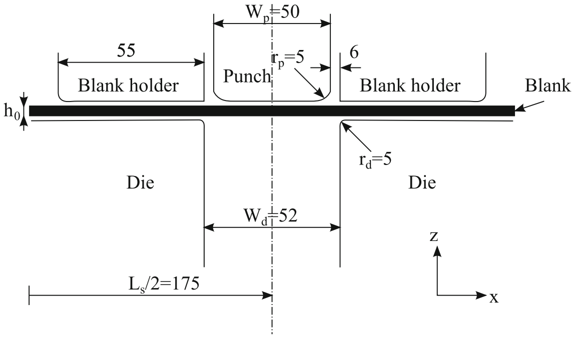


Fig. 10 Geometrical configuration of the modeled Numisheet'93 benchmark (values in mm)

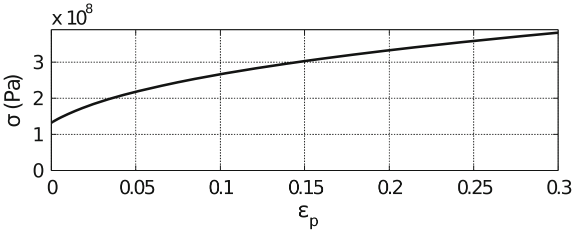


Fig. 11 Representation of the Swift hardening law for the parameters described in Table 1

6.2 2D Deep Drawing Process

We model the variability of the springback parameter of a 2D deep drawn U-shaped metal sheet from the B3 Numisheet'93 [58] benchmark.

The overall geometrical configuration of the deep drawing process is illustrated in Fig. 10.

6.2.1 Numerical Experiment Description

The process is modeled using a legacy software [59] using appropriate symmetry boundary conditions. A single row of 175 first-order shell elements is used to model the blank with Simpson integration rule using 10 integration points across the thickness. As the problem is essentially in plane strain state (the width of the blank is 35 mm and its thickness nominal value is 0.8 mm), corresponding boundary conditions are applied on each node (Fig. 11). The blank is made of mild steel modeled as an elastic-plastic material. Isotropic elasticity and the Swift isotropic hardening law are considered

$$\sigma = K_0(\epsilon_0 + \epsilon_p)^{n_0}. \tag{26}$$

Table 1 Nominal geometrical, material, loading and contact parameter of the U-shaped B-U-T model

Geometry	Material	Loading	Contact
L_s : 300 mm	E : 71 GPa	F_b : 300 N	μ : 0.15
h_0 : 0.81 mm	ν : 0.342	s : 60 mm	–
W_s : 1 mm	ρ : 2700 kg/m^3	–	–
r_p : 10 mm	K_0 : 576.79 MPa	–	–
W_d : 62 mm	ϵ_0 : 0.3593	–	–
r_d : 10 mm	n_0 : 0.01658	–	–

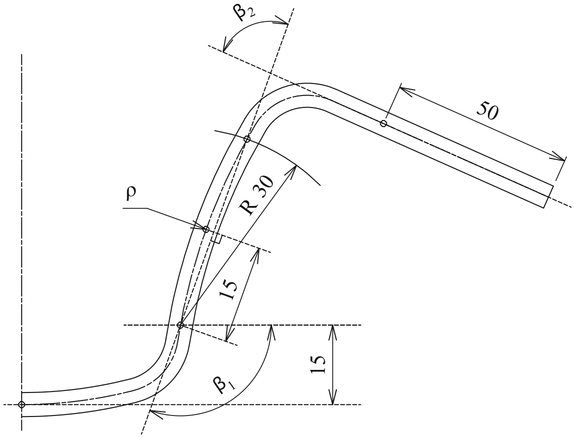


Fig. 12 Definition of springback parameters, ρ (mm), β_1 and β_2 (degree)

The value of the geometrical, material, loading and contact parameters are summarized in Table 1.

The tools (punch, blank holder and die) are modeled as rigid body surfaces. The punch velocity is taken here as 15 m/s and its displacement is $s = 70$ mm. The blank holder force is defined as $F_b = 2.45$ kN. The whole deep drawing process is simulated in two steps.

1. The forming phase is modeled using the explicit dynamic approach. The blank holder force is applied with a smooth ramp to minimize the inertia effect and the punch velocity follows a triangle step starting and ending with 0 velocity and reaching the 15 m/s at the half run. The contact occurring during forming phase is modeled using contact pairs.
2. The springback phase is modeled using an implicit approach. At the end at this phase, the springback shape parameters (output functions of interest): the curvature ρ (in mm), the angles β_1 and β_2 (in °) are measured (Fig. 12).

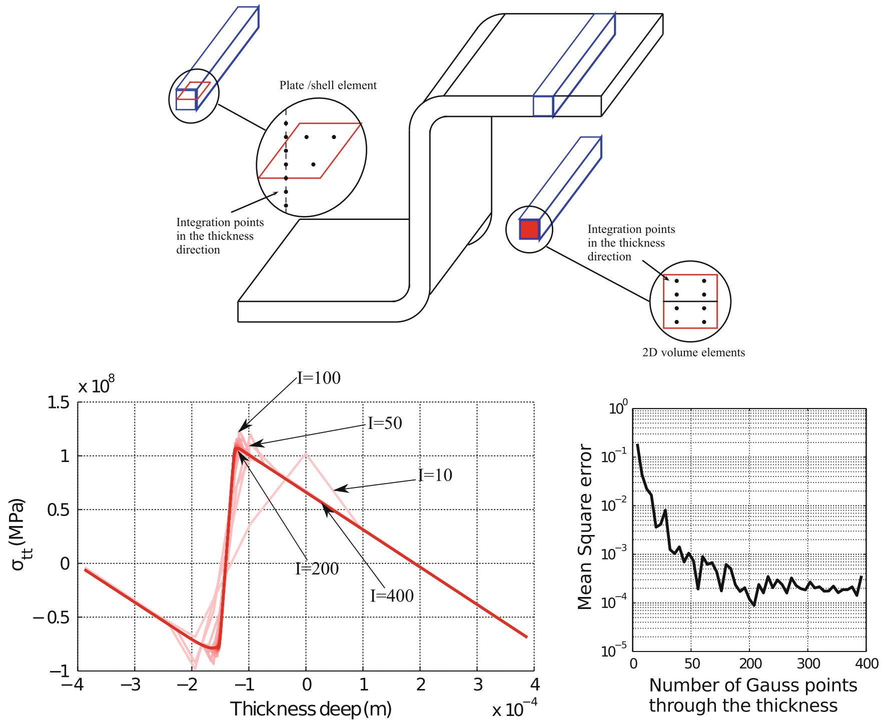


Fig. 13 Evolution of the stress across the section for different number of integration points (thickness 0.81 mm)

6.2.2 Numerical Instability

When the material undergoes plastic deformation, the stress profile becomes non-smooth due to the load path. We exhibit this phenomenon by modelling a section of the metal sheet submitted to a typical 2D deep drawing process undergoing bending-unbending loading path and focus on the in-plane stress σ_{tt} distribution across the thickness [34, 60]. When increasing the number of integration points across the cross-section profile, the in-plane stress σ_{tt} reaches numerical convergence (Fig. 13 (left)) when the number of integration points through the section is increased from 2 to around 200 (Fig. 13 (right)). The convergence is assessed using the following mean square error:

$$Err = \frac{\sum_{i=1}^{I}(\sigma_i - \sigma_i^{\mathrm{ref}})^2}{\sum_{i=1}^{I}(\sigma_i^{\mathrm{ref}})^2}. \tag{27}$$

Due to the non-smoothness of the stress profile for lower (realistic) number of integration points (here 10) leads to error in bending moment and numerical instabilities occur for small variation of parameters.

Table 2 Full stochastic model for the deep drawing process application

ξ	min	max	E(ξ)	$cv(\xi)$ (Coeff. of variation)
Thickness (h_0)	0.805	0.815	0.81 mm	$5/3 \times 10^{-3}$ mm
Young's modulus (E_b)	70.5	71.5	71 GPa	0.5/3 GPa
K_0	575.79	577.79	576.79 MPa	1/3 MPa
ϵ_0	0.3493	0.3693	0.3593	0.01/3
ν_0	0.01558	0.01758	0.01658	0.001/3
Poisson's coefficients (ν)	0.325	0.335	0.33	0.005/3
Friction coefficients (μ)	0.152	0.172	0.162	$0.01/3 \times 10^{-3}$
Clamp force (F)	34.5e3	35.5e3	35e3 kN	$0.5/3 \times 10^3$ kN

All random variables are assumed to be Gaussian and independent

6.2.3 Stochastic Model

Table 2 gives the set of independent random variables considered in the model. If the variation range of the parameters may be considered as realistic, the Gaussian hypothesis is only illustrative.

The mean values correspond to the nominal values, and the standard deviations are adjusted so that all the possible values lie in the 99, 7 % confidence interval: $\xi_{\text{min}} = \xi_{\text{mean}} - 3\sigma$ and $\xi_{\text{max}} = \xi_{\text{mean}} + 3\sigma$.

6.2.4 2D Validation Test Case

In this paragraph, we consider only 2 random variables: the blank thickness and the Young modulus. We start by investigations of sensitivity in Sect. 6.2.4. Then we apply the sampling methodology developed in Sect. 4 yielding $S_{ub} = 343$ samples respecting the sensitivity criterion. These samples have been obtained by applying the Fat-LHS algorithm using ranges of variation of both variables (thickness and Young modulus), their associated probability distribution and the observed model resolution associated with each of these parameters using $IT = 10,000$.

Sensitivity analysis. The sensitivity study is performed using the finite difference scheme described in Sect. 2 under small variations of the thickness (h_0) and of Young modulus (E) for the $1/\rho$ response. Figure 14 quantitatively illustrates the FEM numerical instability of simulation of the deep drawing process of 2-D U-shaped metal sheet introduced in Fig. 1. Responses are given in Fig. 13a, b, d, e and numerical sensitivities in Fig. 13c–f with respects to different orders of magnitude of thickness variation (Table 3).

The comparison of solid lines in Fig. 14c, f reveals local noisy behavior in the range of variation of the random parameters below the model resolution threshold (Table 4).

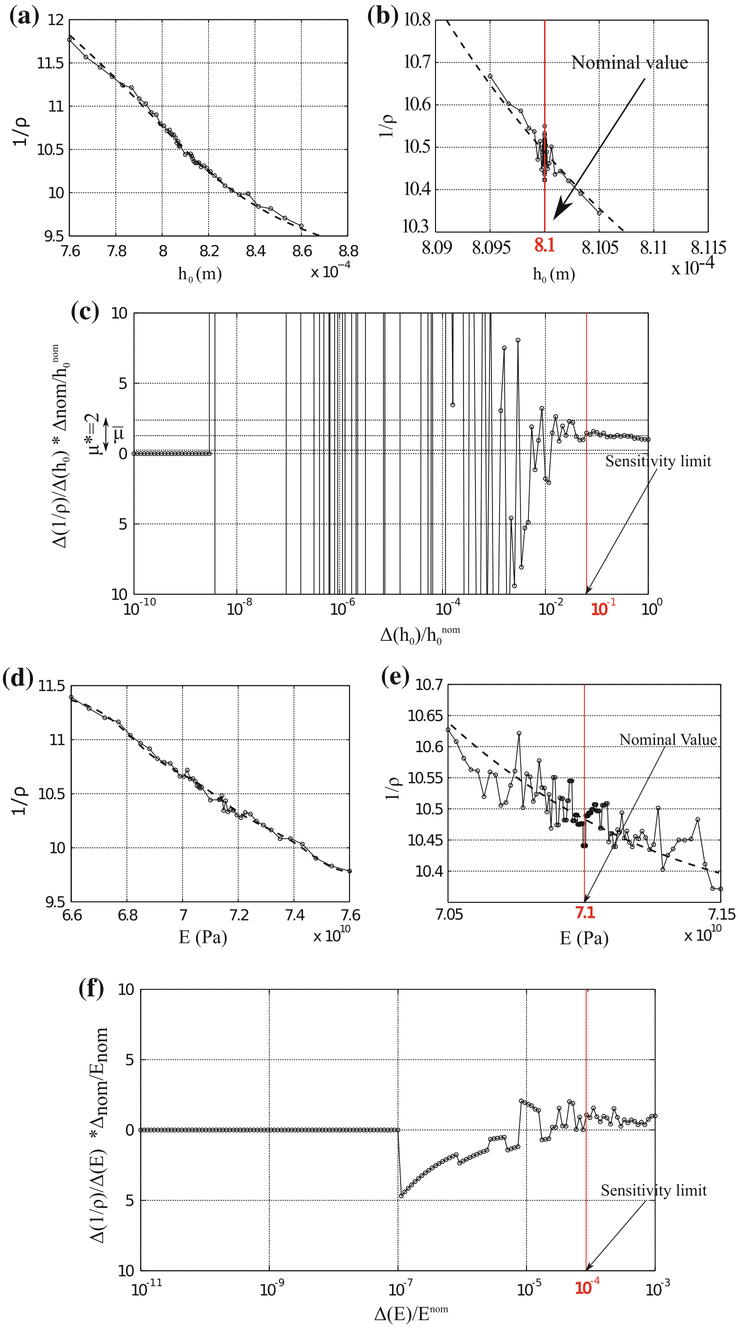


Fig. 14 Numerical instability for FEM simulations. **a** Global behavior. **b** Zoom around the nominal value

Table 3 Relative resolution threshold (order of magnitude)

Variables relative resolution	Responses		
	ρ	β_1	β_2
Thickness $(\Delta h_0^*/h_0^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
Young's Modulus $(\Delta E^*/E^{\text{nom}})$	1×10^{-4}	1×10^{-4}	1×10^{-4}

Table 4 Relative resolution threshold (order of magnitude)

Variables (relative)	Response resolution		
	ρ	β_1	β_2
Thickness $(\Delta h_0^*/h_0^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
Young's Modulus $(\Delta E^*/E^{\text{nom}})$	1×10^{-4}	1×10^{-4}	1×10^{-4}
K_0 $(\Delta K_0^*)/K_0^{\text{nom}})$	1×10^{-3}	1×10^{-3}	1×10^{-3}
ϵ_0 $(\Delta \epsilon_0^*/\epsilon^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
ν_0 $(\Delta \nu_0^*/\nu_0^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
Poisson's coefficient $(\Delta \nu^*/\nu^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
Friction coefficient $(\Delta \mu^*/\mu^{\text{nom}})$	1×10^{-1}	1×10^{-1}	1×10^{-1}
Clamp force $(\Delta F^*/F^{\text{nom}})$	1×10^{-2}	1×10^{-2}	1×10^{-2}

Illustration of Fat-LHS. Fig. 15a shows 343 LHS samples obtained without taking into account the model resolution (with illegal neighbors marked in red) and Fig. 15b shows the equivalent Fat-LHS pattern (without illegal neighbors.)

We compare then the two first statistical moments of $1/\rho$ for both samplings when increasing the number of points (β_1, β_2) exhibits similar trends (Fig. 16).

The Fig. 16 is obtained using the classical LHS and Fat-LHS. When the number of sampling becomes high, the classical LHS mean and standard deviation converges to a stable value produced by the Fat-LHS for a smaller number of sampled points.

In Fig. 14a–e solid lines are obtained without taking into account the model resolution. Dashed lines plot a bivariate 4th PCE is trained on sufficiently spaced points issued from the Fat-LHS sample (Fig. 15). We stress that the proposed approach allows to retrieve the underlying tendencies approximated by a relatively low 4th order PCE.

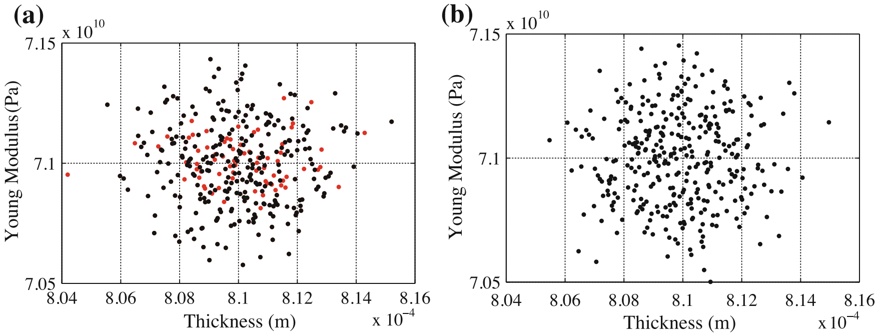


Fig. 15 Illustration of LHS obtained with taking (**a**) and not taking (**b**) into account the sensitivity constraint for 343 samplings. In *red* appears the illegal neighbors (color online)

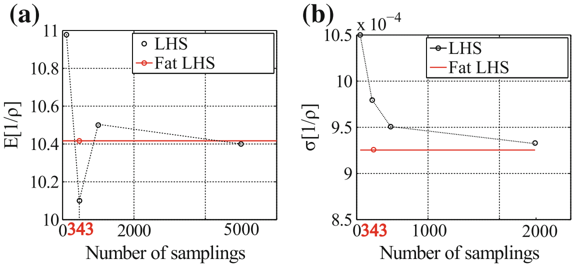


Fig. 16 Evolution of the mean (**a**) and of the standard deviation (**b**) of curvature against LHS/Fat-LHS size

6.2.5 8D Example

Sensitivity analysis of the full stochastic model. We now consider 8 random variables. These are reported in Table 4 which summarizes the relative response resolution threshold.

Comparison of PCE truncation schemes. We here compare the convergence results of the different truncation strategies for different 2-variate PCE of increasing order. We consider a limited number of 457 simulations issued from the Fat-LHS previously described.

We choose a polynomial order $N \in \{1, \ldots, n\}$ for the PCE approximation. For each PCE order N we apply 3 truncation strategies:

1. Select the Q-norm parameters such that $S_{lb} \leq S \leq S_{ub}$. For all the possible q parameters compute the approximate model $y_{\mathcal{A}_q^M}$, and retain the one with the lowest LOO error.
2. Use LARS on the classical truncation scheme and retain the best model.
3. Combine the Q-norm and LARS and retain the best model.

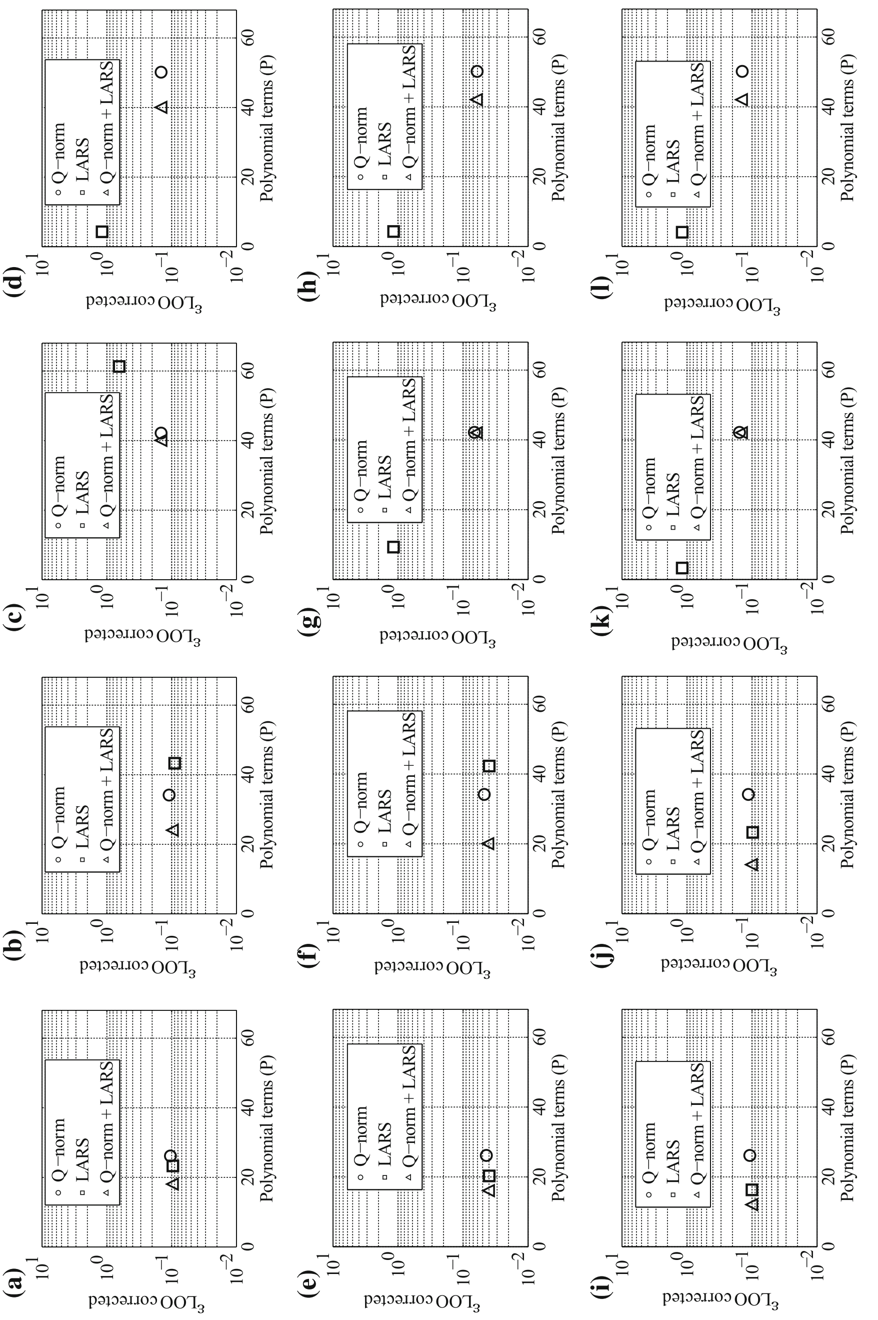


Fig. 17 Comparison of the evolution of the LOO error corrected for different polynomial degree with regards to springback parameter ρ (**a**), β_1 (**b**), β_2 (**c**) for random variables in Table 2. **a** ρ, $N = 3$, **b** ρ, $N = 4$, **c** ρ, $N = 5$, **d** ρ, $N = 6$, **e** β_1, $N = 3$, **f** β_1, $N = 4$, **g** β_1, $N = 5$, **h** β_1, $N = 6$, **i** β_2, $N = 3$, **j** β_2, $N = 4$, **k** β_2, $N = 5$, **l** β_2, $N = 6$

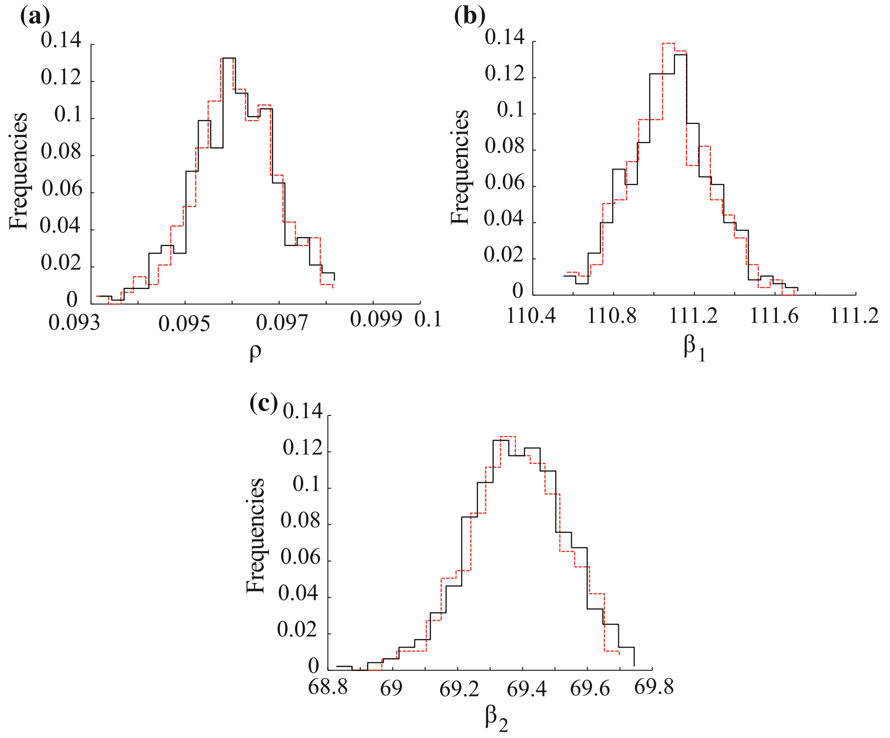


Fig. 18 Histograms obtained for the modified cross validation strategy for the 8-variate case using the $\epsilon_{\text{LOO corrected}}$. **a** ρ, $N = 3$, $\epsilon_{\text{LOO}} = 9.5e - 2$, $P_{\epsilon_{\text{LOO}}} = 17$,$q_{\epsilon_{\text{LOO}}} = 0.70$, **b** β_1, $N = 3$, $\epsilon_{\text{LOO}} = 4.2e - 2$, $P_{\epsilon_{\text{LOO}}} = 15$, $q_{\epsilon_{\text{LOO}}} = 0.70$, **c** β_2, N=3, $\epsilon_{\text{LOO}} = 8.5e - 2$, $P_{\epsilon_{\text{LOO}}} = 10$, $q_{\epsilon_{\text{LOO}}} = 0.5$

Figure 17 shows the evolution of the LOO corrected error with regards to the number of terms contained in the best PCE approximation. Each point refers to the best model obtained during the truncation for different PCE order. LARS alone provides the worst results, most of the time less accurate and more costly than the two other methods. Comparing the Q-norm and the combined LARS+Q-norm we observe similar results in terms of accuracy. However, a slight advantage may be given to the Combined LARS+Q-norm as it gives similar accuracy for a sparse PCE expansion. In addition, we note that due to smooth training data, we finally obtain the best results for truncated low order PCE.

The histograms (Fig. 18) illustrate the variability obtained for the best model obtained using the **Algorithm** 2. A good agreement is observed with the exact "high-fidelity" simulations as the relative error in mean is close to 0 and in standard deviation lower than 1 % (Table 5).

Table 5 Relative error in mean and standard deviation obtained for each response

	$\frac{\Delta E}{E_{\text{high-fidelity}}}$	$\frac{\Delta\sigma}{\sigma_{\text{high-fidelity}}}$
ρ	$\approx 10^{-9}$	3.7×10^{-3}
β_1	$\approx 10^{-7}$	4.5×10^{-3}
β_2	$\approx 10^{-12}$	8.5×10^{-3}

7 Conclusions and Prospects

We highlighted a fundamental limitation of the surrogate-based approach for uncertainty propagation. We showed that when using non linear FE “high-fidelity” simulations involving contact/friction and material non linearities, small variations of input parameters may lead to unstable training data sets and distort the statistical quantities of interest. We propose a sampling scheme named *Fat-LHS* allowing us to filter numerical instability. This heuristic strategy provides the maximum number of simulations available considering the finite model sensitivity. We then use this limited number of non-noisy samples to build a “best available” sparse PCE according to LOO error estimator. The comparison of LARS, Q-norm and LARS+Q-norm shows that generally the Q-norm+LARS hybrid is more efficient. Further work is required to

- economically compute the model sensitivity threshold,
- generate more space-filling LHS design (only the min-max strategy has been tested)

Finally, to assess the PCE accuracy, we use a LOO corrected PCE error in order to

- assess the goodness of fit of the PCE
- and to limit the overfitting phenomenon simultaneously.

It is difficult to separate which part of the inaccuracy comes from the model misspecification and which part comes from the overfitting phenomenon. A formulation of an overfitting measure for PCE approximation inspired from [61] may open a new way to efficiently select the most significant polynomials terms in a sparse PCE expansion.

Moreover, the use of structured sampling grid such as those provided by fully tensorized and sparse grid of Gaussian quadrature rules is usually limited to low order polynomial surrogate model with a limited number of random variables. In fact the associated computational cost increases exponentially with the number of variables and the polynomial order. In the 8-variate case studied, the results provided by the proposed iterative approach show that a low order polynomial provides the best results. A structured sampling grids may constitute an alternative approach in this particular case. However, in a more general way the proposed method may allow to build high order polynomial expansion with low order crossed polynomial terms (sparsity of effect principle) at lower computational cost than fully tensorized high-order quadrature rules.

Acknowledgments This research was conducted as part of the OASIS project, supported by OSEO within the contract FUI no. F1012003Z. This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). The first author is grateful to Innoviris for its support to BB2B project entitled multicriteria optimization with uncertainty quantification applied to the building industry.

References

1. G. G. Wang, S. Shan, Review of metamodeling techniques in support of engineering design optimization, Journal of Mechanical Design 129 (2007) 370.
2. T. W. Simpson, V. Toropov, V. Balabanov, F. A. Viana, Design and analysis of computer experiments in multidisciplinary design optimization: a review of how far we have come or not, in: 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 2008, pp. 10–12.
3. N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green, P. A. Newman, Approximation and model management in aerodynamic optimization with variable-fidelity models, Journal of Aircraft 38 (6) (2001) 1093–1101.
4. R. Ghanem, Ingredients for a general purpose stochastic finite elements implementation, Computer Methods in Applied Mechanics and Engineering 168 (1) (1999) 19–34.
5. R. Ghanem, Probabilistic characterization of transport in heterogeneous media, Computer Methods in Applied Mechanics and Engineering 158 (3) (1998) 199–220.
6. B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering & System Safety 93 (7) (2008) 964–979.
7. T. Crestaux, O. Le MaıÀÜtre, J.-M. Martinez, Polynomial chaos expansion for sensitivity analysis, Reliability Engineering & System Safety 94 (7) (2009) 1161–1172.
8. M. Eldred, Design under uncertainty employing stochastic expansion methods, International Journal for Uncertainty Quantification 1 (2) (2011) 119–146.
9. N. M. Alexandrov, R. M. Lewis, An overview of first-order model management for engineering optimization, Optimization and Engineering 2 (4) (2001) 413–430.
10. M. Eldred, A. Giunta, S. Collis, N. Alexandrov, R. Lewis, Second-order corrections for surrogate-based optimization with model hierarchies, in: Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, Aug, 2004.
11. T. de Souza, B. Rolfe, Multivariate modelling of variability in sheet metal forming, Journal of Material Processing Technology 203 (2008) 1–12.
12. N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green, P. A. Newman, Optimization with variable-fidelity models applied to wing design, Tech. rep., DTIC Document (1999).
13. R. Vitali, R. T. Haftka, B. V. Sankar, Multi-fidelity design of stiffened composite panel with a crack, Structural and Multidisciplinary Optimization 23 (5) (2002) 347–356.
14. G. Sun, G. Li, S. Zhou, W. Xu, X. Yang, Q. Li, Multi-fidelity optimization for sheet metal forming process, Structural and Multidisciplinary Optimization 44 (1) (2011) 111–124.
15. M. H. Bakr, J. W. Bandler, K. Madsen, J. E. Rayas-Sanchez, J. Sondergaard, Space-mapping optimization of microwave circuits exploiting surrogate models, Microwave Theory and Techniques, IEEE Transactions on 48 (12) (2000) 2297–2306.
16. M. Eldred, D. Dunlavy, Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models, in: Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, number AIAA-2006-7117, Portsmouth, VA, Vol. 199, 2006.
17. A. I. Forrester, A. Sóbester, A. J. Keane, Multi-fidelity optimization via surrogate modelling, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 463 (2088) (2007) 3251–3269.

18. L. Leifsson, S. Koziel, Multi-fidelity design optimization of transonic airfoils using physics-based surrogate modeling and shape-preserving response prediction, Journal of Computational Science 1 (2) (2010) 98–106.
19. C. Bucher, U. Bourgund, A fast and efficient response surface approach for structural reliability problems, Structural safety 7 (1) (1990) 57–66.
20. P. Breitkopf, A. Rassineux, P. Villon, An introduction to moving least squares meshfree methods, Revue européenne des éléments finis 11 (7–8) (2002) 826–867.
21. P. Breitkopf, H. Naceur, P. Villon, Moving least squares response surface approximation: Formulation and metal forming applications, Computers and Structures and Structures 83 (2005) 1411–1428.
22. V. Dubourg, B. Sudret, J. Bourinet, Reliability-based design optimization using kriging surrogates and subset simulation, Structural and Multidisciplinary Optimization 44 (5) (2011) 673–690.
23. R. G. Ghanem, R. M. Kruger, Numerical solution of spectral stochastic finite element systems, Computer Methods in Applied Mechanics and Engineering 129 (3) (1996) 289–303.
24. M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element : a non intrusive approach by regression, Revue Européenne de Mécanique Numérique 15 (2006) 81–92.
25. A. Doostan, R. G. Ghanem, J. Red-Horse, Stochastic model reduction for chaos representations, Computer Methods in Applied Mechanics and Engineering 196 (37) (2007) 3951–3966.
26. G. Blatman, B. Sudret, Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, Comptes Rendus Mécanique 336 (2008) 518–523.
27. P. G. Constantine, M. S. Eldred, E. T. Phipps, Sparse pseudospectral approximation method, Computer Methods in Applied Mechanics and Engineering 229 (2012) 1–12.
28. H. Ou, P. Wang, B. Lu, H. Long, Finite element modelling and optimization of net-shape metal forming processes with uncertainties, Computers and Structures 90–91 (2012) 13–27.
29. T. Jansson, L. Nilsson, R. Moshfegh, Reliability analysis of a sheet metal forming process using monte carlo analysis and metamodels, Journal of Material Processing Technology 202 (2008) 255–268.
30. W. Donglai, C. Zhenshan, C. Jun, Optimization and tolerance prediction of sheet metal forming process using response surface model, Computational Materials Science 42 (2007) 228–233.
31. S.-C. Kang, H.-M. Koh, J. F. Choo, An efficient response surface method using moving least squares approximation for structural reliability analysis, Probabilistic Engineering Mechanics 25 (4) (2010) 365–371.
32. C. J. Roy, W. L. Oberkampf, A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Computer Methods in Applied Mechanics and Engineering 200 (25) (2011) 2131–2144.
33. V. A. Yastrebov, Numerical Methods in Contact Mechanics, John Wiley & Sons, 2013.
34. T. Meinders, I. Burchitz, M. Bonte, R. Lingbeek, Numerical product design: Springback prediction, compensation and optimization, International Journal of Machine Tools and Manufacture 48 (5) (2008) 499–514.
35. F. Pukelsheim, Optimal design of experiments, Vol. 50, SIAM, 1993.
36. T. Hastie, R. Tibshirani, J. Friedman, Linear Methods for Regression, Springer, 2009.
37. M. McKay, W. Conover, R. Beckman, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (1979) 239–245.
38. R. Ghanem, Stochastic finite Elements: A spectral approach, Springer, 1991.
39. D. Xiu, G. E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24 (2) (2002) 619–644.
40. R. Cameron, W. Martin, Transformations of wiener integrals under translations, The annals of Mathematics 45 (2) (1944) 386–396.
41. A. Nataf, Détermination des distributions de probabilités dont les marges sont données, Comptes Rendus des l'Acadamie des Sciences 225 42–43.
42. M. Rosenblatt, Remarks on a multivariate transformation, The annals of mathematical statistics 23 (3) (1952) 470–472.

43. B. Sudret, Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods, habilitation à diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France (2007).
44. K. R. Dalbey, G. N. Karystinos, Generating a maximally spaced set of bins to fill for high-dimensional space-filling latin hypercube sampling, International Journal for Uncertainty Quantification 1 (3) (2011) 241–255.
45. F. A. Viana, G. Venter, V. Balabanov, An algorithm for fast optimal latin hypercube design of experiments, International journal for numerical methods in engineering 82 (2) (2010) 135–156.
46. G. D. Wyss, K. H. Jorgensen, A user's guide to lhs: Sandia's latin hypercube sampling software, SAND98-0210, Sandia National Laboratories, Albuquerque, NM.
47. P. Z. Qian, Nested latin hypercube designs, Biometrika 96 (4) (2009) 957–970.
48. M. Stein, Large sample properties of simulations using latin hypercube sampling, Technometrics 29 (2) (1987) 143–151.
49. P. Z. Qian, M. Ai, C. Wu, Construction of nested space-filling designs, The Annals of Statistics 37 (6A) (2009) 3616–3643.
50. P. Z. Qian, B. Tang, C. Jeff Wu, Nested space-filling designs for computer experiments with two levels of accuracy, Statistica Sinica 19 (1) (2009) 287.
51. S. J. Bates, J. Sienz, V. V. Toropov, Formulation of the optimal latin hypercube design of experiments using a permutation genetic algorithm, in: Proceedings of the 5th ASMO-UK/ISSMO Conference on Engineering Design Optimization, 2004.
52. K. R. Dalbey, G. N. Karystinos, Fast generation of spacefilling latin hypercube sample designs, in: Proc. of the 13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 2010.
53. P. Zhang, P. Breitkopf, C. Knopf-Lenoir, W. Zhang, Diffuse response surface model based on moving latin hypercube patterns for reliability-based design optimization of ultrahigh strength steel nc milling parameters, Structural and Multidisciplinary Optimization 44 (5) (2011) 613–628.
54. B. Sudret, A. Der Kiureghian, Stochastic finite elements methods and reliability- a state of the art, Tech. rep., Department of civil and environmental engineering (2000).
55. G. Blatman, B. Sudret, Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, Comptes-Rendus Mécanique 336 (2008) 518–523.
56. R. T. Hastie, J. Tibshirani, G. Friedman, The Elements of Statistical Learning, Data Mining, Inference and Prediction, Springer, 2009.
57. T. Hesterberg, N. Choi, L. Meier, C. Fraley, Least angle and l_1 penalized regression: A review, Statistics Surveys 2 (2008) 61–93.
58. A. Makinouchi, E. Nakamachi, E. Onate, R. Wagoner (Eds.), Numisheet'93 2nd International Conference, Numerical Simulatoin of 3-D Sheet Metal Forming Process - Verification of Simulation with Experiment., Riken Tokyo, 1993.
59. LSTC, LS-Dyna Keyword User's Manual, Livermore Software Technology Corporation (LSTC), 2009th Edition.
60. J. Lebon, G. Le Quilliec, P. Breitkopf, R. F. Coelho, P. Villon, A two-pronged approach for springback variability assessment using sparse polynomial chaos expansion and multi-level simulations, International Journal of Material Forming (2013) 1–13.
61. M. Bilger, W. G. Manning, Measuring overfitting and mispecification in nonlinear models, Tech. rep., HEDG, c/o Department of Economics, University of York (2011).

Multiscale Atomistic-to-Continuum Reduced Models for Micromechanical Systems

Eduard Marenić and Adnan Ibrahimbegovic

Abstract In this work we discuss the multiscale reduced models and computational strategy for micromechanical systems. The main focus is upon the interplay between the fine-scale atomistic model and the corresponding model at coarse-scale reduced by homogenization and placed within the continuum mechanics framework. Two mainstream multiscale methods, the quasi-continuum and bridging domain, are compared and brought to bear on the chosen model reduction strategy. Consequently, these two methods are further advanced from their standard formulation to a unified coupling and implementation strategy. An illustrative example of a defect-damaged grahene sheet is presented in order to confirm an excellent performance of such a multiscale solution strategy.

1 Introduction

An increased competition in consumer electronics has pushed the boundaries of technological development towards miniaturisation, with weight/size limitations and increasing power demands being the two most stringent requirements. Consequently, the mainstream scientific research in material science is currently shifting from micro- and meso-scale to the study of the behavior of materials at the atomistic or nano-scale. The main novelty with nano-scale is that the effects related to single atom, individual molecule, or nano-structural features (like lattice defects) may come to dominate the material behaviour. Once the occurring dimensions reach the submicron length scale, many interesting processes can no longer be described nor completely understood within the continuum mechanics modeling framework. For example, the computational modeling using the molecular simulation [1, 2] has been

E. Marenić (✉) · A. Ibrahimbegovic
Chaire de Mécanique, Lab. Roberval de Mécanique, Centre de Recherche Royallieu,
Sorbonne Universités / Université de Technologie Compiègne, 60203 Compiègne, France
e-mail: eduard.marenic@utc.fr

A. Ibrahimbegovic
e-mail: adnan.ibrahimbegovic@utc.fr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_9

used for dealing with materials failure and plasticity [3]. Besides these important studies of the nano-scale phenomena effect upon the irreversible deformation processes of bulk materials, another equally important reason pertains to study of the nano-scaled materials like graphene. Graphene represents a conceptually new class of materials that are only one atom thick [4]. Moreover, this is a perfect two-dimensional crystal exhibiting exceptionally high mechanical and electronic qualities. Advances in the synthesis of nanoscale materials have stimulated ever-broader research activities in science and engineering devoted entirely to these materials and their applications. This is mostly due to the combination of their expected structural perfection, small size, low density, high stiffness, high strength and excellent electronic properties [5]. As a result, nano-scale materials may find use in a wide range of applications from composite design, i.e., material reinforcement, nanoelectronics to sensors and medical diagnostic [6, 7]. These and other examples stem from different domains of application ranging from physics, biology, and chemistry to modern material sciences.

1.1 Models at Atomistic Scale

One of the most common tools used for the modeling of nano-materials is molecular dynamics (MD) (e.g. [3, 8–10]). MD is a common name for the computer simulation technique where the time evolution of a set of interacting atoms is determined by integrating their equations of motion. These equations are usually given in terms of the second Newton's law expressing the well known proportionality between force and acceleration. This way each atom is considered as a classical particle. Treating atomistic system using classical mechanics laws, and not by using Schrödinger equation and quantum mechanics is simplification related to the complexity of the Schrödinger equation and high dimension of the space in which the equation is posed. This simplification is based on the fact that the electron mass is much smaller than the mass of the nuclei. The idea is to split the Schrödinger equation, which describes the state of both the electrons and nuclei, with a separation approach into two coupled equations. The influence of the electrons on the interaction between the nuclei is then described by an effective potential. The latter is based on the simplification that restricts the whole electronic wave function to a single state, typically the ground state. This approximation is justified as long as the difference in energy between the ground state and the first excited state is everywhere large enough compared to the thermal energy (given as a product of Boltzman constant and absolute temperature $k_B T$) so that transitions to excited states do not play a significant role. The validity of this approximation is usually based on the de Broglie thermal wavelength (see [7, 10] and references therein) since the ground state is an eigenstate with the smallest energy level.

Concerning the approximation, the nuclei are moved according to the classical Newton's equations using either effective potentials which result from quantum mechanical computations (and include the effects of the electrons) or empirical potentials. The latter have been fitted to the results of quantum mechanical computations or to the results of experiments. Note that usage of the effective potential precludes the approximation errors to be rigorously controlled [7]. Moreover, quantum mechanical effects, and therefore chemical reactions are completely excluded. Nevertheless, the method has been proven successful, in particular in the computation of macroscopic properties (which is our concern in this work).

In this work we focus on the quasi-static problems, i.e., on the minimization of the potential energy of the system. Energy minimization corresponds to the physical situation of the system at absolute zero temperature. Methods in which the deformation behavior of the nano-structure is probed during continuous energy minimization is also referred to as molecular statics or molecular mechanics (MM). A variety of algorithms exist to perform energy minimization, most notably conjugate gradient methods or steepest descent methods [7]. However, in this work we will use Newton's incremental-iterative algorithm which is usually implemented as a solver in finite element codes, e.g. [11].

The main challenge related to fully atomistic simulations is that atomistic models typically contain extremely large number of particles, even though the actual physical dimension may be quite small. For instance, a crystal with dimensions below a few micrometers side-length has several tens of billions of atoms. Thus, in spite of the fact that micro-scale systems and processes are becoming more viable for engineering applications, our ability to model their performance is limited, and fully atomistic simulations remain out of reach for systems of practical interest. We thus assume that the calculation of specific quantities from the fine scale solution can be accurately approximated by replacing the particle model by a reduced model defined at coarser scale, within the sub-domains where the deformation remains sufficiently smooth. Thus, the idea is to use atomistic representation only in the localized region in which the position of each individual atom is important and to use reduced model, here continuum mechanics combined with the FE method, where the deformation is homogeneous and smooth.

1.2 Concurrent Atomistic-to-Continuum Methods

Multiscale modeling methods have recently emerged as the tool of choice to link the mechanical behavior of materials from the smallest scale of atoms to the scale of structures, thus being named atomistic-to-continuum multiscale (MS) methods. The approach where the fine scale model is processed simultaneously and directly coupled to the coarse scale, reduced model is usually called concurrent MS method. In order to reduce the computational cost, the molecular model must be limited to small cluster(s) of atoms in the vicinity of a domain of interest where high resolution models are necessary and a continuum method should be used for the rest of the domain.

Extensive work has been done in the development of atomistic-to-continuum MS modelling approaches, starting with early works by Mullins and Dokainish [12] and

Kohlhoff et al. [13]. Mullins simulated 2D cracks in B.C.C crystal in the context of a quasi static calculation with the atomistic scale models, and due to the restrictions of the computational power the question was how to connect the atomistic model and surrounding continuum. The basic idea is that the stresses are evaluated from the interatomic potential under the imposing strains stemming from the FE nodal displacements. Furthermore, these stresses are translated into nodal forces. Kohlhoff et al. proposed somewhat new method for combined FE and atomistic analysis of crystal defects, called FEAt. Here, an atomistic model is surrounded by a FE mesh with a small overlap region enforcing boundary condition on the atomistic as well as on the continuum domain. In particular, Kohlhoff et al. tried to overcome the capturing problem described in [12] by a refinement of the FE mesh down to the atomistic scale with nodal positions dictated by the crystal lattice structure. Both early works dealt with the problem of proper treatment of the transition between the lattice and continuum.

These early works initiated further development of a great number of MS methods, see e.g. some of the numerous reviews in [14–20]. Most frequently mentioned methods are: quasicontinuum (QC) method, bridging domain/Arlequin method (often abbreviated as BD or BD/A), concurrent coupling of length scales (CLS) [21], bridging scale (BS) method [22–24], coupled atomistics and discrete dislocations (CADD) [25], atomistic-to-continuum coupling (AtC) [26–28], macroscopic, atomistic, ab initio dynamics (MAAD) [29–31]. This list is by no means exhaustive. For instance, there is a recent effort of coupling non-local to local continuum, see [32, 33].

We focus in this work on the bridging domain (BD) method developed by Belytschko and Xiao [34]. The BD method is in essence a partially overlapping domain decomposition scheme used for atomistic-to-continuum coupling. The main idea is to divide the problem in the atomistic and continuum domains which partially overlap, and this overlap is called bridging domain. Moreover, there is a novel idea to draw attention towards a special role of adaptivity in providing an optimal form of the atomistic-to-continuum coupling based on the overlapping domain decomposition. For motivation, we consider the quasi-continuum (QC) method developed by Tadmor [35]. The QC method could be thought of as adaptive coarse graining approach and is used as a reference for adaptive strategy. Thus, BD and QC are described in detail, and compared.

This chapter is organized as follows.

Following this introduction, In Sect. 2 we give a brief insight in multiscale problems reaching from atomistic-to-continuum. The atomistic model problem is firstly introduced, following the formulation details of the two mainstream representatives QC and BD methods. In Sect. 3 a brief comparison of the two MS methods is given focusing of the adaptivity performance. Following this comparison, a possible unified formulation is given. A numerical example of the BD-based coupling with the adaptivity featuring goal oriented error estimates is shown in Sect. 4 on the defected graphene sheet. The concluding remarks are given in Sect. 5.

2 Problem Definition with Multiple Scales

2.1 Atomistic Model Problem

We focus in this work upon the MM neglecting both the dynamic effects and the thermal effects, used for quasi-static loading applications with the assumption of the zero Kelvin temperature. The equilibrium configuration corresponds to a state of minimum energy of the particle system, and we assume here that the initial configuration is at equilibrium. We consider a domain Ω^a in a 3-dimensional Euclidian space $\mathbb{R}^3$, which is occupied by N atoms placed within graphene microstructure. Let $\mathbf{R}_i$ and $\mathbf{r}_i$ denote, respectively, the position vectors in the reference and the current configurations of atom i, where $i = 1, \ldots, N$. The corresponding displacement vector of atom i is given by $\mathbf{d}_i = \mathbf{r}_i - \mathbf{R}_i$. Thus the displacement of the atoms is conveniently represented in compact form by means of vector $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N]$ from the space $\mathcal{V}^a = \{\mathbf{d} \in \mathbb{R}^{3\times N}\}$. The boundary conditions ought to be defined atom-wise, such that either the displacement $\bar{\mathbf{d}}_i$ or the external point force $\bar{\mathbf{f}}_i$ takes an imposed value. These conditions are imposed in quasi-static manner, with the corresponding incremental sequence.

As described in the Introduction, the nature of atomic interactions is governed by quantum effects taking place at the subatomic level and governing the chemical properties such as valence and bond energy [5, 7, 10]. Quantum mechanics-based description of atomic interaction is not discussed in this work, emphasis is rather on the empirical interaction models that can be derived as the result of such computations, i.e. from experimental observations. Classical potential is designed to account for the quantum effects in the average sense. Let $U(\mathbf{r}_i, \mathbf{r}_j^{el})$ denote the microscopic energy function that explicitly account for each atom i with coordinates $\mathbf{r}_i$, and each electronic degree of freedom $\mathbf{r}_j^{el}$. Then the classical potential (used in this work) pertain to the approximation which considers that the electronic degrees of freedom are completely removed, which can be written as

$$U(\mathbf{r}_i, \mathbf{r}_j^{el}) \to U_{approx}(\mathbf{r}_i). \tag{1}$$

Many different expressions $U(\mathbf{r}_i)$ can be fit to closely reproduce the energy predicted from quantum mechanics methods, while retaining computational efficiency [2, 3, 36]. There is no single, universal approach that is suitable for all materials and for all different phenomena of material behavior. The choice of the interatomic potential depends very strongly on both the particular application and the material. Thus, the heart of the MM model is the potential which governs the atomic interaction and it's choice is really important. The general structure of the potential energy function for a system of N atoms is

$$U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = \sum_{i}^{N} V_1(\mathbf{r}_i) + \sum_{i,j>i}^{N} V_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_{i,j>i,k>i}^{N} V_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \cdots, \tag{2}$$

where the function V_m, $\forall m = 1, 2, \ldots$ is the m-body potential depending on $\mathbf{r}_i$, the position vector of the atom i in current configuration. The first term on the right hand side of Eq. (2) indicates the effect of an external force field on the system where it is immersed, such as gravitational or electrostatic. This term is usually ignored in practice, [5]. The second term V_2 shows pair-wise interaction which is usually given in terms of the atom pair separation $r_{ij} = \|\mathbf{r}_j - \mathbf{r}_i\|$. Thus this term is usually denoted as V_{ij} or V_p. The best known examples of pair-wise potentials are Mie, Lennard-Jones, Morse [37] etc. The three-body term involves energy that characterizes angle-dependent forces, whereas four-body term includes torsion effects. m-body potential terms for $m > 2$ are usually called multi-body potentials. Apart from V_2, which depends on only one independent variable, each further term has $3m - 6$ variables.

The total energy E^a_{tot} of the atomic microstructure is given by

$$E^a_{tot} = U(\mathbf{d}_1, \ldots, \mathbf{d}_N) - \sum_i^N \bar{\mathbf{f}}_i \cdot \mathbf{d}_i, \tag{3}$$

where U denotes the energy stored in the atomic bonds, and the second term on the right-hand side represents the external energy E^{ext}. The state of equilibrium of the atomistic system corresponds to the minimum of the total energy which can be given in the weak form:

Find $\mathbf{d} \in \mathscr{V}^a$ such that

$$G^a(\mathbf{d}; \mathbf{w}) := \sum_i^N \frac{\partial E^a_{tot}}{\partial \mathbf{d}_i} \cdot \mathbf{w}_i - \sum_i^N \mathbf{f}_i \cdot \mathbf{w}_i = 0, \quad \forall \mathbf{w} \in \mathscr{V}^a_0. \tag{4}$$

In the above equation $\mathbf{w}_i$ represents the kinematically admissible virtual movement from the set of $\mathscr{V}^a_0 \subset \mathbb{R}^{3\times N}$, vanishing on the Dirichlet boundary. Linearising (4) and writing the result in matrix notation leads to

$$\mathbf{K}^{(k)} \Delta\mathbf{d}^{(k)} = \mathbf{F}^{(k)}, \tag{5}$$

where $\Delta\mathbf{d}^{(k)}$ is displacement increment corresponding to the k-th load increment, whereas $\mathbf{K}^{(k)}$ and $\mathbf{F}^{(k)}$ are the tangent stiffness and residual vector, respectively. The latter can be explicitly defined as

$$\mathbf{K}_{ij} = \frac{\partial^2 U}{\partial \mathbf{r}_i \partial \mathbf{r}_j}, \quad \mathbf{F}_i = \frac{\partial U}{\partial \mathbf{r}_i} - \bar{\mathbf{f}}_i. \tag{6}$$

An incremental-iterative solver is needed to solve system in (5) due to the nonlinear nature of the interatomic potential and geometrically nonlinear kinematics. For each load increment several Newton iterations are performed until convergence criteria are met in terms of energy test. At each iteration (k) the atomic positions are updated as follows

$$\mathbf{r}_i^{(k+1)} = \mathbf{r}_i^{(k)} + \Delta\mathbf{d}_i^{(k)}. \tag{7}$$

The initial iteration $(k) = 0$ starts at the initial configuration of the atomic system, with the position vector $\mathbf{r}_i^{(0)} = \mathbf{R}_i$. The procedure is terminated when the convergence is achieved for the last load increment.

2.2 QC and BD Formulations

In this section the formulation of QC and the BD/A methods are given, with only sufficient details. The goal is to show the evolution of the BD/Arlequin based coupling and to compare these methods regarding the ability to adapt.

2.2.1 Quasicontinuum Method

The Quasicontinuum (QC) method is originally proposed in late 90s by Tadmor et al. [35]. Since then it has seen a great deal of development and application by a number of researchers. The QC method has been used to study a variety of fundamental aspects of deformation in crystalline solids, including fracture [38–40], grain boundary slip and deformation [41]. The nano-indentation [42] and similar applications are examples where neither atomistic simulation nor continuum mechanics alone were appropriate, whereas the QC was able to effectively combine the advantages of both models. The main goal of the QC method is to provide a seamless link of the atomistic and continuum scales, and this coupling is further explained. The total energy of the coupled system consists of the energy of both domains.

> In QC the conceptual advantage in developing the coupled energy equation pertains to the fact that there is *no distinction between atoms and nodes*. This goal is achieved by the three main building blocks [43, 44]:
>
> 1. Reduction of degrees of freedom (DOF) by coarse-graining of fully atomistic resolution via kinematic constraints. The fully atomistic description is retained only in the regions of interest.
> 2. An approximation of the energy in the coarse grained region via numerical quadrature. The main idea is to avoid the need to calculate the energy of all the atoms, but retain only a few so-called rep-atoms.
> 3. Ability of the fully refined, atomistic region to evolve with deformation, where adaptivity is directed by suitable refinement indicator.

Model Reduction or Coarse Graining

If the deformation changes gradually on the atomistic scale, it is not necessary to explicitly track the displacement of every atom in the region. Instead it is sufficient to consider some selected atoms, often called representative atoms or rep-atoms. This process is in essence the model reduction via coarse graining. Only rep-atoms have independent DOF while all other atoms are forced to follow the interpolated motion of the rep-atoms. The QC incorporates such a scheme by means of the interpolation functions of the FE method, and thus the FE triangulation has to be performed with rep-atoms as FE mesh nodes. This way continuum assumption is implicitly introduced in QC method.

Let the total potential energy E_{tot} be given as a function of displacement $\mathbf{u}$ (similarly as in (3))

$$E_{tot}(\mathbf{u}) = U(\mathbf{u}) - \sum_{i=1}^{N} \bar{\mathbf{f}}_i \mathbf{u}_i, \tag{8}$$

where $\bar{\mathbf{f}}_i$ is the external force on the atom i and U is an atomistic internal energy, i.e. the energy stored in atomistic bonds. We assume further that the internal energy can be given as the sum of atom energies (E_i)

$$U = \sum_{i=1}^{N} E_i(\mathbf{u}). \tag{9}$$

Next, the kinematic constraint mentioned above is accomplished by replacing U with U^h

$$U^h = \sum_{i=1}^{N} E_i(\mathbf{u}^h), \tag{10}$$

where $\mathbf{u}^h$ is the approximated displacement field. The displacement approximation is given via standard FE interpolation

$$\mathbf{u}^h = \sum_{i=1}^{N_{rep}} \mathbf{N}_i \mathbf{u}_i, \tag{11}$$

where $\mathbf{N}_i$ is a shape function and $\mathbf{u}_i$ is the displacement for the node/rep-atom i. Clearly, the constraints introduced by the interpolation of the displacements is some level of approximation. The density of rep-atoms vary in space according to the considered problem. In the vicinity of region of interest every atom is considered as rep-atom (fully refined) and in region of more slowly varying deformation gradient, only a few atoms are chosen.

Efficient Energy Calculation via Cauchy-Born Rule, Local Approach

Described kinematic constraint on most of the atoms in the body will achieve the goal of reducing the number of degrees of freedom in the problem. However, for the purpose of energy minimization the energy of all the atoms (not just rep-atoms) has to be computed. The way to avoid visiting every atom is the Cauchy-Born (CB) rule [45–47]. The CB rule postulates that when a simple, mono-atomic crystal is subjected to small displacement on its boundary, all the atoms in the bulk will follow this displacement. In QC method this rule is implemented in that every atom in the region subjected to a uniform deformation gradient is taken to be energetically equivalent. Thus, energy within an element e can be estimated by computing the energy of one, single atom in the deformed state. The estimation is performed simply by multiplying the single atom energy by the number of atoms in the element e.

Let $\mathbf{F}$ be the deformation gradient and E_0 the energy of the unit cell when its lattice vectors are distorted according to the given deformation gradient. The strain energy density (SED) of the element can then be expressed as:

$$W(\mathbf{F}) = \frac{E_0(\mathbf{F})}{\Omega_0}, \tag{12}$$

where Ω_0 is the volume of the unit cell. Having this result in hand, the sum in Eq. (10) where $i = 1 \ldots N$ is reduced to number of FEs (N_{elem}) as

$$U^h \approx U^{h'} = \sum_{e=1}^{N_{elem}} \Omega_e W(\mathbf{F}_e). \tag{13}$$

In the above equation, the element volume and unit cell volume are related as $n_e \Omega_0 = \Omega_e$, and n_e is the number of atoms contained in element e. Using the CB rule, the QC can be thought of as a purely continuum formulation (local QC), but with a constitutive law that is based on atomistic model rather than on an assumed phenomenological form [44]. For a given deformation gradient $\mathbf{F}$ the lattice vectors in a unit cell are deformed according to given $\mathbf{F}$ and the SED is obtained according to Eq. (12). The main limitation pertaining to the CB rule is that it is valid only for simple lattices. In the original QC formulation the constant strain triangle (CST) elements (2D) are used with the linear shape functions to interpolate the displacement field within each element. In this case the deformation gradient is uniform. This boils down to the following: the Cauchy-Born rule assumes that a uniform deformation gradient at the macro-scale can be mapped directly to the same uniform deformation on the micro-scale. We will use this in sequel for the unified coupling formulation.

Non-local QC and Local/Non-local Coupling

In settings where the deformation is varying slowly and the FE size is adequate with respect to the variations of the deformation, the local QC is sufficiently accurate and very effective. In the non-local regions, which can be eventually refined to fully atomistic resolution, the energy in (10) can be calculated by explicitly computing the energy of the rep-atoms by numerical quadrature

$$U^h \approx U^{h'} = \sum_{i=1}^{N_{rep}} n_i E_i(\mathbf{u}^h), \tag{14}$$

where n_i is the weight for the rep-atom i. The value of the weight is high for rep-atoms in regions of low rep-atom density, and low for the region of the high density. Thus, n_i is the number of the atoms represented by the i-th rep-atom with the limiting case of $n_i = 1$ for fully atomistic region and consistency requirement

$$\sum_{i=1}^{N_{rep}} n_i = N. \tag{15}$$

The main advantage of the non-local QC is that when it is refined down to the atomistic scale, it reduces exactly to lattice statics, given in (3). High accuracy of non-local formulation can be combined with the high efficiency of the local formulation. The rep-atom can be chosen as local or non-local depending on its deformation environment giving $N_{rep} = N_{loc} + N_{nonloc}$. The total energy (10) is then approximated as

$$U^h = \sum_{i=1}^{N_{nonloc}} n_i E_i(\mathbf{u}^h) + \sum_{i=1}^{N_{loc}} n_i E_i(\mathbf{u}^h), \tag{16}$$

The above equation is yet another way of writing that the internal energy of the coupled system is a sum of atomistic (non-local) and continuum (local, here CB-based) energies, respectively. Regarding the calculation of the weights n_i in the above equation, for both local or non-local rep-atom, the Voronoi tessellation is used to create the cells around each rep-atom. Given that the cell of atom i contains n_i atoms, and n_i^e of these atoms reside in FE e adjacent to rep-atom i, the weighted energy contribution of rep-atom i is then found by applying the CB rule within each element adjacent to i such that

$$n_i E_i = \sum_{e}^{N_{el}^i} n_i \Omega_{0c} W(\mathbf{F}_e), \qquad n_i = \sum_{e}^{N_{el}^i} n_i^e, \tag{17}$$

where Ω_{0c} is the cell volume for single atom, and N_{el}^i is the number of FE adjacent to atom i.

Local/Non-local Criterion

The criterion to trigger the non-local treatment is based on the significant variation of deformation gradient. Precisely, we say that the state of deformation near a representative atom is nearly homogeneous if the deformation gradients that it senses from the different surrounding elements are nearly equal. The non-locality criterion is then:

$$\max_{a,b,k} |\lambda_k^a - \lambda_k^b| < \varepsilon_c, \tag{18}$$

where λ_k^a is the k-th eigenvalue of the right stretch tensor for element a, $k = 1 \ldots 3$ and indices a and b ($a \neq b$) refers to the neighboring elements of rep-atom. The rep-atom will be made local if this inequality is satisfied, and non-local otherwise, depending on the empirical constant ε_c.

Adaptivity

Without a priori knowledge of where the deformation field will require fine-scale resolution, it is necessary that the method should have a built-in, automatic way to adapt the finite element mesh through the addition or removal of rep-atoms. This is a feature that is in QC inherent from the FE literature, where considerable attention has been given to adaptive meshing techniques for many years, e.g. [48]. Typically in FE techniques, a scalar measure is defined to quantify the error introduced into the solution by the current density of nodes (or rep-atoms in the QC). Elements in which this error estimator is higher than some prescribed tolerance are targeted for adaptation, while at the same time the error estimator can be used to remove unnecessary nodes from the model.

The error estimator in terms of deformation gradient is defined as the difference between the actual solution and the estimate of the higher order (index '*ho*') solution (see [44])

$$\varepsilon_{\mathbf{F}}^e = \sqrt{\frac{1}{\Omega^e} \int_{\Omega^e} (\mathbf{F}^{ho} - \mathbf{F}^e)^2 d\Omega^e}, \tag{19}$$

where Ω^e is the volume of the element e, $\mathbf{F}^e$ is the solution for the deformation gradient in element e, and $\mathbf{F}^{ho} = \mathbf{N}\mathbf{F}_{avg}$ is the higher order estimate obtained by interpolating nodal values $\mathbf{F}_{avg}$, which simply represents the average of the deformation gradients of the elements sharing the given node. If this error is small, it implies that the higher order solution is well represented by the lower order elements in the region, and thus no refinement is required, while the elements for which the error is greater than the error tolerance are targeted for refinement. The refinement process advances by adding three new rep-atoms at the atomic sites closest to the mid-sides of the targeted elements (the constant strain triangle (CST) elements are used). If the nearest atomic sites to the mid-sides of the elements are the atoms at the element corners, the region is fully refined and no new rep-atoms can be added. The same error estimator is used in the QC to remove unnecessary rep-atoms from the mesh. In this process, a rep-atom is temporarily removed from the mesh and the surrounding region is locally re-meshed (i.e. nodal connectivity table is rebuilt). If all of the elements produced by this re-meshing process have a value of the error estimator below the threshold, the rep-atom can be eliminated. Essentially, the idea is to examine the necessity of each node. To prevent excessive coarsening of the mesh far from defects the nodes corresponding to the initial mesh are usually protected from deletion [41].

With these ideas in hand we turn to introduce the BD method. Note that initially emphasis of the research related to atomistic-to-continuum MS methods, namely BD method, was to make the coupling of the two different models as seamless as possible. No special attention was devoted to the question how to adaptively refine the model around the region of interest and where to position the coupling zone, i.e. how far from the region of interest. This issue is related to the adaptivity feature, and will be presented in sequel comparing the QC and BD methods.

2.2.2 Bridging Domain/Arlequin Method

The Bridging domain (BD) method is developed by Belytschko and Xiao in [34] for the static, and [49] for dynamical problems (see also more recent developments [50–52]). The compatibility in the overlapping domain is enforced by Lagrange multipliers. More precisely, the domain Ω is divided in three subdomains, atomistic, continuum and their overlap, as shown in Fig. 1. This overlapping region is also called handshake, bridging or coupling domain. The atomistic domain Ω^a is treated with MM, as described in Sect. 2.1, whereas the discretization in the continuum domain Ω^c is usually but not necessarily carried out by FEs. The atomistic and continuum domains overlap is denoted as $\Omega^b = \Omega^a \cap \Omega^c$. Before proceeding to BD governing equations and coupling, we will first recall the solution strategy related to the continuum part.

Continuum Solution Strategy

The role of the continuum mechanics formalism is to provide the reduced model replacing the molecular model with a coarser, and computationally much cheaper, model in $\Omega^c \subset \Omega$. The intention is to propagate only the large-scale information of the nanostructure, i.e. to be "compatibile" to the underlying lattice. Thus, the material parameters of the continuum constitutive model should be calibrated accordingly.

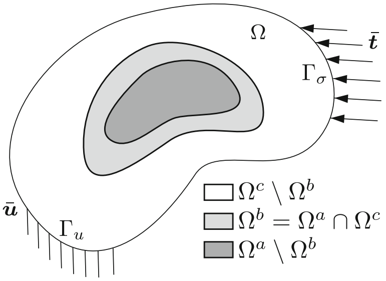

Ω
$\bar{t}$
Γ_σ
$\bar{u}$
Γ_u
□ $\Omega^c \setminus \Omega^b$
▨ $\Omega^b = \Omega^a \cap \Omega^c$
■ $\Omega^a \setminus \Omega^b$

Fig. 1 Scheme of the coupled model in BD method denoting the domain partitioning and overlap

This calibration is usually performed through numerical homogenization and virtual experiments on the RVE, see e.g. [53].

We consider a deformable solid body $\varOmega^c$ where the position of each material point is denoted with $\mathbf{X}$ in reference and with $\mathbf{x}$ in current configuration. The displacement vector is given as $\mathbf{u}(\mathbf{X}) = \mathbf{x} - \mathbf{X}$. We will consider further in the numerical example the geometrically linear theory of solid mechanics, assuming the hypothesis of small displacement gradients $\|\nabla\mathbf{u}(\mathbf{X})\| \ll 1$, which further allows us to use symmetric part of displacement gradient tensor as appropriate strain measure, $\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$. Let $W(\boldsymbol{\varepsilon}(\mathbf{X}), \mathbf{X})$ represent the continuum potential in terms of strain energy density (SED) which is given as

$$W = \frac{1}{2}\boldsymbol{\varepsilon}(\mathbf{X}) \cdot \mathbb{C}(\mathbf{X})\boldsymbol{\varepsilon}(\mathbf{X}), \tag{20}$$

with $\mathbb{C}$ being an elasticity tensor $\mathbb{C} = \frac{\partial^2 W(\cdot)}{\partial \boldsymbol{\varepsilon}^2}$ that is calibrated by homogenisation.

We now construct the weak form of the continuum boundary value problem in $\varOmega^c$, satisfying the equilibrium only in average sense. We assume that Dirichlet boundary conditions $\mathbf{u} = \bar{\mathbf{u}}$ are prescribed on the part $\varGamma_u$ of the boundary $\varGamma$. The nanostructure system represented as continuum is in general subjected to tractions $\bar{\mathbf{t}}$ on the part $\varGamma_\sigma$ of the boundary and to a volume forces $\mathbf{b}$ in $\varOmega$. We introduce the space of admissible solutions $\mathscr{V} = \{\mathbf{u} \in H^1(\varOmega); \mathbf{u} = \bar{\mathbf{u}} \text{ on } \varGamma_u\}$ and space of virtual displacement field $\mathscr{V}_0 = \{\mathbf{v} \in H^1(\varOmega); \mathbf{v} = \mathbf{0} \text{ on } \varGamma_u\}$, where each component v_i takes a zero value on the $\varGamma_u$ i.e. $\mathscr{V}_0 := \{v_i : \varOmega \mapsto \mathbb{R} \mid [v_i]_{\varGamma_{u_i}} = 0\}$. This choice of real and virtual displacement fields ensuring sufficient regularity ($\mathbf{u}, \mathbf{w} \in H^1(\varOmega)$) should also satisfy the weak form of equilibrium equation

$$0 = G(\mathbf{u}; \mathbf{v}) := \int_{\varOmega} \nabla^s\mathbf{v} \cdot \boldsymbol{\sigma}(\nabla^s\mathbf{u}) d\varOmega - \int_{\varOmega} \mathbf{v} \cdot \mathbf{b} d\varOmega - \int_{\varGamma_\sigma} \mathbf{v} \cdot \bar{\mathbf{t}} d\varGamma, \tag{21}$$

where $\nabla^s(\cdot) = sym[\nabla(\cdot)]$. Under the assumption of hyperelastic material with (20), the weak form in (21) is identical to the condition of the minimum of the total potential energy, given as

$$E^c_{tot} := \int_{\varOmega} W(\nabla^s\mathbf{u}) d\varOmega - \int_{\varOmega} \mathbf{u} \cdot \mathbf{b} d\varOmega - \int_{\varGamma_\sigma} \mathbf{u} \cdot \bar{\mathbf{t}} d\varGamma. \tag{22}$$

The weak form given in (21) is used as the basis for constructing the finite element approximation.

Governing Equations and Coupling

In QC method the total potential energy is composed of local and non-local parts (16), which correspond to continuum and atomistic description. Following the idea of the seamless transition, this approach somewhat hides the true coupling between the two descriptions. In BD method the system can be clearly decomposed into continuum and atomistic parts which are "glued" together in the coupling domain Ω^b. Thus, the total potential energy (with index w denoting that the energy term is weighted in Ω^b) of the system considering (3) and (22) may be written as

$$E_{tot,w} = E^a_{tot,w}(\mathbf{d}) + E^c_{tot,w}(\mathbf{u}) = w^a E^a_{tot}(\mathbf{d}) + w^c E^c_{tot}(\mathbf{u}), \tag{23}$$

where $\mathbf{d}$ and $\mathbf{u}$ are displacement vectors in the atomistic and continuum domains, respectively.

In the bridging domain the two models overlap, and the weighting functions w^a and w^c from (23) partition the energy. More precisely, weighting function serves to blend the behaviour from the continuum model (w^c) and the atomistic model (w^a) as well as to avoid the double counting of the energy in the bridging domain. The use of an overlapping subdomain *obviates the need for the FE nodes of the continuum model to correspond to the atomic positions*.

The weighting functions w^c and w^a define a *partition of unity* of the energy in the bridging domain as follows:

$$\begin{aligned} w^c(\mathbf{X}) &= 1 \quad \text{for } \mathbf{X} \in \Omega^c \setminus \Omega^b, \\ w^a(\mathbf{X}) &= 1 \quad \text{for } \mathbf{X} \in \Omega^a \setminus \Omega^b, \\ w^c(\mathbf{X}) + w^a(\mathbf{X}) &= 1 \quad \text{for } \mathbf{X} \in \Omega^b. \end{aligned} \tag{24}$$

The energy weighting functions are usually taken to be constant, linear (ramp) or cubic functions in Ω^b.

As mentioned, the Lagrange multiplier (LM) method is used to achieve the coupling, and to convert the problem of constrained minimization into finding the energy minimum of the larger, unconstrained problem. Thus, we introduce the space of LM as $\mathscr{M} = H^1(\Omega^b)$, and denote LM with $\boldsymbol{\lambda} \in \mathscr{M}$. In order to enforce the compatibility between the atomistic and continuum domains, the coupling term C in terms of energy is added to total energy forming so called Lagrangian

$$W_L := E_{tot,w} + C. \tag{25}$$

The choice of the coupling term determines which quantities and in which fashion should be coupled. Namely, we can choose whether only displacement or both the displacement and the displacement gradients are coupled. We will present two types: the strong (or discrete), and weak coupling. In the former, coupling of the atomistic and continuum models is achieved by enforcing (only) displacement compatibility in the bridging domain as $\mathbf{u}(\mathbf{X} = \mathbf{X}_i) = \mathbf{d}_i$, $\forall i \in \Omega^b$. The compatibility constraint between each atomistic displacement (discrete) and the continuum displacement field can be written as [34, 50]

$$C_1 := \sum_{i \in \Omega^b} \int_{\Omega^b} \boldsymbol{\lambda}(\mathbf{X}) \cdot [\mathbf{u}(\mathbf{X}) - \mathbf{d}_i]\, \delta(\mathbf{X} - \mathbf{X}_i) \mathrm{d}\Omega, \tag{26}$$

where $\delta(\cdot)$ is Dirac delta function. Note that the right hand side in the above equation is left in the integral form because the Lagrange multipliers is approximated as a field.

The evolution of BD method has much in common and parallels with recent works in the FE community on the coupling of nonconforming meshes in the overlapping subdomain, what is known as Arlequin method [54, 55]. In Arlequin method the coupling is given in the weak sense, and can be generalised as

$$C_2 := \int_{\Omega^b} \alpha_1 \boldsymbol{\lambda} \cdot (\mathbf{u} - \mathbf{d}^b) + \alpha_2 \nabla \boldsymbol{\lambda} (\nabla \mathbf{u} - \nabla \mathbf{d}^b) d\Omega, \tag{27}$$

where the choice of the weighting parameters α_1 and α_2 determines the coupling by mixing the displacement and strain coupling terms, and $\mathbf{d}^b(\mathbf{X})$ is the interpolated atomistic displacement field in Ω^b. The two versions of coupling, named L^2 and H^1, arise for the value of the weighting given $(\alpha_1, \alpha_2) = (1, 0)$, and $(\alpha_1, \alpha_2) = (1, 1)$, respectively. Note also, that the names L^2 and H^1, originate from the fact that they define the scalar products in Lebesgue (L^2) and Sobolev (H^1) spaces [54], respectively, and can be defined as

$$(\boldsymbol{\lambda}, \mathbf{u} - \mathbf{d}^b)_{L^2} := \int_{\Omega^b} \boldsymbol{\lambda} \cdot (\mathbf{u} - \mathbf{d}^b) d\Omega, \tag{28}$$

$$(\boldsymbol{\lambda}, \mathbf{u} - \mathbf{d}^b)_{H^1} := \int_{\Omega^b} \boldsymbol{\lambda} \cdot (\mathbf{u} - \mathbf{d}^b) + l^2 \nabla \boldsymbol{\lambda} (\nabla \mathbf{u} - \nabla \mathbf{d}^b) d\Omega, \tag{29}$$

where l is usually taken as the width of the bridging zone. An interpolated atomic displacement field is needed for this formulation of coupling, as well as its derivative. The interpolation of the discrete atom displacement is obtained by interpolant (Φ)

$$\mathbf{d}^b(\mathbf{X}) = \Phi \mathbf{d}_i, \qquad \forall i \in \Omega^b. \tag{30}$$

The first choice of Φ is naturally FE shape function, but the interpolant based on moving least squares approximation can also be used [50, 56].

Having these results in hand, we present next the weak form of the coupling problem. Denoting the space of LM with $\mathscr{M} = \{\boldsymbol{\lambda}, \boldsymbol{\mu} \in H^1(\Omega)\}$, we proceed to the minimising of the functional in (25) with the coupling term defined as (27), i.e. (28) or (29). This leads to the saddle point problem, which can be written in terms of its weak form:

Find $(\mathbf{u}, \mathbf{d}, \boldsymbol{\lambda}) \in \mathscr{V} \times \mathscr{V}^a \times \mathscr{M}$ such that

$$\begin{aligned} G^c_w(\mathbf{u}; \mathbf{v}) + G^a_w(\mathbf{d}; \mathbf{w}) + (\boldsymbol{\lambda}, \mathbf{v} - \Phi \mathbf{w}_{i|_{i\in\Omega^b}})_{L^2 \,\mathrm{or}\, H^1} &= 0 \qquad \forall(\mathbf{v}, \mathbf{w}) \in \mathscr{V}_0 \times \mathscr{V}_0^a, \\ (\boldsymbol{\mu}, \mathbf{u} - \mathbf{d}^b)_{L^2 \,\mathrm{or}\, H^1} &= 0 \qquad \forall \boldsymbol{\mu} \in \mathscr{M}, \end{aligned} \tag{31}$$

where G^c_w and G^a_w are scaled forms of (21) and (4), following the scaling given in (23).

The numerical implementation of the given coupling formulation with the coupling term C_2 (27) further resides to choice of the approximations fields for $\mathbf{u}$, $\mathbf{d}^b$ and LM field $\boldsymbol{\lambda}$.

Adaptivity and Error Estimate

Before proceeding with the numerical examples we will revisit the adaptive features related to the BD method. Apart from the advances in the coupling itself which is mostly related to the development of the Arlequin method advocated in initial work by Ben Dhia [54, 55] and its further application to the atomistic-to-continuum coupling [56–62], this method is acquiring the ability to accommodate the model and decrease the error in chosen quantity of interest. That is, the adaptivity described above for the QC method was included in the BD/Arlequin. This evolution parallels recent development in goal oriented error estimate theory as discussed in sequel.

In computer simulations of physical models we encounter usually approximation error due to the discretization, and modeling error related to the model simplification or in general to the natural imperfections in abstract models of actual physical phenomena. We focus here on the estimation and control of modeling error.

This subject has been introduced in recent years and was initially devoted to estimating global modeling error e.g. [63]. Since then, extensions to a posteriori error estimates in specific quantities of interest (QOI) have been proposed [64–66], with the idea to estimate upper and lower bounds of error in linear functionals. In [67] the error estimates are related to the error between discrete models (lattice) and homogenized model. Finally, the developments regarding the goal oriented error estimates, were employed in the coupling of atomic and continuum models. The difficulty in the use of such coupling methods is to decide where to locate the overlap region between the two models so as to control the accuracy of the solution with respect to the fully atomistic model. Initial convergence studies from [58] and later in [60, 62] are the basis for the development of the adaptive strategy in the Arlequin based coupled atomistic-to-continuum modeling. Refinement is related to the decrease of the modeling error in each iteration by locally enriching the surrogate model, i.e. by locally switching on the atomic model in the subregions where the reduced model is not accurate enough.

3 Comparison and Unified Formulation with Reduced Model

In this section, we briefly discuss how to achieve the appropriate combination of both atomistic-to-continuum multiscale method and successfully construct a reduced model to enhance the computational efficiency with no essential sacrifice of the computed results accuracy. QC method is in essence an adaptive FE approach, and adaptivity is intrinsically in the formulation. BD/A method, on the other hand, was initially assumed as approach to couple two different models. Nevertheless, the described evolution associated with the goal oriented error estimate theory, with the strong mathematical foundations, improved the method so that it shows good performance in the sense of model adaptivity. However, the choice where to place the atomistic and where to remain with reduced model, and how to provide the appropriate evolution of that region is still among the most important open questions. The idea of model adaptivity is shown schematically in the Fig. 2 for the 1D truss-like case. In this scheme we suppose that the strain field is perturbed in the left end, and the adaptive procedure advances from some initial model shown on the top.

For QC approach this procedure looks similar to a mesh refinement, however, the main goal is to address the possibility of model adaptivity in terms of substitution of the reduced model, based upon continuum mechanics, instead of the atomic one. As described in QC section, adapting process in this method advances by selecting new atoms as rep-atoms/nodes in the area where deformation gradient changes severely. Note that the continuum model, represented with a line Fig. 2 is present in the whole domain.

In the BD/A-based method, on the other hand, adaptive process concerns the switch from continuum to atomistic model cell by cell (see Fig. 2 on the right), in order to deliver accurate results regarding the selected QOI. Note that the overlap region has to be reconfigured, but the continuum and atomistic domains are separated except in the overlap.

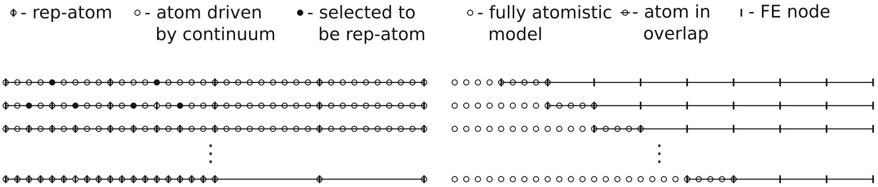


Fig. 2 Scheme of the adaptive procedure for the QC (*left*) and BD (*right*) method in 1D setting. A perturbation in the strain field which initiates the model adaptation is assumed on the left end of the truss-like model

3.1 Unified Coupling Formulation

The idea about model switch introduced above will be further exploited for the proposition of unified formulation in the step of the incremental loading sequence. Let us introduce the pseudo-time parameter denoted as t, as is customary in the incremental analysis. The choice of the load increments in a given load program is handled through increments in $t \in [0, T]$ according to $[0, T] = \bigcup_{n=1}^{n_{\mathrm{inc}}}[t_n, t_{n+1}]$. What we would like to point out is the similarity between the BD coupling and the adaptive, coarse graining procedure performed in the typical step of the incremental analysis between t_n and t_{n+1} in QC method, as schematically depicted in the Fig. 3. It is not directly obvious that the Cauchy-Born (CB) rule as the main ingredient of the QC method can be regarded as *homogenisation approach*, and as a *kinematic constraint* where the continuum is imposing the displacement gradient to the atoms. In the time step t_n we check for the error estimator $\varepsilon^e_{\mathbf{F}}$, and if the adaptation criteria is met we change the model as described in the QC method. Considering that the deformation gradient is related to the displacement gradient as $\mathbf{F} = \mathbf{I} + \nabla\mathbf{u}$ allows us to formulate the model change as the following coupling term

$$(\bar{\boldsymbol{\lambda}}, \mathbf{u} - \mathbf{d}^b)_{QC} = \int_{\Omega^e \in \Omega^b} \bar{\boldsymbol{\lambda}} \left(\nabla\mathbf{u} - \nabla\mathbf{d}^b\right) d\Omega, \tag{32}$$

similarly like in BD method with (28) for the case $(\alpha_1, \alpha_2) = (0, 1)$. In the above equation $\bar{\boldsymbol{\lambda}}$ is the LM field to impose the constraint, and $\mathbf{d}^b$ is the interpolated displacement of the atoms in the element e (where $\Omega^e \in \Omega^b$) which is being adapted and in which we want to achieve the match of the displacement gradients of the two domains (Ω^a and Ω^c). Next, selecting the LM mesh to correspond the lattice $\bar{\boldsymbol{\lambda}} = \delta(\mathbf{X} - \mathbf{X}_i)\bar{\boldsymbol{\lambda}}_i, \ \forall i \in \Omega^b$ gives

$$(\bar{\boldsymbol{\lambda}}, \mathbf{u} - \mathbf{d}^b)_{QC} = \int_{\Omega^e \in \Omega^b} \delta(\mathbf{X} - \mathbf{X}_i)\bar{\boldsymbol{\lambda}}_i \left(\nabla\mathbf{u} - \nabla\mathbf{d}^b\right) d\Omega, \tag{33}$$

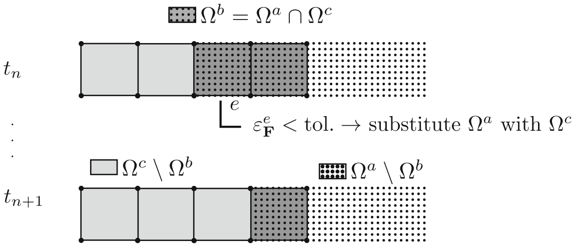

$\Omega^b = \Omega^a \cap \Omega^c$
t_n
e
$\varepsilon^e_{\mathbf{F}}$ < tol. → substitute Ω^a with Ω^c
$\Omega^c \setminus \Omega^b$
$\Omega^a \setminus \Omega^b$
t_{n+1}

Fig. 3 Converting atomistic to continuum in the solution step of the incremental analysis between t_n and t_{n+1}. The bridging domain Ω^b is where we perform model switch (following the logic from BD method) by formally imposing deformation gradient coupling (following the strategy from QC method)

which boils down to the strong form of coupling of deformation gradients of the two displacement fields

$$\bar{\boldsymbol{\lambda}}_i \left(\nabla \mathbf{u} - \nabla \mathbf{d}^b\right) = 0. \tag{34}$$

Needless to say, this kind of coupling naturally leads to the extremely expensive model, adding one unknown variable for every atomistic degree of freedom, and is not performed this way in practice neither in QC method nor in other MS methods. For instance in [68, 69] or [11, 70] similar coupling as (34) is implemented. This coupling is rather implemented by a priori taking the inherent property of the selected finite element. For instance, taking linear displacement distribution on the edges of the 4-node quadrilateral elements ($\mathbf{u}^e$) (as schematically shown on Fig. 3), one can form a direct transformation matrix T_{ij} which maps kinematic constraints to underlying atoms giving

$$d_{j|_{\Omega^e}} = T_{ij} u_i^e, \qquad \forall e \in \Omega^b. \tag{35}$$

Presented unified interpretation of the coupling gives a new look that allows to conclude the following. For BD method the inspiration from the QC-based adaptive strategy shows that overlapping zone Ω^b can and should move from step to step ($t_n \rightarrow t_{n+1}$). Furthermore, the choice of the LM field as $\bar{\boldsymbol{\lambda}} \sim \delta(\cdot)$ the 'direct' solution should be obtained by enforcing the constraint explicitly (see [68]) and not by using additional unknowns. Regarding the BD-based coupling in the context of the QC method shows that it is possible to couple Ω^a and Ω^c not only for $\mathbf{F} = cst.$ but also for non-homogeneous deformation.

4 Numerical Example

Presented numerical example is concerning a 2D case, namely the graphene sheet with the crack-like defect. Note that this example resembles the well known example of the through-thickness crack in an plate from linear elastic fracture mechanics, making it simple enough to have the theoretical, closed form solution, and at the same time, complex enough to present the performance of the presented MS method. Moreover, this example considers the problem of large practical interest related to two-dimensional crystal named graphene [71, 72]. Graphene is a new class of nanomaterial with remarkable properties whose immense potential for applications is driven by an intense current research. In the following example we are showing the performance of the MS strategy based on BD method presented above (and implemented in the in-house MATLAB code). The atomistic part of the model is used to properly capture the heterogeneous strain field produced by the defect, whereas the continuum is used for the part where the strain field is close enough to homogeneous state. We will be using the fully atomistic model treated with MM as the reference model for comparison. We note in passing that for the real problems this reference model is unavailable. On Fig. 4 both models the reference, and the coupled are shown, in the undeformed configuration. The former consists of 10,960 atoms, while the

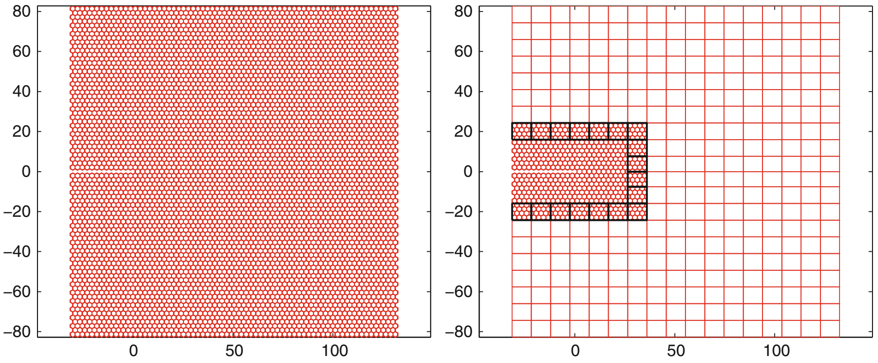


Fig. 4 Graphene sheet with a hypothetical initial crack modelled using the fully atomistic model (*left*) consisting of 10,960 atoms and coupled model (*right*) with the size of atomistic domain 67.4 × 48.7 Å

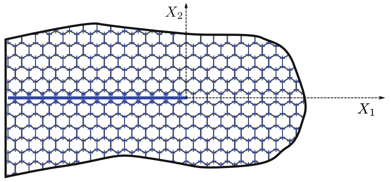


Fig. 5 A detail of the rectangular graphene sheet near the *left edge*. The atomistic model Ω^a is represented with the pair bonds between the neighbouring carbon atoms forming the honeycomb structure. The bonds parallel with the X_2 direction between atoms denoted with ($*$) are removed along the *blue line* in order to model the crack-like defect (Color Online)

latter provides considerable saving with 2080 atoms with the size of atomistic domain 67.4 × 48.7 Å. On the lattice level, crack-like defect is modelled simply by removing a line of bonds parallel with the X_2 direction, see Fig. 5. This configuration leads to the introduction of two free edges which stop at the single bond being at the crack tip. The continuum mesh M^c is represented with the red squares on the right plot in Fig. 4. The thick lines around Ω^a denotes mesh used for LM interpolation, where M^λ coincides with the FE mesh M^c used for the interpolation of continuum displacement $\mathbf{u}$. Young's modulus (E) and Poisson's ratio (ν) used to describe the linear elastic behaviour of continuum model have been determined by means of virtual experiments performed on the atomistic lattice [53].

On the edges of rectangular domain $\Omega = \Omega^a \cup \Omega^c$ the displacement boundary conditions are imposed. They correspond to mode I (K_I), near-tip displacement field [73] given as

$$\bar{u}_1(r,\theta) = \frac{K_I}{2G}\sqrt{\frac{r}{2\pi}}\left[\kappa - 1 + 2\sin^2\left(\frac{\theta}{2}\right)\right], \tag{36}$$

$$\bar{u}_2(r,\theta) = \frac{K_I}{2G}\sqrt{\frac{r}{2\pi}}\left[\kappa + 1 - 2\cos^2\left(\frac{\theta}{2}\right)\right], \tag{37}$$

where $\bar{u}_1$ and $\bar{u}_2$ are the displacement in the X_1 and X_2 directions, respectively, G is the shear modulus, and $\kappa = (3-\nu)/(1+\nu)$ for plain stress, r and θ denote the polar coordinates of boundary nodes/atoms measured from the crack tip. The given geometrical and load data is as follows: the overall size of the graphene sample is 163.7134 × 165.4100 Å, the crack length is 31.3 Å, while the stress intensity factor is set to $K_I = 177.8\ \text{GPa}\sqrt{\text{Å}}$. Deformed shapes obtained for the fully atomistic computation and for the coupled model are depicted in the Fig. 6. Note that in both cases the atomic interaction is governed by the modified Morse potential. The potential parameters are tuned to model the carbon-carbon bonds properly, see [34, 53].

In order to quantify the quality of the proposed modeling strategy, we define the atom-wise relative displacement error as

$$e_{u,i} = \frac{\sqrt{(\mathbf{d}_i - \mathbf{d}_i^{\text{ref}})^{\text{T}}(\mathbf{d}_i - \mathbf{d}_i^{\text{ref}})}_{\Omega^a}}{\|\mathbf{d}^{\text{ref}}\|_{\Omega}}, \qquad \forall i \in \Omega^a \tag{38}$$

where the norm is defined as follows

$$\|\mathbf{d}\| = \frac{1}{n_a}\sum_{i}^{n_a}\sqrt{\mathbf{d}_i^{\text{T}}\mathbf{d}_i}. \tag{39}$$

In the equations above, $\mathbf{d}_i$ and $\mathbf{d}_i^{\text{ref}}$ are the displacement of atom i, $\forall i \in \Omega^a$ related to the coupled, and fully atomistic model, respectively. The contour plot of displacement

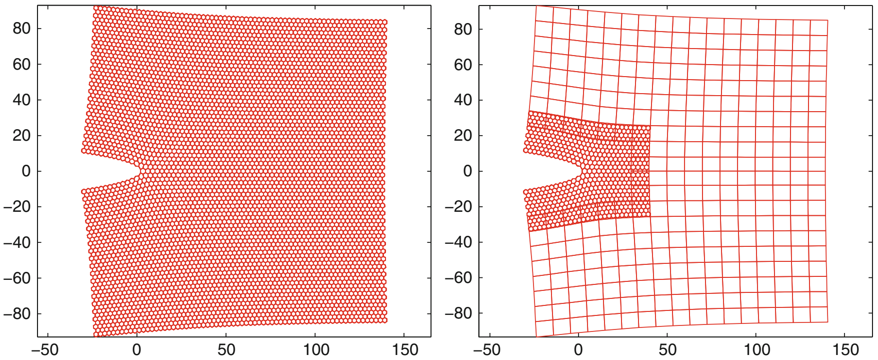


Fig. 6 Deformed shape of the graphene sheet with crack modelled using the fully atomistic model (*left*) and coupled model (*right*) with the size of atomistic domain 67.4 × 48.7 Å. Deformation scale factor is set to 20

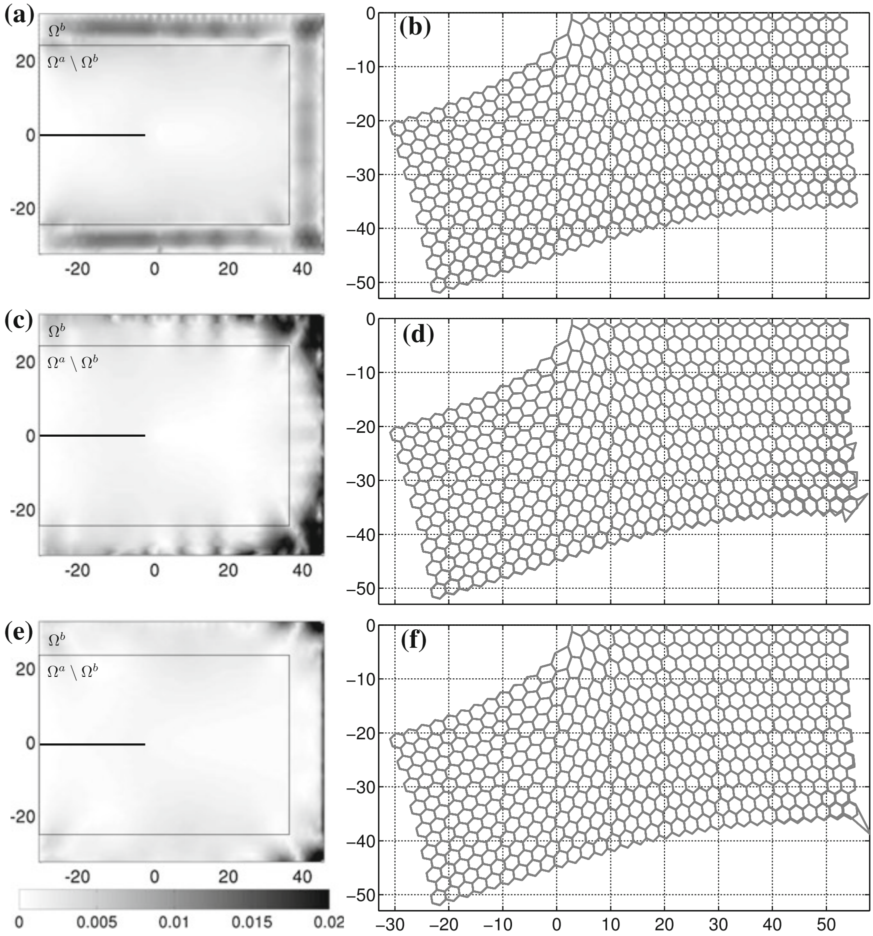


Fig. 7 The distribution of the local displacement error ($e_{u,i}$) on the contour of the domain Ω^a is given in the plots on the *left*. The corresponding deformed shapes (for coupled and reference models in overlap) are given on the plots on the *right*. Only half of the deformed plots is given due to symmetry, with rather large amplification magnitude factor of 40. The uppermost, middle and lower plots correspond to H^1-constant, H^1-linear, and L^2-linear couplings, respectively. The results are presented for the coupled model with the size of atomistic domain 67.4 × 48.7 Å

error (38) is given on the Fig. 7a, c and e for the three coupling options. Namely, the L^2 (28) with linear weighting, and the H^1 coupling (29) with both constant and linear weighting in Ω^b. The scale maximum is set to 2 % on the contour plots. The latter shows that H^1 coupling with constant weighting (Fig. 7a) the displacement error is noticeable in the entire bridging zone, being in general small and just slightly more noticeable in the corners. This results with the deformed shape which shows almost no difference from the reference, fully atomistic model (see Fig. 7b). For the coupling with linear weighting (Fig. 7c–f) of H^1 or L^2 type, the error is noticeable only in the corners. However, this error is somewhat larger, which is visible on the corners of superimposed deformed plots, see Fig. 7d, f. Note that the displacement is exaggerated with the deformation scaling factor of 40. The error in the corners of Ω^b is related to the problems of the integration of the coupling term for the case of the linear weighting function. Nevertheless, the error in zone of interest for the given example (Ω^a), is negligible.

4.1 Error Convergence

We proceed here with the adaptivity performance for the BD/A coupling. More precisely, we will use the size of the fine-scale model as a parameter that needs to be adapted. For that purpose we choose QOI in terms of relative energy related to the zone of interest $\Omega^a \setminus \Omega^b$. Let the global relative error in terms of the displacement be defined as

$$e_u = \frac{\|\mathbf{d} - \mathbf{d}^{\mathrm{ref}}\|_{\Omega^a \setminus \Omega^b}}{\|\mathbf{d}^{\mathrm{ref}}\|_{\Omega}}, \tag{40}$$

with $\|\mathbf{d}\|$ defined as in (39). The description of the models presented in the following results is given in Table 1. For the three different sizes of the Ω^a, and fully atomistic model we give the number of atoms n_a, number of nodes n_n, number of LM nodes n_λ, number of degrees of freedom $n_{dof} = 2(n_a + n_n + n_\lambda)$. The convergence in terms of the selected QOI is presented in Fig. 8. We seek here to show the displacement error in the zone of interest. Not surprisingly, the convergence is achieved as the size of the atomistic domain is increased. From Fig. 8 we can conclude that for the coupling with

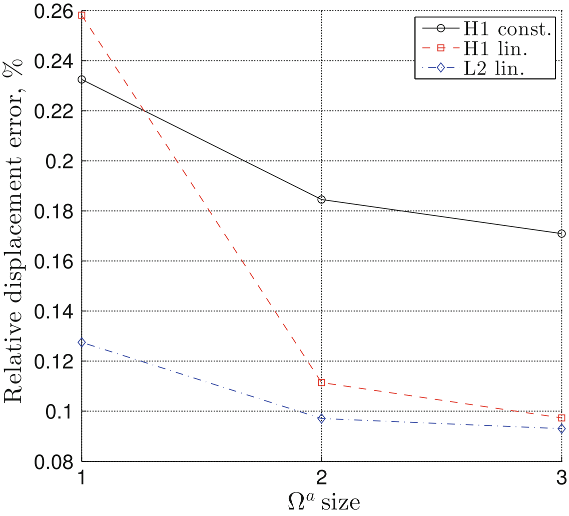
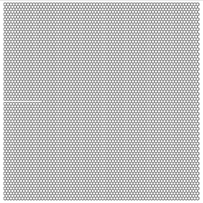
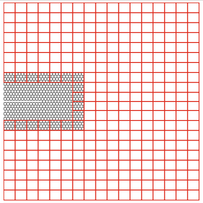
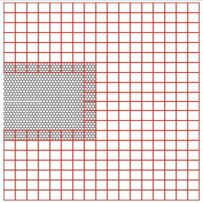
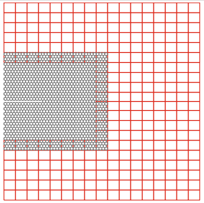


Fig. 8 Convergence of the global relative error in displacement e_u given for three different atomistic domain dimensions (given in Table 1) and the different couplings

Table 1 The data for the models named as ‘1’, ‘2’ and ‘3’ used in convergence study

				
id	–	1	2	3
n_a	10,960	1368	2080	2920
n_n	0	360	343	322
n_λ	0	38	46	54
n_{dof}	21,920	3532	4938	6592
L_1	163.7134	67.4100	77.0400	86.6700
L_2	165.4100	48.6500	65.3300	82.0100

The size of the atomistic domain is defined by $L_1 \times L_2$ and given in Å

H^1 the error in terms of the selected QOI is in general higher. Moreover, H^1-linear coupling decreases rapidly with the Ω^a increase, and it equates the coupling with L^2 for the model denoted as '3' in Table 1. It is important to note that for the coupling H^1 with added complexity regarding the calculation of the atomistic displacement gradient, no advantage over L^2 is visible, which leads to conclusion that for the MS modeling of deformation process of defected graphene in quasi-static application the L^2-linear coupling performs better. Note also that for the model denoted as '1' in the Table 1 the number of degrees of freedom is reduced by 84 %. Still the corresponding solution yields negligible error (less then 0.25 %) with respect to the fully atomistic model. Thus, both considered energy scaling constant and linear and coupling types that have been investigated show really good performance.

5 Conclusion

The MS atomistic-to-continuum modeling approach is shown to be an elegant way to keep the atomistic model of lattice structure and retain the computational affordability of reduced model based on continuum mechanics in such a way that atomistic representation is maintained only in the localized region around defect and is coupled to related equivalent continuum model. There is a number of available multiscale methods, however we focused on the bridging domain which enables inclusion of the atomistic submodel or patch in the continuum model. We showed that the perturbation caused by the coupling of the atomistic and continuum models in the overlapping zone is localized. Next, we confronted bridging domain method with one of the most prominent multiscale methods of this type, the quasicontinuum method, emphasising the adaptivity features. We implemented model adaptivity algorithm based on the a posteriori error estimates, and tested its performance choosing quantity of interest in terms of displacement. Moreover, a unified coupling formulation is proposed which shows that the two mentioned mainstream multiscale methods are similar, even though on the implementation level they may seem completely different. A good performance is presented on the problem of defected graphene sheet.

Acknowledgments The authors would like to acknowledge the collaboration with Pierre-Alain Guidault from LMT, ENS-Cachan, and the post-doctoral fellowship support from Labex MS2T-Maîtrise de Systèmes de Systèmes Technologiques.

References

1. D. C. Rapaport. *The Art of Molecular Dynmics Simulations*. Cambridge University Press, 2004.
2. Robert Phillips. *Crystals, Defects and Microstructures Modeling Across Scales*. Cambridge University Press, 2004.
3. Markus J. Buehler. *Atomistic Modeling of Materials Failure*. Springer US, 2008.

4. A. K. Geim and K. S. Novoselov. The rise of graphene. *Nature Materials*, 6:183–191, March 2007.
5. Wing Kam Liu, Eduard G. Karpov, and Harold S. Park. *Nano Mechanics and Materials Theory, Multiscale Methods and Applications*. John Wiley & Sons, Ltd, 2006.
6. Andrew N. Cleland. *Foundations of Nanomechanics From Solid-State Theory to Device Applications*. Springer, 2003.
7. M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Springer, Berlin, Heidelberg, 2007.
8. Lammps www site. http://lammps.sandia.gov/index.html, 2013.
9. Steve Plimpton. Fast parallel algorithms for shortrange molecular dynamics. *Journal of Computational Physics*, 117:1–19, 1995.
10. Furio Ercolessi. A molecular dynamics primer, June 1997.
11. Adnan Ibrahimbegovic. *Nonlinear Solid Mechanics*. Springer, 2009.
12. M. Mullins and M.A. Dokainish. Simulation of the (001) plane crack in alpha-iron employing a new boundary scheme. *Philosophical Magazine A*, 46:771–787, 1982.
13. S. Kohlhoff, P. Gumbsch, and H. F. Fischmeister. Crack propagation in b.c.c. crystals studied with a combined finite-element and atomistic model. *Philosophical Magazine A*, 64:4:851–878, 1991.
14. Deepak Srivastava and Satya N. Atluri. Computational nanotechnology: A current perspective. *CMES*, 3:531–538, 2002.
15. W A Curtin and Ronald E Miller. Atomistic/continuum coupling in computational materials science. *Modelling Simul. Mater. Sci. Eng.*, 11:33–68, 2003.
16. W. K. Liu, E. G. Karpov, S. Zhang, and H. S. Park. An introduction to computational nanomechanics and materials. *Computer Methods in Applied Mechanics and Engineering*, 193(17–20):1529–1578, 2004.
17. NM Ghoniem, EP Busso, N Kioussis, and H Huang. Multiscale modelling of nanomechanics and micromechanics: an overview. *Philosophical Magazine*, 83:3475–3528, 2003.
18. Jeremy Q. Broughton, Farid F. Abraham, Noam Bernstein, and Efthimios Kaxiras. Concurrent coupling of length scales: Methodology and application. *Phys. Rev. B*, 60(4):2391–2403, Jul 1999.
19. Jacob Fish. *Multiscale Methods Bridging the Scales in Science and Engineering*. Oxford Univeristy press, 2009.
20. J. M. Wernik and S. A. Meguid. Coupling atomistics and continuum in solids: status, prospects, and challenges. *International Journal of Mechanics and Materials in Design*, 5:79–110, 2009.
21. R.E. Rudd and J.Q. Broughton. Concurrent coupling of length scales in solid state systems. *physica status solidi (b)*, 217:251–291, 2000.
22. Wing Kam Liu Harold S. Park. An introduction and tutorial on multiple-scale analysis in solids. *Computer Methods in Applied Mechanics and Engineering*, 193:1733–1772, 2004.
23. E.G. Karpov, H. Yu, H.S. Park, Wing Kam Liu, Q. Jane Wang, and D. Qian. Multiscale boundary conditions in crystalline solids: Theory and application to nanoindentation. *International Journal of Solids and Structures*, 43(21):6359–6379, 2006.
24. Dong Qian, Gregory J. Wagner, and Wing Kam Liu. A multiscale projection method for the analysis of carbon nanotubes. *Computer Methods in Applied Mechanics and Engineering*, 193(17–20):1603–1632, 2004.
25. L. E. Shilkrot, W. A. Curtin, and R. E. Miller. A coupled atomistic/continuum model of defects in solids. *Journal of the Mechanics and Physics of Solids*, 50(10):2085–2106, 2002.
26. Jacob Fish, Mohan A. Nuggehally, Mark S. Shephard, Catalin R. Picu, Santiago Badia, Michael L. Parks, and Max Gunzburger. Concurrent AtC coupling based on a blend of the continuum stress and the atomistic force. *Computer Methods in Applied Mechanics and Engineering*, 196(45–48):4548–4560, 2007.
27. Santiago Badia, Pavel Bochev, Richard Lehoucq, Michael Parks, Jacob Fish, Mohan A. Nuggehally, and Max Gunzburger. A force-based blending model foratomistic-to-continuum coupling. *International Journal for Multiscale Computational Engineering*, 5(5):387–406, 2007.

28. S. Badia, M. Parks, P. Bochev, M. Gunzburger, and R. Lehoucq. On atomistic-to-continuum coupling by blending. *Multiscale modeling and simulation*, 7-1:381–406, 2008.
29. Farid F. Abraham, J. Q. Broughton, N. Bernstein, and E. Kaxiras. Spanning the continuum to quantum length scales in a dynamic simulation of brittle fracture. *Europhysics Letters*, 44(6):783–787, 1998.
30. Farid F. Abraham, Robert Walkup, Huajian Gao, Mark Duchaineau, Tomas Diaz De La Rubia, and Mark Seage. Simulating materials failure by using up to one billion atoms and the world's fastest computer: Brittle fracture. *PNAS*, 99:5777–5782, 2002.
31. Farid F. Abraham, Robert Walkup, Huajian Gao, Mark Duchaineau, Tomas Diaz De La Rubia, and Mark Seage. Simulating materials failure by using up to one billion atoms and the world's fastest computer: Work-hardening. *PNAS*, 99:5783–5787, 2002.
32. Fei Han and Gilles Lubineau. Coupling of nonlocal and local continuum models by the arlequin approach. *International Journal for Numerical Methods in Engineering*, 89(6):671–685, 2012.
33. Gilles Lubineau, Yan Azdoud, Fei Han, Christian Rey, and Abe Askari. A morphing strategy to couple non-local to local continuum mechanics. *Journal of the Mechanics and Physics of Solids*, 60(6):1088–1102, 2012.
34. T. Belytschko and S. P. Xiao. Coupling methods for continuum model with molecular model. *International Journal for Multiscale Computational Engineering*, 1:12, 2003.
35. E. B. Tadmor, M. Ortiz, and R. Phillips. Quasicontinuum analysis of defects in solids. *Philosophical Magazine A*, 73:1529–1563, 1996.
36. M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Oxford Univeristy press, 1987.
37. Philip M. Morse. Diatomic molecules according to the wave mechanics. ii. vibrational levels. *Phys. Rev.*, 34:57–64, Jul 1929.
38. R. Miller, E. B. Tadmor, R. Phillips, and M. Ortiz. Quasicontinuum simulation of fracture at the atomic scale. *Modeling and Simulation in Materials Science and Engineering*, 6:607–638, 1998.
39. R. Miller, M. Ortiz, R. Phillips, V. Shenoy, and E. B. Tadmor. Quasicontinuum models of fracture and plasticity. *Engineering Fracture Mechanics*, 61(3-4):427–444, 1998.
40. S. Hai and E. B. Tadmor. Deformation twinning at aluminum crack tips. *Acta Materialia*, 51(1):117–131, 2003.
41. V. B. Shenoy, R. Miller, E. b. Tadmor, D. Rodney, R. Phillips, and M. Ortiz. An adaptive finite element approach to atomic-scale mechanics–the quasicontinuum method. *Journal of the Mechanics and Physics of Solids*, 47(3):611–642, 1999.
42. Vivek B. Shenoy, Rob Phillips, and Ellad B. Tadmor. Nucleation of dislocations beneath a plane strain indenter. *Journal of the Mechanics and Physics of Solids*, 48(4):649–673, 2000.
43. Bernhard Eidel, Alexander Hartmaier, and Peter Gumbsch. Atomistic simulation methods and their application on fracture. In Reinhard Pippan, Peter Gumbsch, Friedrich Pfeiffer, Franz G. Rammerstorfer, Jean Salenon, Bernhard Schrefler, and Paolo Serafini, editors, *Multiscale Modelling of Plasticity and Fracture by Means of Dislocation Mechanics*, volume 522 of *CISM Courses and Lectures*, pages 1–57. Springer Vienna, 2010.
44. Ronald E. Miller and E. B. Tadmor. The quasicontinuum method: Overview, applications and current directions. *Journal of Computer-Aided Materials Design*, 9:203–239, 2002.
45. J. L. Ericksen. The cauchy and born hypotheses for crystals. *Phase transformation and material instabilities in solids - from book 'Mechanics and Mathematics of Crystals: Selected Papers of J. L. Ericksen' by Millard F. Beatty and Michael A. Hayes*, page 61–77, 1984.
46. J.L. Ericksen. On the cauchyborn rule. *Mathematics and Mechanics of Solids*, 13:199–220, 2008.
47. Giovanni Zanzotto. The cauchy-born hypothesis, nonlinear elasticity and mechanical twining in crystals. *Acta Crystallographica*, A52:839–849, 1996.
48. Zienkiewicz and Taylor. *The finite element method*. McGraw-Hill, 1994.
49. S. P. Xiao and T. Belytschko. A bridging domain method for coupling continua with molecular dynamics. *Computer Methods in Applied Mechanics and Engineering*, 193(17–20):1645–1669, 2004.

50. Sulin Zhang, Roopam Khare, Qiang Lu, and Ted Belytschko. A bridging domain and strain computation method for coupled atomistic-continuum modelling of solids. *International Journal for Multiscale Computational Engineering*, 70:913–933, 2007.
51. Guillaume Anciaux, Olivier Coulaud, Jean Roman, and Gilles Zerah. Ghost force reduction and spectral analysis of the 1d bridging method. Research Report RR-6582, INRIA, 2008.
52. Ted Belytschko, Robert Gracie, and Mei Xu. A continuum-to-atomistic bridging domain method for composite lattices. *International Journal for Numerical Methods in Engineering*, 81:1635–1658, 2010.
53. Eduard Marenić, Adnan Ibrahimbegovic, Jurica Sorić, and Pierre-Alain Guidault. Homogenized elastic properties of graphene for small deformations. *Materials: Special Issue "Computational Modeling and Simulation in Materials Study"*, 6(9):3764–3782, 2013.
54. Hashmi Ben Dhia and Guillaume Rateau. The Arlequin method as a flexible engineering design tool. *International Journal for Numerical Methods in Engineering*, 62:1442–1462, 2005.
55. Hachmi Ben Dhia, Nadia Elkhodja, and Franois-Xavier Roux. Multimodeling of multi-alterated structures in the Arlequin framework. solution with a domain-decomposition solver. *European Journal of Computational Mechanics*, 17:969–980, 2008.
56. P.A. Guidault and T. Belytschko. Bridging domain methods for coupled atomistic-continuum models with l^2 or h^1 couplings. *International Journal for Numerical Methods in Engineering*, 77-11:1566–1592, 2009.
57. Paul T. Bauman, Hachmi Ben Dhia, Nadia Elkhodja, J. Tinsley Oden, and Serge Prudhomme. On the application of the arlequin method to the coupling of particle and continuum models. *Computational Mechanics*, 42:511–530, 2008.
58. S. Prudhomme, H. Ben Dhia, P.T. Bauman, N. Elkhodja, and J.T. Oden. Computational analysis of modeling error for the coupling of particle and continuum models by the Arlequin method. *Computer Methods in Applied Mechanics and Engineering*, 197(41–42):3399–3409, 2008. Recent Advances in Computational Study of Nanostructures.
59. Paul T. Bauman, J. Tinsley Oden, and Serge Prudhomme. Adaptive multiscale modeling of polymeric materials with arlequin coupling and goals algorithms. *Computer Methods in Applied Mechanics and Engineering*, 198:799–818, 2009.
60. Serge Prudhomme, Ludovic Chamoin, Hachmi Ben Dhia, and Paul T. Bauman. An adaptive strategy for the control of modeling error in two-dimensional atomic-to-continuum coupling simulations. *Computer Methods in Applied Mechanics and Engineering*, 198(21–26):1887–1901, 2009. Advances in Simulation-Based Engineering Sciences - Honoring J. Tinsley Oden.
61. L. Chamoin, S. Prudhomme, H. Ben Dhia, and T. Oden. Ghost forces and spurious effects in atomic-to-continuum coupling methods by the arlequin approach. *International Journal for Numerical Methods in Engineering*, 83:1081–1113, 2010.
62. H. Ben Dhia, Ludovic Chamoin, J. Tinsley Oden, and Serge Prudhomme. A new adaptive modeling strategy based on optimal control for atomic-to-continuum coupling simulations. *Computer Methods in Applied Mechanics and Engineering*, In Press, Corrected Proof:–, 2010.
63. Mark Ainsworth and J.Tinsley Oden. A posteriori error estimation in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 142(12):1–88, 1997.
64. J.Tinsley Oden and Kumar S. Vemaganti. Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials: I. error estimates and adaptive algorithms. *Journal of Computational Physics*, 164(1):22–47, 2000.
65. J.Tinsley Oden and Serge Prudhomme. Estimation of modeling error in computational mechanics. *Journal of Computational Physics*, 182(2):496–515, 2002.
66. Serge Prudhomme, J. Tinsley Oden, Tim Westermann, Jon Bass, and Mark E. Botkin. Practical methods for a posteriori error estimation in engineering applications. *International Journal for Numerical Methods in Engineering*, 56(8):1193–1224, 2003.
67. J.T. Oden, S. Prudhomme, and P. Bauman. On the extension of goal-oriented error estimation and hierarchical modeling to discrete lattice models. *Computer Methods in Applied Mechanics and Engineering*, 194(34–35):3668–3688, 2005.
68. Adnan Ibrahimbegovic and Damijan Markovic. Strong coupling methods in multi-phase and multi-scale modeling of inelastic behavior of heterogeneous structures. *Computer Methods in Applied Mechanics and Engineering*, 192(28–30):3089–3107, 2003.

69. Damijan Markovic and Adnan Ibrahimbegovic. On micro-macro interface conditions for micro scale based FEM for inelastic behavior of heterogeneous materials. *Computer Methods in Applied Mechanics and Engineering*, 193(48–51):5503–5523, 2004.
70. M. Hautefeuille, J.-B. Colliat, A. Ibrahimbegovic, H.G. Matthies, and P. Villon. A multi-scale approach to model localized failure with softening. *Computers & Structures*, 9495(0):83–95, 2012.
71. K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov, and A. K. Geim. Two-dimensional atomic crystals. *PNAS*, 102-30:10451–10453, 2005.
72. Virendra Singh, Daeha Joung, Lei Zhai, Soumen Das, Saiful I. Khondaker, and Sudipta Seal. Graphene based materials: Past, present and future. *Progress in Materials Science*, 56(8):1178–1271, 2011.
73. Ted L. Anderson. *Fracture Mechanics: Fundamentals and Applications*. CRC Press; 3 edition, 2004.

Inverse Problems in a Bayesian Setting

Hermann G. Matthies, Elmar Zander, Bojana V. Rosić, Alexander Litvinenko and Oliver Pajonk

Abstract In a Bayesian setting, inverse problems and uncertainty quantification (UQ)—the propagation of uncertainty through a computational (forward) model—are strongly connected. In the form of conditional expectation the Bayesian update becomes computationally attractive. We give a detailed account of this approach via conditional approximation, various approximations, and the construction of filters. Together with a functional or spectral approach for the forward UQ there is no need for time-consuming and slowly convergent Monte Carlo sampling. The developed sampling-free non-linear Bayesian update in form of a filter is derived from the variational problem associated with conditional expectation. This formulation in general calls for further discretisation to make the computation possible, and we choose a polynomial approximation. After giving details on the actual computation in the framework of functional or spectral approximations, we demonstrate the workings of the algorithm on a number of examples of increasing complexity. At last, we compare the linear and nonlinear Bayesian update in form of a filter on some examples.

1 Introduction

Inverse problems deal with the determination of parameters in computational models, by comparing the prediction of these models with either real measurements or observations, or other, presumably more accurate, computations. These parameters can typically not be observed or measured directly, only other quantities which are

H.G. Matthies (✉) · E. Zander · B.V. Rosić
TU Braunschweig, Brunswick, Germany
e-mail: wire@tu-bs.de

A. Litvinenko
KAUST, Thuwal, Saudi Arabia

O. Pajonk
Elektrobit, Braunschweig, Germany

O. Pajonk
Schlumberger Information Solutions AS, Instituttveien 8, Kjeller, Norway

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_10

somehow connected to the one for which the information is sought. But it is typical that we can compute what the observed response should be, under the assumption that the unknown parameters have a certain value. And the difference between predicted or forecast response is obviously a measure for how well these parameters were identified.

There are different ways of attacking the problem of parameter identification theoretically and numerically. One way is to define some measure of discrepancy between predicted observation and the actual observation. Then one might use optimisation algorithms to make this measure of discrepancy as small as possible by changing the unknown parameters. Classical least squares approaches start from this point. The parameter values where a minimum is attained is then usually taken as the 'best' value and regarded as close to the 'true' value.

One of the problems is that for one the measure of discrepancy crops pretty arbitrarily, and on the other hand the minimum is often not unique. This means that there are many parameter values which explain the observations in a 'best' way. To obtain a unique solution, some kind of 'niceness' of the optimal solution is required, or mathematically speaking, for the optimal solution some regularity is enforced, typically in competition with discrepancy measure to be minimised. This optimisation approach hence leads to regularisation procedures, a good overview of which is given by [5].

Here we take another tack, and base our approach on the Bayesian idea of updating the knowledge about something like the unknown parameters in a probabilistic fashion according to Bayes's theorem. In order to apply this, the knowledge about the parameters has to be described in a *Bayesian* way through a probabilistic model [16, 40, 41]. As it turns out, such a probabilistic description of our previous knowledge can often be interpreted as a regularisation, thus tying these differing approaches together.

The Bayesian way is on one hand difficult to tackle, i.e. finding a computational way of doing it; and on the other hand often becomes computationally very demanding. One way the Bayesian update may be achieved computationally is through sampling. On the other hand, we shall here use a functional approximation setting to address such stochastic problems. See [26] for a synopsis on our approach to such parametric problems.

It is well-known that such a Bayesian update is in fact closely related to *conditional expectation* [2, 11], and this will be the basis of the method presented. For these and other probabilistic notions see for example [30] and the references therein.

The functional approximation approach towards stochastic problems is explained e.g. in [24]. These approximations are in the simplest case known as Wiener's so-called *homogeneous* or *polynomial chaos* expansion [43], which are polynomials in independent Gaussian RVs—the 'chaos'—and which can also be used numerically in a Galerkin procedure [10, 24, 25]. This approach has been generalised to other types of RVs [44]. It is a computational variant of *white noise analysis*, which means analysis in terms of independent RVs, hence the term 'white noise' [13–15], see also [8, 25, 33] for here relevant results on stochastic regularity. Here we describe a computational extensions of this approach to the inverse problem of Bayesian updating, see also [28, 29, 34, 35].

To be more specific, let us consider the following situation: we are investigating some physical system which is modelled by an evolution equation for its state:

$$\frac{\mathrm{d}}{\mathrm{d}t}u = A(q; u(t)) + \eta(q; t); \quad u(0) = u_a \text{ given.} \tag{1}$$

where $u(t) \in \mathcal{U}$ describes the state of the system at time $t \in [0, T]$ lying in a Hilbert space $\mathcal{U}$ (for the sake of simplicity), A is a—possibly non-linear—operator modelling the physics of the system, and $\eta \in \mathcal{U}^*$ is some external influence (action/excitation/loading). Both A and ℓ may involve some *noise*—i.e. a random process—so that (1) is a stochastic evolution equation.

Assume that the model depends on some parameters $q \in \mathcal{Q}$, which are uncertain. These may actually include the initial conditions for the state, u_a. To have a concrete example of Eq. (1), consider the diffusion equation

$$\frac{\partial}{\partial t}u(x, t) - \mathrm{div}(\kappa(x)\nabla u(x, t)) = \eta(x, t), \quad x \in \mathcal{G}, \tag{2}$$

with appropriate boundary and initial conditions, where $\mathcal{G} \subset \mathbb{R}^n$ is a suitable domain. The diffusing quantity is $u(x, t)$ (heat, concentration) and the term $\eta(x, t)$ models sinks and sources. Similar examples will be used for the numerical experiments in Sects. 5 and 6. Here $\mathcal{U} = H_E^1(\mathcal{G})$, the subspace of the Sobolev space $H^1(\mathcal{G})$ satisfying the essential boundary conditions, and we assume that the diffusion coefficient $\kappa(x)$ is uncertain. The parameters could be the positive diffusion coefficient field $\kappa(x)$, but for reasons to be explained fully later we prefer to take $q(x) = \log(\kappa(x))$, and assume $q \in \mathcal{Q} = L_2(\mathcal{G})$.

The updating methods have to be well defined and stable in a continuous setting, as otherwise one can not guarantee numerical stability with respect to the PDE discretisation refinement, see [40] for a discussion of related questions. Due to this we describe the update before any possible discretisation in the simplest Hilbert space setting.

On the other hand, no harm will result for the basic understanding if the reader wants to view the occurring spaces as finite dimensional Euclidean spaces. Now assume that we observe a function of the state $Y(u(q), q)$, and from this observation we would like to identify the corresponding q. In the concrete example Eq. (2) this could be the value of $u(x_j, t)$ at some points $x_j \in \mathcal{G}$. This is called the *inverse* problem, and as the mapping $q \mapsto Y(q)$ is usually not invertible, the inverse problem is *ill-posed*. Embedding this problem of finding the best q in a larger class by modelling our knowledge about it with the help of probability theory, then in a Bayesian manner the task becomes to estimate conditional expectations, e.g. see [16, 40, 41], and the references therein. The problem now is *well-posed*, but at the price of 'only' obtaining probability distributions on the possible values of q, which now is modelled as a $\mathcal{Q}$-valued random variable (RV). On the other hand one naturally also obtains information about the remaining uncertainty. Predicting what the measurement $Y(q)$ should be from some assumed q is computing the *forward* problem. The *inverse* problem is then approached by comparing the forecast from the forward problem with the actual information.

Since the parameters of the model to be estimated are uncertain, all relevant information may be obtained via their stochastic description. In order to extract information from the posterior, most estimates take the form of expectations w.r.t. the posterior. These expectations—mathematically integrals, numerically to be evaluated by some quadrature rule—may be computed via asymptotic, deterministic, or sampling methods. In our review of current work we follow our recent publications [28, 29, 34, 35].

One often used technique is a Markov chain Monte Carlo (MCMC) method [9, 21], constructed such that the asymptotic distribution of the Markov chain is the Bayesian posterior distribution; for further information see [34] and the references therein.

These approaches require a large number of samples in order to obtain satisfactory results. Here the main idea here is to perform the Bayesian update directly on the polynomial chaos expansion (PCE) without any sampling [26, 28, 29, 34, 35]. This idea has appeared independently in [1] in a simpler context, whereas in [37] it appears as a variant of the Kalman filter (e.g. [17]). A PCE for a push-forward of the posterior measure is constructed in [27].

From this short overview it may already have become apparent that the update may be seen abstractly in two different ways. Regarding the uncertain parameters

$$q : \Omega \to \mathcal{Q} \text{ as a RV on a probability space } (\Omega, \mathfrak{A}, \mathbb{P}) \tag{3}$$

where the set of elementary events is Ω, $\mathfrak{A}$ a σ-algebra of events, and $\mathbb{P}$ a probability measure, one set of methods performs the update by changing the probability measure $\mathbb{P}$ and leaving the mapping $q(\omega)$ as it is, whereas the other set of methods leaves the probability measure unchanged and updates the function $q(\omega)$. In any case, the push forward measure $q_*\mathbb{P}$ on $\mathcal{Q}$ defined by $q_*\mathbb{P}(\mathcal{R}) := \mathbb{P}(q^{-1}(\mathcal{R}))$ for a measurable subset $\mathcal{R} \subset \mathcal{Q}$ is changed from prior to posterior. For the sake of simplicity we assume here that $\mathcal{Q}$—the set containing possible realisations of q—is a Hilbert space. If the parameter q is a RV, then so is the state u of the system Eq. (1). In order to avoid a profusion of notation, unless there is a possibility of confusion, we will denote the random variables q, f, u which now take values in the respective spaces $\mathcal{Q}, \mathcal{U}^*$ and $\mathcal{U}$ with the same symbol as the previously deterministic quantities in Eq. (1).

In our overview on [34] spectral methods in identification problems, we show that Bayesian identification methods [11, 16, 40, 41] are a good way to tackle the identification problem, especially when these latest developments in functional approximation methods are used. In the series of papers [26, 29, 34, 35], Bayesian updating has been used in a linearised form, strongly related to the Gauss-Markov theorem [20], in ways very similar to the well-known Kalman filter [17]. These similarities ill be used to construct an abstract linear filter, which we term the **Gauss-Markov-Kalman** filter (GMKF). This turns out to be a linearised version of *conditional expectation*.

Here we want to extend this to a non-linear form, and show some examples of linear (LBU) and non-linear (QBU) Bayesian updates.

The organisation of the remainder of the paper is as follows: in Sect. 2 we review the Bayesian update—classically defined via conditional probabilities—and recall the link between conditional probability measures and conditional expectation. In the Sect. 3, the abstract version of the conditional expectation into real computational procedures.

We show how to approximate the conditional expectation up to any desired polynomial degree, not only the linearised version [17, 20] which was used in [26, 28, 29, 34, 35].

The numerical realisation in terms of a functional or spectral approximations—here we use the well known Wiener-Hermite chaos—is shortly sketched in Sect. 4. In Sect. 5 we then show some computational examples with the *linear version (LBU)*, whereas in Sect. 6 we show how to compute with the non-linear or quadratic (QBU) version. Some concluding remarks are offered in Sect. 7.

2 Bayesian Updating

Here we shall describe the frame in which we want to treat the problem of Bayesian updating, namely a dynamical system with time-discrete observations and updates. After introducing the setting in Sect. 2.1, we recall Bayes's theorem in Sect. 2.2 in the formulation of Laplace, as well as its formulation in the special case where densities exist, e.g. [2]. The next Sect. 2.3 treats the more general case and its connection with the notion of *conditional expectation*, as it was established by Kolmogorov, e.g. [2]. This notion will be the basis of our approach to characterise a RV which corresponds to the posterior measure.

2.1 Setting

In the setting of Eq. (1) consider the following problem: one makes observations y_n at times $0 < t_1 < \cdots < t_n \cdots \in [0, T]$, and from these one would like to infer what q (and possibly $u(q;t)$) is. In order to include a possible identification of the state $u(q;t_n)$, we shall define a new variable $x = (u, q)$, which we would thus like to identify:

Assume that $U : \mathcal{U} \times \mathcal{Q} \times [0, T] \ni (u_a, q, t) \mapsto u(q;t) \in \mathcal{U}$ is the flow or solution operator of Eq. (1), i.e. $u(q;t) = U(u_a, t_a, q, t)$, where u_a is the initial condition at time t_a. We then look at the operator which advances the variable $x = (u, q) \in \mathcal{X} = \mathcal{U} \times \mathcal{Q}$ from $x_n = (u_n, q)$ at time t_n to $x_{n+1} = (u_{n+1}, q)$ at t_{n+1}, where the Hilbert space $\mathcal{X}$ carries the natural inner product implied from $\mathcal{U}$ and $\mathcal{Q}$,

$$x_n = (u_n, q) \mapsto x_{n+1} = (u_{n+1}, q) = (U(u_n, t_n, q, t_{n+1}), q) \in \mathcal{X},$$

or a bit more generally encoded in an operator $\hat{f}$:

$$\forall n \in \mathbb{N}_0 : \quad x_{n+1} = \hat{f}(x_n, w_n, n); \quad x_0 = x_a \in \mathcal{X} \text{ given.} \tag{4}$$

This is a discrete time step advance map, for example of the dynamical system Eq. (1), where a random 'error' term w_n is included, which may be used to model randomness in the dynamical system per se, or possible discretisation errors, or both, or similar things. Most dynamical—and also quasi-static and stationary systems, considering different loadings as a sequence in some pseudo-time—can be put in the form Eq. (4) when observed at discrete points in time. Obviously, for fixed model parameters like q in Eq. (1) the evolution is trivial and does not change anything, but the Eq. (4) allows to model everything in one formulation.

Often the dependence on the random term is assumed to be linear, so that one has

$$\forall n \in \mathbb{N}_0 : \quad x_{n+1} = f(x_n) + \varepsilon S_x(x_n) w_n; \quad x_0 = x_a \text{ given,} \tag{5}$$

where the scalar $\varepsilon \geq 0$ explicitly measures the size of the random term w_n, which is now assumed to be discrete white noise of unit variance and zero mean, and possible correlations are introduced via the linear operator $S_x(x_n)$.

But one cannot observe the entity q or $u(q;t)$, i.e. $x = (q, u)$ directly—like in Plato's cave allegory we can only see a 'shadow'—here denoted by a vector $y \in \mathcal{Y}$—of it, formally given by a 'measurement operator'

$$Y : \mathcal{X} = \mathcal{Q} \times \mathcal{U} \ni (q, u(t_n)) \mapsto y_{n+1} = Y(q; u(t_n)) \in \mathcal{Y}, \tag{6}$$

where for the sake of simplicity we assume $\mathcal{Y}$ to be a Hilbert space.

Typically one considers also some observational 'error' εv_n, so that the observation may be expressed as

$$y_{n+1} = H(Y(q; u(t_n)), \varepsilon v_n) = \hat{h}(x_n, \varepsilon v_n), \tag{7}$$

where similarly as before v_n is a discrete white noise process, and the observer map H resp. $\hat{h}$ combines the 'true' quantity $Y(q; u(t_n))$ to be measured with the error, to give the observation y_n.

Translating this into the notation of the discrete dynamical system Eq. (4), one writes

$$y_{n+1} = \hat{h}(x_n, \varepsilon v_n) \in \mathcal{Y}, \tag{8}$$

where again the operator $\hat{h}$ is often assumed to be linear in the noise term, so that one has similarly to Eq. (5)

$$y_{n+1} = h(x_n) + \varepsilon S_y(x_n) w_n \in \mathcal{Y}. \tag{9}$$

The mappings Y in Eq. (6), H in Eq. (7), $\hat{h}$ in Eq. (8), resp. h Eq. (9) are usually not invertible and hence the problem is called *ill-posed*. One way to address this is via regularisation (see e.g. [5]), but here we follow a different track. Modelling our lack-of-knowledge about q and $u(t_n)$ in a Bayesian way [41] by replacing them with a $\mathcal{Q}$- resp. $\mathcal{U}$-valued random variable (RV), the problem becomes well-posed [40]. But of course one is looking now at the problem of finding a probability distribution that best fits the data; and one also obtains a probability distribution, not just *one* pair $x_n = (q, u(t_n))$.

We shall allow for $\mathcal{X}$ to be an infinite-dimensional space, as well as for $\mathcal{Y}$; although practically in any real situation only finitely many components are measured. But by allowing for the infinite-dimensional case, we can treat the case of partial differential equations—PDE models—like Eq. (1) directly and not just their discretisations as it often done, and we only use arguments which are independent on the number of observations. In particular this prevents hidden dependencies on local compactness, the dimension of the model, or the number of measurements, and the possible break-down of computational procedures as these dimensions grow, as they will be designed for the infinite-dimensional case. The procedure practically performed in a real computation on a finite-dimensional model and a finite-dimensional observation may then be seen as an approximation of the infinite-dimensional case, and analysed as such.

Here we focus on the use of a Bayesian approach inspired by the 'linear Bayesian' approach of [11] in the framework of 'white noise' analysis [13–15, 22, 43, 44]. Please observe that although the unknown 'truth' x_n may be a deterministic quantity, the model for the observed quantity y_{n+1} involves randomness, and it therefore becomes a RV as well.

To complete the mathematical setup we assume that Ω is a measure space with σ-algebra $\mathfrak{A}$ and with a probability measure $\mathbb{P}$, and that $x : \Omega \to \mathcal{X}$ and similarly q, u, and y are random variables (RVs). The corresponding *expectation* will be denoted by $\bar{x} = \mathbb{E}(x) = \int_\Omega x(\omega)\, \mathbb{P}(\mathrm{d}\omega)$, giving the mean $\bar{x}$ of the random variable, also denoted by $\langle x \rangle := \bar{x}$. The quantity $\tilde{x} := x - \bar{x}$ is the zero-mean or fluctuating part of the RV x.

The space of vector valued RVs, say $x : \Omega \to \mathcal{X}$, will for simplicity only be considered in the form $\mathscr{X} = \mathcal{X} \otimes \mathcal{S}$, where $\mathcal{X}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$, $\mathcal{S}$ is a Hilbert space of scalar RVs—here we shall simply take $\mathcal{S} = L_2(\Omega, \mathfrak{A}, \mathbb{P})$—with inner product $\langle \cdot, \cdot \rangle_{\mathcal{S}}$, and the tensor product signifies the Hilbert space completion with the scalar product as usually defined for elementary tensors $x_1 \otimes s_1, x_2 \otimes s_2 \in \mathscr{X}$ with $x_1, x_2 \in \mathcal{X}$ and $s_1, s_2 \in \mathcal{S}$ by

$$\langle\!\langle x_1 \otimes s_1, x_2 \otimes s_2 \rangle\!\rangle_{\mathscr{X}} := \langle x_1, x_2 \rangle_{\mathcal{X}} \langle s_1, s_2 \rangle_{\mathcal{S}},$$

and extended to all of $\mathscr{X}$ by linearity.

Obviously, we may also consider the expectation not only as a linear operator $\mathbb{E} : \mathscr{X} \to \mathcal{X}$, but, as $\mathcal{X}$ is isomorphic to the subspace of constants $\mathscr{X}_c := \mathcal{X} \otimes \operatorname{span}\{1\} \subset \mathscr{X}$, also as an orthogonal projection onto that subspace $\mathbb{E} = P_{\mathscr{X}_c}$, and we have the orthogonal decomposition

$$\mathscr{X} = \mathscr{X}_c \oplus \mathscr{X}_c^{\perp}, \text{ with } \mathscr{X}_c^{\perp} =: \mathscr{X}_0,$$

where $\mathscr{X}_0$ is the zero-mean subspace, so that

$$\forall x \in \mathscr{X} : \quad \bar{x} = \mathbb{E}(x) = P_{\mathscr{X}_c} x \in \mathscr{X}_c, \ \tilde{x} = (I - P_{\mathscr{X}_c})x \in \mathscr{X}_0.$$

Later, the covariance operator between two Hilbert-space valued RVs will be needed. The covariance operator between two RVs x and y is denoted by

$$C_{xy} : \mathcal{Y} \ni v \mapsto \mathbb{E}(\tilde{x} \langle \tilde{y}, v \rangle_{\mathscr{Y}}) \in \mathcal{X} \cong \mathscr{X}_c.$$

For $x \in \mathcal{X} \otimes \mathcal{S}$ and $y \in \mathcal{Y} \otimes \mathcal{S}$ it is also often written as $C_{xy} = \mathbb{E}(\tilde{x} \otimes \tilde{y})$.

2.2 Recollection of Bayes's Theorem

Bayes's theorem is commonly accepted as a consistent way to incorporate new knowledge into a probabilistic description [16, 41], and its present mathematical form is due to Laplace, so that a better denomination would be the *Bayes-Laplace* theorem.

The elementary textbook statement of the theorem is about conditional probabilities

$$\mathbb{P}(\mathcal{I}_x | \mathcal{M}_y) = \frac{\mathbb{P}(\mathcal{M}_y | \mathcal{I}_x)}{\mathbb{P}(\mathcal{M}_y)} \mathbb{P}(\mathcal{I}_x), \quad \mathbb{P}(\mathcal{M}_y) > 0, \tag{10}$$

where $\mathcal{I}_x \subseteq \mathcal{X}$ is some measurable subset of possible x's, and the measurable subset $\mathcal{M}_z \subseteq \mathcal{Y}$ is the information provided by the measurement. Here the conditional probability $\mathbb{P}(\mathcal{I}_x | \mathcal{M}_y)$ is called the *posterior* probability, $\mathbb{P}(\mathcal{I}_x)$ is called the *prior* probability, the conditional probability $\mathbb{P}(\mathcal{M}_y | \mathcal{I}_x)$ is called the *likelihood*, and $\mathbb{P}(\mathcal{M}_y)$ is called the evidence. The Eq. (10) is only valid when the set $\mathcal{M}_y$ has non-vanishing probability measure, and becomes problematic when $\mathbb{P}(\mathcal{M}_y)$ approaches zero, cf. [16, 32]. This arises often when $\mathcal{M}_y = \{y_m\}$ is a one-point set representing a measured value $y_m \in \mathcal{Y}$, as such sets have typically vanishing probability measure. In fact the well-known *Borel-Kolmogorov paradox* has led to numerous controversies and shows the possible ambiguities [16]. Typically the posterior measure is singular w.r.t. the prior measure, precluding a formulation in densities. Kolmogorov's resolution of this situation shall be sketched later.

One well-known very special case where the formulation in densities is possible, which has particular requirements on the likelihood, is when $\mathcal{X}$—as here—is a metric space, and there is a background measure μ on $(\mathcal{X}, \mathfrak{B}_{\mathcal{X}})$—$\mathfrak{B}_{\mathcal{X}}$ is the Borel-σ-algebra of $\mathcal{X}$—and similarly with ν and $(\mathcal{Y}, \mathfrak{B}_{\mathcal{Y}})$, and the RVs x and y have probability density functions (pdf) $\pi_x(x)$ w.r.t. μ and $\pi_y(y)$ w.r.t. ν resp., and a joint density $\pi_{xy}(x, y)$ w.r.t. $\mu \otimes \nu$. Then the theorem may be formulated as ([41] Chap. 1.5, [16, 32])

$$\pi_{(x|y)}(x|y) = \frac{\pi_{xy}(x, y)}{\pi_y(y)} = \frac{\pi_{(y|x)}(y|x)}{Z_y}\pi_x(x), \tag{11}$$

where naturally the marginal density $Z_y := \pi_y(y) = \int_{\mathcal{X}} \pi_{xy}(x, y)\, \mu(\mathrm{d}x)$ (from German *Zustandssumme*) is a normalising factor such that the conditional density $\pi_{(x|y)}(\cdot|y)$ integrates to unity w.r.t x. In this case the limiting case where $\mathbb{P}(\mathcal{M}_y)$ vanishes may be captured via the metric [16, 32]. The joint density

$$\pi_{xy}(x, y) = \pi_{(y|x)}(y|x)\pi_x(x)$$

may be factored into the likelihood function $\pi_{(y|x)}(y|x)$ and the prior density $\pi_x(x)$, like $\pi_y(y)$ a marginal density, $\pi_x(x) = \int_{\mathcal{Y}} \pi_{xy}(x, y)\nu(\mathrm{d}y)$. These terms in the second equality in Eq. (11) are in direct correspondence with those in Eq. (10). Please observe that the model for the RV representing the error in Eq. (8) determines the likelihood functions $\mathbb{P}(\mathcal{M}_y|\mathcal{I}_x)$ resp. $\pi_{(y|x)}(y|x)$. To require the existence of the joint density is quite restrictive. As Eq. (8) shows, y is a function of x, and a joint density on $\mathcal{X} \times \mathcal{Y}$ will generally not be possible as $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are most likely on a sub-manifold; but the situation of Eq. (9) is one possibility where a joint density may be established. The background densities are typically in finite dimensions the Lebesgue measure on $\mathbb{R}^d$, or more general Haar measures on locally compact Lie-groups [39]. Most computational approaches determine the pdfs [18, 23, 40].

However, to avoid the critical cases alluded to above, Kolmogorov already defined conditional probabilities via conditional expectation, e.g. see [2]. Given the conditional expectation operator $\mathbb{E}\left(\cdot|\mathcal{M}_y\right)$, the conditional probability is easily recovered as $\mathbb{P}(\mathcal{I}_x|\mathcal{M}_y) = \mathbb{E}\left(\chi_{\mathcal{I}_x}|\mathcal{M}_y\right)$, where $\chi_{\mathcal{I}_x}$ is the characteristic function of the subset $\mathcal{I}_x$. It may be shown that this extends the simpler formulation described by Eq. (10) or Eq. (11) and is the more fundamental notion, which we examine next. Its definition will lead directly to practical computational procedures.

2.3 Conditional Expectation

The easiest point of departure for conditional expectation [2] in our setting is to define it not just for one piece of measurement $\mathcal{M}_y$—which may not even be possible unambiguously—but for sub-σ-algebras $\mathfrak{S} \subset \mathfrak{A}$ on Ω. A sub-σ-algebra $\mathfrak{S}$ is a mathematical description of a reduced possibility of randomness—the smallest sub-σ-algebra $\{\emptyset, \Omega\}$ allows only the constants in $\mathscr{X}_c$—as it contains fewer events than the full algebra $\mathfrak{A}$. The connection with a measurement $\mathcal{M}_y$ is to take $\mathfrak{S} := \sigma(y)$, the σ-algebra generated by the measurement $y = \hat{h}(x, \varepsilon v)$ from Eq. (8). These are all events which are consistent with possible observations of some value for y. This means that the observation of y allows only a certain 'fineness' of information to be obtained, and this is encoded in the sub-σ-algebra $\mathfrak{S}$.

2.3.1 Scalar Random Variables

For scalar RVs—functions $r(x)$ of x with finite variance, i.e. elements of $\mathcal{S} := L_2(\Omega, \mathfrak{A}, \mathbb{P})$—the subspace corresponding to the sub-σ-algebra $\mathcal{S}_\infty := L_2(\Omega, \mathfrak{S}, \mathbb{P})$ is a closed subspace [2] of the full space $\mathcal{S}$. One example of such a scalar RV is the function

$$r(x) := \chi_{\mathcal{I}_x}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{I}_x, \\ 0 & \text{otherwise,} \end{cases}$$

mentioned at the end of Sect. 2.2 used to define conditional probability of the subset $\mathcal{I}_x \subseteq \mathcal{X}$ once a conditional expectation operator is defined: $\mathbb{P}(\mathcal{I}_x|\mathfrak{S}) = \mathbb{E}\left(\chi_{\mathcal{I}_x}|\mathfrak{S}\right)$.

Definition 1 For scalar functions of x—scalar RVs $r(x)$—in $\mathcal{S}$, the conditional expectation $\mathbb{E}\,(\cdot|\mathfrak{S})$ is defined as the orthogonal projection onto the closed subspace $\mathcal{S}_\infty$, so that $\mathbb{E}\,(r(x)|\mathfrak{S}) \in \mathcal{S}_\infty$, e.g. see [2].

The question is now on how to characterise this subspace $\mathcal{S}_\infty$, in order to make it more accessible for possible numerical computations. In this regard, note that the *Doob-Dynkin* lemma [2] assures us that if a RV $s(x)$—like $\mathbb{E}\,(r(x)|\mathfrak{S})$—is in the subspace $\mathcal{S}_\infty$, then $s(x) = \varphi(y)$ for some $\varphi \in L_0(\mathcal{Y})$, the space of measurable scalar functions on $\mathcal{Y}$. We state this key fact and the resulting new characterisation of the conditional expectation in

Proposition 2 *The subspace* $\mathcal{S}_\infty$ *is given by*

$$\mathcal{S}_\infty = \overline{\text{span}}\{\varphi \mid \varphi(\hat{h}(x, \varepsilon v));\;\; \varphi \in L_0(\mathcal{Y}) \;\; \textit{and}\; \varphi \in \mathcal{S}\}. \tag{12}$$

The conditional expectation of a scalar RV $r(x) \in \mathcal{S}$*, being the orthogonal projection, minimises the distance to the original RV over the whole subspace:*

$$\mathbb{E}\,(r(x)|\mathfrak{S}) := P_{\mathcal{S}_\infty}(r(x)) := \arg\min_{\tilde{r} \in \mathcal{S}_\infty} \|r(x) - \tilde{r}\|_{\mathcal{S}}, \tag{13}$$

where $P_{\mathcal{S}_\infty}$ *is the orthogonal projector onto* $\mathcal{S}_\infty$*. The Eqs. (12) and (13) imply the existence of a optimal map* $\phi \in L_0(\mathcal{Y})$ *such that*

$$\mathbb{E}\,(r(x)|\mathfrak{S}) = P_{\mathcal{S}_\infty}(r(x)) = \phi(\hat{h}(x, \varepsilon v)). \tag{14}$$

In Eq. (13), one may equally well minimise the square of the distance, the loss-function

$$\beta_{r(x)}(\tilde{r}) = \frac{1}{2}\,\|r(x) - \tilde{r}\|^2_{\mathcal{S}}. \tag{15}$$

Taking the vanishing of the first variation/Gâteaux derivative of the loss-function Eq. (15) as a necessary condition for a minimum leads to a simple geometrical interpretation: the difference between the original scalar RV $r(x)$ *and its projection has to be perpendicular to the subspace:*

$$\forall \tilde{r} \in \mathcal{S}_\infty : \ \langle r(x) - \mathbb{E}\left(r(x)|\mathfrak{S}\right), \tilde{r}\rangle_{\mathcal{S}} = 0, \ i.e. r(x) - \mathbb{E}\left(r(x)|\mathfrak{S}\right) \in \mathcal{S}_\infty^\perp. \tag{16}$$

Rephrasing Eq. (13) with account to Eqs. (16) and (15) leads for the optimal map $\phi \in L_0(\mathcal{Y})$ to

$$\mathbb{E}\left(r(x)|\sigma(y)\right) = \phi(\hat{h}(x, \varepsilon v)) := \arg\min_{\varphi \in L_0(\mathcal{Y})} \beta_{r(x)}(\varphi(\hat{h}(x, \varepsilon v))), \tag{17}$$

and the orthogonality condition of Eq. (17) which corresponds to Eq. (16) leads to

$$\forall \varphi \in L_0(\mathcal{Y}) : \ \langle r(x) - \phi(\hat{h}(x, \varepsilon v)), \varphi(\hat{h}(x, \varepsilon v))\rangle_{\mathcal{S}} = 0. \tag{18}$$

Proof The Eq. (12) is a direct statement of the Doob-Dynkin lemma [2], and the Eq. (13) is equivalent to the definition of the conditional expectation being an orthogonal projection in $L_2(\Omega, \mathfrak{A}, \mathbb{P})$—actually an elementary fact of Euclidean geometry.

The existence of the optimal map ϕ in Eq. (14) is a consequence of the minimisation of a continuous, coercive, and strictly convex function—the norm Eq. (13)—over the closed set $\mathcal{S}_\infty$ in the complete space $\mathcal{S}$. The equivalence of minimising the norm Eqs. (13) and (15) is elementary, which is re-stated in Eq. (17).

The two equivalents statements—the 'Galerkin orthogonality' conditions—Eqs. (16) and (18) follow not only from requiring the Gâteaux derivative of Eq. (15) to vanish, but also express an elementary fact of Euclidean geometry.

The square of the distance $r(x) - \phi(y)$ may be interpreted as a difference in variance, tying conditional expectation with variance minimisation; see for example [2, 30], and the references therein for basic descriptions of conditional expectation. See also [20].

2.3.2 Vector Valued Random Variables

Now assume that $R(x)$ is a function of x which takes values in a vector space $\mathcal{R}$, i.e. a $\mathcal{R}$-valued RV, where $\mathcal{R}$ is a Hilbert space. Two simple examples are given by the conditional mean where $R(x) := x \in \mathcal{X}$ with $\mathcal{R} = \mathcal{X}$, and by the conditional variance where one takes $R(x) := (x - \bar{x}) \otimes (x - \bar{x}) = (\tilde{x}) \otimes (\tilde{x})$, where $\mathcal{R} = \mathscr{L}(\mathcal{X})$. The Hilbert tensor product $\mathscr{R} = \mathcal{R} \otimes \mathcal{S}$ is again needed for such vector valued RVs, where a bit more formalism is required, as we later want to take linear combinations of RVs, but with linear operators as 'coefficients' [20], and this is most clearly expressed in a component-free fashion in terms of L-invariance, where we essentially follow [3, 4]:

Definition 3 Let $\mathscr{V}$ be a subspace of $\mathscr{R} = \mathcal{R} \otimes \mathcal{S}$. The subspace is called *linearly closed*, *L-closed*, or *L-invariant*, iff $\mathscr{V}$ is closed, and $\forall v \in \mathscr{V}$ and $\forall L \in \mathscr{L}(\mathcal{R})$ it holds that $Lv \in \mathscr{V}$.

In finite dimensional spaces one can just apply the notions for the scalar case in Sect. 2.3.1 component by component, but this is not possible in the infinite dimensional case. Of course the vectorial description here collapses to the scalar case upon taking $\mathcal{R} = \mathbb{R}$. From [3] one has the following

Proposition 4 *It is obvious that the whole space $\mathscr{R} = \mathcal{R} \otimes \mathcal{S}$ is linearly closed, and that for a linearly closed subspace $\mathscr{V} \subseteq \mathscr{R}$ its orthogonal complement $\mathscr{V}^\perp$ is also linearly closed. Clearly, for a closed subspace $\mathcal{S}_a \subseteq \mathcal{S}$, the tensor space $\mathcal{R} \otimes \mathcal{S}_a$ is linearly closed, and hence the space of constants $\mathscr{R}_c = \mathcal{R} \otimes \operatorname{span}\{\chi_\Omega\} \cong \mathcal{R}$ is linearly closed, as well as its orthogonal complement $\mathscr{R}_0 = \mathscr{R}_c^\perp$, the subspace of zero-mean RVs.*

Let $v \in \mathscr{R}$ be a RV, and denote by

$$\mathcal{R}_v := \overline{\operatorname{span}}\; v(\Omega), \quad \sigma(v) := \{v^{-1}(B) \;:\; B \in \mathfrak{B}_\mathcal{R}\} \tag{19}$$

the closure of the span of the image of v and the σ-algebra generated by v, where $\mathfrak{B}_\mathcal{R}$ is the Borel-σ-algebra of $\mathcal{R}$. Denote the closed subspace generated by $\sigma(v)$ by $\mathcal{S}_v := L_2(\Omega, \sigma(v), \mathbb{P}) \subseteq \mathcal{S}$. Let $\mathscr{R}_{Lv} := \overline{\operatorname{span}}\; \{Lv \;:\; L \in \mathscr{L}(\mathcal{R})\} \subseteq \mathscr{R}$, the linearly closed subspace generated by v, and finally denote by $\mathscr{R}_v := \operatorname{span}\{v\} \subseteq \mathscr{R}$, the one-dimensional ray and hence closed subspace generated by v. Obviously it holds that

$$v \in \mathscr{R}_v \subseteq \mathscr{R}_{Lv} \subseteq \mathcal{R} \otimes \mathcal{S}_v \subseteq \mathscr{R}, \quad \text{and } \bar{v} \in \mathcal{R}_v,$$

and $\mathcal{R} \otimes \mathcal{S}_v$ is linearly closed according to Proposition 4.

Definition 5 Let $\mathscr{V}$ and $\mathscr{W}$ be subspaces of $\mathscr{R}$, and $v, w \in \mathscr{R}$ two RVs.

- The two subspaces are *weakly orthogonal* or simply just *orthogonal*, denoted by $\mathscr{V} \perp \mathscr{W}$, iff $\forall v \in \mathscr{V}, \forall w \in \mathscr{W}$ it holds that $\langle\!\langle v, w \rangle\!\rangle_\mathscr{R} = 0$.
- A RV $v \in \mathscr{R}$ is *weakly orthogonal* or simply just *orthogonal* to the subspace $\mathscr{W}$, denoted by
  $v \perp \mathscr{W}$, iff $\mathscr{R}_v \perp \mathscr{W}$, i.e. $\forall w \in \mathscr{W}$ it holds that $\langle\!\langle v, w \rangle\!\rangle_\mathscr{R} = 0$.
- Two RVs $v, w \in \mathscr{R}$ are *weakly orthogonal* or as usual simply just *orthogonal*, denoted by
  $v \perp w$, iff $\langle\!\langle v, w \rangle\!\rangle_\mathscr{R} = 0$, i.e. $\mathscr{R}_v \perp \mathscr{R}_w$.
- The two subspaces $\mathscr{V}$ and $\mathscr{W}$ are *strongly orthogonal* or *L-orthogonal*, iff they are linearly closed—Definition 3—and it holds that $\langle\!\langle Lv, w \rangle\!\rangle_\mathscr{R} = 0$, $\forall v \in \mathscr{V}, \forall w \in \mathscr{W}$ and $\forall L \in \mathscr{L}(\mathcal{R})$. This is denoted by
  $\mathscr{V} \underline{\perp} \mathscr{W}$, and in other words $\mathscr{L}(\mathcal{R}) \ni C_{vw} = \mathbb{E}(v \otimes w) = 0$.
- The RV v is *strongly orthogonal* to a linearly closed subspace $\mathscr{W} \subseteq \mathscr{R}$, denoted by
  $v \underline{\perp} \mathscr{W}$, iff $\mathscr{R}_{Lv} \underline{\perp} \mathscr{W}$, i.e. $\forall w \in \mathscr{W}$ it holds that $C_{vw} = 0$.
- The two RVs v, w are *strongly orthogonal* or simply just *uncorrelated*, denoted by
  $v \underline{\perp} w$, iff $C_{vw} = 0$, i.e. $\mathscr{R}_{Lv} \underline{\perp} \mathscr{R}_{Lw}$.
- Let $\mathfrak{C}_1, \mathfrak{C}_2 \subseteq \mathfrak{A}$ be two sub-σ-algebras. They are *independent*, denoted by
  $\mathfrak{C}_1 \perp\!\!\!\perp \mathfrak{C}_2$, iff the closed subspaces of $\mathcal{S}$ generated by them are orthogonal in $\mathcal{S}$:
  $L_2(\Omega, \mathfrak{C}_1, \mathbb{P}) \perp L_2(\Omega, \mathfrak{C}_2, \mathbb{P})$.

- The two subspaces $\mathscr{V}$ and $\mathscr{W}$ are *stochastically independent*, denoted by $\mathscr{V} \perp\!\!\!\perp \mathscr{W}$, iff the sub-$\sigma$-algebras generated are: $\sigma(\mathscr{V}) \perp\!\!\!\perp \sigma(\mathscr{W})$.
- The two RVs v, w are *stochastically independent*, denoted by $v \perp\!\!\!\perp w$, iff $\sigma(v) \perp\!\!\!\perp \sigma(w)$, i.e. $\mathcal{S}_v \perp \mathcal{S}_w$.

Proposition 6 *Obviously $\mathscr{R}_c \underline{\perp} \mathscr{R}_0$. It is equally obvious that for any two closed subspaces $\mathcal{S}_a, \mathcal{S}_b \subseteq \mathcal{S}$, the condition $\mathcal{S}_a \perp \mathcal{S}_b$ implies that the tensor product subspaces are strongly orthogonal:*

$$\mathcal{R} \otimes \mathcal{S}_a \underline{\perp} \mathcal{R} \otimes \mathcal{S}_b.$$

This implies that for a closed subspace $\mathcal{S}_s \subseteq \mathcal{S}$ the subspaces $\mathscr{R}_s = \mathcal{R} \otimes \mathcal{S}_s \subseteq \mathscr{R}$ and its orthogonal complement $\mathscr{R}_s^{\perp} = \mathcal{R} \otimes \mathcal{S}_s^{\perp}$ are linearly closed and strongly orthogonal.

We note from [3, 4] the following results which we collect in

Proposition 7 *Let $v, w \in \mathscr{R}_0$ be two zero-mean RVs. Then*

$$v \perp\!\!\!\perp w \Rightarrow v \underline{\perp} w \Rightarrow v \perp w.$$

Strong orthogonality in general does not imply independence, and orthogonality does not imply strong orthogonality, unless $\mathcal{R}$ is one-dimensional.

If $\mathscr{S} \subseteq \mathscr{R}$ is linearly closed, then

$$v \perp \mathscr{S} \Rightarrow v \underline{\perp} \mathscr{S}, \text{ i.e. } \mathscr{R}_v \perp \mathscr{S} \Rightarrow \mathscr{R}_v \underline{\perp} \mathscr{S} \Rightarrow \mathscr{R}_{Lv} \underline{\perp} \mathscr{S}.$$

From this we obtain the following:

Lemma 8 *Set $\mathscr{R}_\infty := \mathcal{R} \otimes \mathcal{S}_\infty$ for the $\mathcal{R}$-valued RV $R(x)$ with finite variance on the sub-σ-algebra $\mathfrak{S}$, representing the new information.*

Then $\mathscr{R}_\infty$ is L-invariant or strongly closed, and for any zero mean RV $v \in \mathscr{R}$:

$$v \in \mathscr{R}_\infty^{\perp} \Leftrightarrow v \perp \mathscr{R}_\infty \Rightarrow v \underline{\perp} \mathscr{R}_\infty. \tag{20}$$

In addition, it holds—even if $v \in \mathscr{R}$ is not zero mean—that

$$v \in \mathscr{R}_\infty^{\perp} \Leftrightarrow v \perp \mathscr{R}_\infty \Rightarrow \forall w \in \mathscr{R}_\infty : \ \mathbb{E}\left(v \otimes w\right) = 0. \tag{21}$$

Proof $\mathscr{R}_\infty$ is of the type $\mathcal{R} \otimes \mathcal{S}_\infty$ where $\mathcal{S}_\infty$ is a closed subspace, and $\mathcal{R}$ is obviously closed. From the remarks above it follows that $\mathscr{R}_\infty$ is L-invariant or linearly resp. strongly closed. The Eq. (20) is a direct consequence of Proposition 7.

To prove Eq. (21), take any $w \in \mathscr{R}_\infty$ and any $L \in \mathscr{L}(\mathcal{R})$, then

$$v \in \mathscr{R}_\infty^{\perp} \Rightarrow 0 = \langle\!\langle v, w \rangle\!\rangle_{\mathscr{R}} = \langle\!\langle v, Lw \rangle\!\rangle_{\mathscr{R}} = \mathbb{E}\left(\langle v, Lw \rangle_{\mathcal{R}}\right).$$

Now, for any $r_1, r_2 \in \mathcal{R}$, take the mapping $L : r_* \mapsto \langle r_2, r_* \rangle_{\mathcal{R}}\, r_1$, yielding

$$\begin{aligned}0 &= \mathbb{E}\left(\langle v, Lw\rangle_{\mathcal{R}}\right) = \mathbb{E}\left(\langle v, \langle r_2, w\rangle_{\mathcal{R}}\, r_1\rangle_{\mathcal{R}}\right)\\ &= \mathbb{E}\left(\langle v, r_1\rangle_{\mathcal{R}}\langle r_2, w\rangle_{\mathcal{R}}\right) = \langle r_1, \mathbb{E}\left(v \otimes w\right)\, r_2\rangle_{\mathcal{R}} \Leftrightarrow \mathbb{E}\left(v \otimes w\right) \equiv 0.\end{aligned}$$

Extending the scalar case described in Sect. 2.3.1, instead of

$$\mathcal{S} = L_2(\Omega, \mathbb{P}, \mathfrak{A}) = L_2(\Omega, \mathbb{P}, \mathfrak{A}; \mathbb{R}) \cong \mathbb{R} \otimes L_2(\Omega, \mathbb{P}, \mathfrak{A}) = \mathbb{R} \otimes \mathcal{S}$$

and its subspace generated by the measurement

$$\mathcal{S}_\infty = L_2(\Omega, \mathbb{P}, \mathfrak{S}) = L_2(\Omega, \mathbb{P}, \mathfrak{S}; \mathbb{R}) \cong \mathbb{R} \otimes L_2(\Omega, \mathbb{P}, \mathfrak{S}) = \mathbb{R} \otimes \mathcal{S}_\infty$$

one now considers the space Eq. (22) and its subspace Eq. (23)

$$L_2(\Omega, \mathbb{P}, \mathfrak{A}; \mathcal{R}) \cong \mathcal{R} \otimes L_2(\Omega, \mathbb{P}, \mathfrak{A}) = \mathcal{R} \otimes \mathcal{S} := \mathscr{R} \text{ and} \tag{22}$$

$$L_2(\Omega, \mathbb{P}, \mathfrak{S}; \mathcal{R}) \cong \mathcal{R} \otimes L_2(\Omega, \mathbb{P}, \mathfrak{S}) = \mathcal{R} \otimes \mathcal{S}_\infty := \mathscr{R}_\infty \subseteq \mathscr{R}. \tag{23}$$

The conditional expectation in the vector-valued case is defined completely analogous to the scalar case, see Definition 1:

Definition 9 For $\mathcal{R}$-valued functions of x—vectorial RVs $R(x)$—in the Hilbert-space $\mathscr{R}$ Eq. (22), the conditional expectation $\mathbb{E}\left(\cdot|\mathfrak{S}\right) : \mathscr{R} \to \mathscr{R}$ is defined as the orthogonal projection onto the closed subspace $\mathscr{R}_\infty$ Eq. (23), denoted by $P_{\mathscr{R}_\infty}$, so that $\mathbb{E}\left(R(x)|\mathfrak{S}\right) = P_{\mathscr{R}_\infty}(R(x)) \in \mathscr{R}_\infty$, e.g. see [2, 3].

From this one may derive a characterisation of the conditional expectation similar to Proposition 2.

Theorem 10 *The subspace $\mathscr{R}_\infty$ is given by*

$$\mathscr{R}_\infty = \{\varphi \mid \varphi(\hat{h}(x, \varepsilon v)) \in \mathscr{R};\ \ \varphi \in L_0(\mathcal{Y}, \mathcal{R})\}. \tag{24}$$

The conditional expectation of a vector-valued RV $R(x) \in \mathscr{R}$, being the orthogonal projection, minimises the distance to the original RV over the whole subspace:

$$\mathbb{E}\left(R(x)|\mathfrak{S}\right) := P_{\mathscr{R}_\infty}(R(x)) := \arg\min_{\tilde{R} \in \mathscr{R}_\infty} \|R(x) - \tilde{R}\|_{\mathscr{R}}, \tag{25}$$

where $P_{\mathscr{R}_\infty}$ is the orthogonal projector onto $\mathscr{R}_\infty$. The Eqs. (24) and (25) imply the existence of a optimal map $\Phi \in L_0(\mathcal{Y}, \mathcal{R})$ such that

$$\mathbb{E}\left(R(x)|\mathfrak{S}\right) = P_{\mathscr{R}_\infty}(R(x)) = \Phi(\hat{h}(x, \varepsilon v)). \tag{26}$$

In Eq. (25), one may equally well minimise the square of the distance, the loss-function

$$\beta_{R(x)}(\tilde{R}) = \frac{1}{2}\, \|R(x) - \tilde{R}\|^2_{\mathscr{R}}. \tag{27}$$

Taking the vanishing of the first variation/Gâteaux derivative of the loss-function Eq. (27) as a necessary condition for a minimum leads to a simple geometrical interpretation: the difference between the original vector-valued RV $R(x)$ and its projection has to be perpendicular to the subspace $\mathscr{R}_\infty$: $\forall \tilde{R} \in \mathscr{R}_\infty$:

$$\langle\!\langle R(x) - \mathbb{E}\left(R(x)|\mathfrak{S}\right), \tilde{R}\rangle\!\rangle_{\mathscr{R}} = 0, \;\; i.e.\;\; R(x) - \mathbb{E}\left(R(x)|\mathfrak{S}\right) \in \mathscr{R}_\infty^\perp. \tag{28}$$

Rephrasing Eq. (25) with account to Eqs. (28) and (27) leads for the optimal map $\Phi \in L_0(\mathcal{Y}, \mathcal{R})$ to

$$\mathbb{E}\left(R(x)|\sigma(y)\right) = \Phi(\hat{h}(x, \varepsilon v)) := \arg\min_{\varphi \in L_0(\mathcal{Y},\mathcal{R})} \beta_{R(x)}(\varphi(\hat{h}(x, \varepsilon v))), \tag{29}$$

and the orthogonality condition of Eq. (29) which corresponds to Eq. (28) leads to

$$\forall \varphi \in L_0(\mathcal{Y}, \mathcal{R}) : \; \langle\!\langle R(x) - \Phi(\hat{h}(x, \varepsilon v)), \varphi(\hat{h}(x, \varepsilon v))\rangle\!\rangle_{\mathscr{R}} = 0. \tag{30}$$

In addition, as $\mathscr{R}_\infty$ is linearly closed, one obtains the useful statement

$$\forall \tilde{R} \in \mathscr{R}_\infty : \; \mathscr{L}(\mathcal{R}) \ni \mathbb{E}\left((R(x) - \mathbb{E}\left(R(x)|\mathfrak{S}\right)) \otimes \tilde{R}\right) = 0. \tag{31}$$

or rephrased $\forall \varphi \in L_0(\mathcal{Y}, \mathcal{R})$:

$$\mathscr{L}(\mathcal{R}) \ni \mathbb{E}\left((R(x) - \Phi(\hat{h}(x, \varepsilon v))) \otimes \varphi(\hat{h}(x, \varepsilon v))\right) = 0. \tag{32}$$

Proof The Eq. (24) is just a version of the Doob-Dynkin lemma again [2], this time for vector-valued functions. The Eqs. (25)–(30) follow just as in the scalar case of Proposition 2.

As $\mathscr{R}_\infty$ is linearly closed according to Lemma 8, the Eq. (20) causes Eq. (28) to imply Eq. (31), and Eq. (30) together with Eq. (21) from Lemma 8 to imply Eq. (32).

Already in [17] it was noted that the conditional expectation is the best estimate not only for the *loss function* 'distance squared', as in Eqs. (15) and (27), but for a much larger class of loss functions under certain distributional constraints. However for the quadratic loss function this is valid without any restrictions.

Requiring the derivative of the quadratic loss function in Eqs. (15) and (27) to vanish may also be characterised by the *Lax-Milgram* lemma, as one is minimising a quadratic functional over the vector space $\mathscr{R}_\infty$, which is closed and hence a *Hilbert* space. For later reference, this result is recollected in

Theorem 11 *In the scalar case, there is a unique minimiser $\mathbb{E}\left(r(x)|\mathfrak{S}\right) = P_{\mathcal{S}_\infty}(r(x)) \in \mathcal{S}_\infty$ to the problem in Eq. (13), and it is characterised by the* orthogonality *condition Eq. (16)*

$$\forall \tilde{r} \in \mathcal{S}_\infty : \quad \langle r(x) - \mathbb{E}\left(r(x)|\mathfrak{S}\right), \tilde{r}\rangle_{\mathcal{S}} = 0. \tag{33}$$

The minimiser is unique as an element of $\mathcal{S}_\infty$, but the mapping $\phi \in L_0(\mathcal{Y})$ in Eq. (17) may not necessarily be. It also holds that

$$\|P_{\mathcal{S}_\infty}(r(x))\|_{\mathcal{S}}^2 = \|r(x)\|_{\mathcal{S}}^2 - \|r(x) - P_{\mathcal{S}_\infty}(r(x))\|_{\mathcal{S}}^2. \tag{34}$$

As in the scalar case, in the vector-valued case there is a unique minimiser $\mathbb{E}\,(R(x)|\mathfrak{S}) = P_{\mathscr{R}_\infty}(R(x)) \in \mathscr{R}_\infty$ to the problem in Eq. (25), which satisfies the orthogonality *condition Eq. (28)*

$$\forall \tilde{R} \in \mathscr{R}_\infty : \quad \langle\!\langle R(x) - \mathbb{E}\,(R(x)|\mathfrak{S})\,, \tilde{R}\rangle\!\rangle_{\mathscr{R}} = 0, \tag{35}$$

which is equivalent to the strong orthogonality condition *Eq. (31)*

$$\forall \tilde{R} \in \mathscr{R}_\infty : \quad \mathbb{E}\left(R(x) - \mathbb{E}\,(R(x)|\mathfrak{S}) \otimes \tilde{R}\right) = 0. \tag{36}$$

The minimiser is unique as an element of $\mathscr{R}_\infty$, but the mapping $\Phi \in L_0(\mathcal{Y}, \mathcal{R})$ in Eq. (29) may not necessarily be. It also holds that

$$\|P_{\mathscr{R}_\infty}(R(x))\|_{\mathscr{R}}^2 = \|R(x)\|_{\mathscr{R}}^2 - \|R(x) - P_{\mathscr{R}_\infty}(R(x))\|_{\mathscr{R}}^2. \tag{37}$$

Proof It is all already contained in Proposition 2 resp. Theorem 10. Except for Eq. (36), this is just a re-phrasing of the *Lax-Milgram* lemma, as the bi-linear functional—in this case the inner product—is naturally coercive and continuous on the subspace $\mathscr{R}_\infty$, which is closed and hence a *Hilbert* space. The only novelty here are the Eqs. (34) and (37) which follow from *Pythagoras's* theorem.

3 Characterising the Posterior

The information contained in the Bayesian update is encoded in the conditional expectation. And it only characterises the distribution of the posterior. A few different ways of characterising the distribution via the conditional expectation are sketched in Sect. 3.1. But in many situations, notably in the setting of Eq. (4) or Eq. (5), with the observations according to Eq. (8) or Eq. (9), we want to construct a new RV $z \in \mathscr{X}$ to serve as an approximation to the solution of Eq. (4) or Eq. (5). This then is a *filter*, and a few possibilities will be given in Sect. 3.2.

3.1 The Posterior Distribution Measure

It was already mentioned at the beginning of Sect. 2.3.1, that the scalar function $r_{\mathcal{I}_x}(x) = \chi_{\mathcal{I}_x}(x)$ may be used to characterise the conditional probability distribution of $x \in \mathscr{X}$. Indeed, if for a RV $R(x) \in \mathscr{R}$ one defines:

$$\forall \mathcal{E} \in \mathfrak{B}_{\mathcal{R}} : \; \mathbb{P}(\mathcal{E}|\mathfrak{S}) := \mathbb{E}\left(\chi_{\mathcal{E}}(R)|\mathfrak{S}\right), \tag{38}$$

one has completely characterised the posterior distribution, a version of which is under certain conditions—[2, 16, 32]—a measure on $\mathcal{R}$, the image space of the RV R.

One may also recall that the *characteristic function* in the sense of stochastics of a RV $R \in \mathscr{R}$, namely

$$\varphi_R : \; \mathcal{R}^* \ni r^* \mapsto \varphi_R(r^*) := \mathbb{E}\left(\exp(\mathrm{i}\,\langle r^*, R\rangle_{\mathcal{R}})\right),$$

completely characterises the distribution of the RV R. As we assume that $\mathcal{R}$ is a Hilbert space, we may identify $\mathcal{R}$ with its dual space $\mathcal{R}^*$, and in this case take φ_R as defined on $\mathcal{R}$. If now a conditional expectation operator $\mathbb{E}\left(\cdot|\mathfrak{S}\right)$ is given, it may be used to define the *conditional* characteristic function $\varphi_{R|\mathfrak{S}}$:

$$\forall r \in \mathcal{R} : \; \varphi_{R|\mathfrak{S}}(r) := \mathbb{E}\left(\exp(\mathrm{i}\,\langle r, R\rangle_{\mathcal{R}})|\mathfrak{S}\right). \tag{39}$$

This again completely characterises the posterior distribution.

Another possible way, actually encompassing the previous two, is to look at all functions $\psi : \mathcal{R} \to \mathbb{R}$, and compute—when they are defined and finite—the quantities

$$\mu_\psi := \mathbb{E}\left(\psi(R)|\mathfrak{S}\right), \tag{40}$$

again completely characterising the posterior distribution. The two previous examples show that not *all* functions of R with finite conditional expectation are needed. The first example uses the set of functions

$$\{\psi \mid \psi(R) = \chi_{\mathcal{E}}(R), \mathcal{E} \in \mathfrak{B}_{\mathcal{R}}\},$$

whereas the second example uses the set

$$\{\psi \mid \psi(R) = \exp(\mathrm{i}\,\langle r, R\rangle_{\mathcal{R}}), r \in \mathcal{R}\}.$$

3.2 A Posterior Random Variable—Filtering

In the context of a situation like in Eq. (4) resp. Eq. (5), which represents the *unknown* system and state vector x_n, and where one observes y_n according to Eq. (8) resp. Eq. (9), one wants to have an estimating or *tracking* model system, with a state estimate z_n for x_n which would in principle obey Eq. (4) resp. Eq. (5) with the noise w_n set to zero—as one only knows the structure of the system as given by the maps $\hat{f}$ resp. f but not the initial condition x_0 nor the noise. The observations y_n can be used to correct the state estimate z_n, as will be shown shortly. The state estimate will be computed via Bayesian updating. But the Bayesian theory, as explained above, only characterises the posterior *distribution*; and there are many random variables which might have a given distribution. To obtain a RV z_n which can be used to predict the next state x_{n+1} through the estimate z_{n+1} one may use a filter based on Bayesian theory. The mean vehicle for this will be the notion of conditional expectation as described in the previous Sect. 2. As we will first consider only one update step, the time index n will be dropped for the sake of ease of notation: The true state is x, its *forecast* is x_f, and the forecast of the measurement is $y_f(x_f)$, whereas the observation is $\hat{y}$.

To recall, according to Definition 9, the Bayesian update is defined via the conditional expectation $\mathbb{E}\left(R(x)|\sigma(y(x))\right)$ through a measurement $y(x)$—which will for the sake of simplicity be denoted just by $\mathbb{E}\left(R(x)|y\right)$—of a $\mathcal{R}$-valued RV $R(x)$ is simply the orthogonal projection onto the subspace $\mathscr{R}_\infty$ in Eq. (24),

$$\mathbb{E}\left(R(x)|y\right) = P_{\mathscr{R}_\infty}(R(x)) = \Phi_R(y(x)),$$

which is given by the optimal map Φ_R from Eq. (26), characterised by Eq. (32), where we have added an index R to signify that this is the optimal map for the conditional expectation of the RV $R \in \mathscr{R}$.

The linearly closed subspace $\mathscr{R}_\infty$ induces a orthogonal decomposition

$$\mathscr{R} = \mathscr{R}_\infty \oplus \mathscr{R}_\infty^\perp,$$

where the orthogonal projection onto $\mathscr{R}_\infty^\perp$ is given by $I - P_{\mathscr{R}_\infty}$. Hence a RV in $\mathscr{R}$ like $R(x)$ can be decomposed accordingly as

$$\begin{aligned} R(x) &= P_{\mathscr{R}_\infty}(R(x)) + \left(I - P_{\mathscr{R}_\infty}\right)(R(x)) \\ &= \Phi_R(y(x)) + (R(x) - \Phi_R(y(x))). \end{aligned} \tag{41}$$

This Eq. (41) is the starting point for the updating. A measurement $\hat{y}$ will inform us about the component in $\mathscr{R}_\infty$, namely $\Phi_R(\hat{y})$, while we leave the component orthogonal to it unchanged: $R(x_f) - \Phi_R(y(x_f))$. Adding these two terms then gives an *updated* or *assimilated* RV $R_a \in \mathscr{R}$:

$$\begin{aligned} R_a &= \Phi_R(\hat{y}) + (R(x_f) - \Phi_R(y(x_f))) = \bar{R}_a^{|\hat{y}} + \tilde{R}_a \\ &= R(x_f) + (\Phi_R(\hat{y}) - \Phi_R(y(x_f))) = R_f + R_\infty, \end{aligned} \tag{42}$$

where $R_f = R(x_f) \in \mathscr{R}$ is the *forecast* and $R_\infty = (\Phi_R(\hat{y}) - \Phi_R(y(x_f))) \in \mathscr{R}$ is the *innovation*. For $\bar{R}_a^{|\hat{y}} = \Phi_R(\hat{y})$ and $\tilde{R}_a = R(x_f) - \Phi_R(y(x_f))$ one has the following result:

Proposition 12 *The assimilated RV R_a from Eq. (42) has the correct conditional expectation*

$$\mathbb{E}\left(R_a|y\right) = \Phi_R(\hat{y}) + \mathbb{E}\left((R(x_f) - \Phi_R(y(x_f)))|y\right) = \mathbb{E}\left(R(x_f)|\hat{y}\right) = \bar{R}_a^{|\hat{y}}, \tag{43}$$

better would be posterior *expectation—after the observation $\hat{y}$.*

Proof Observe that that the conditional expectation of the second term $\tilde{R}_a$ in Eq. (42) vanishes:

$$\begin{aligned}\mathbb{E}\left(\tilde{R}_a|y\right) &= \mathbb{E}\left((R(x) - \Phi_R(y(x)))|y\right)\\ &= P_{\mathscr{R}_\infty}(R(x) - P_{\mathscr{R}_\infty}(R(x))) = P_{\mathscr{R}_\infty}(R(x)) - P_{\mathscr{R}_\infty}(R(x)) = 0.\end{aligned}$$

This means that the conditional expectation of the second term in Eq. (43) is nought, whereas the remaining term $\Phi_R(\hat{y})$ is just $\mathbb{E}\left(R(x_f)|\hat{y}\right)$.

From Eq. (42) one can now construct filters. As often the optimal map Φ_R is often not easy to compute, one may even want to replace it by an approximation, say g_R, so that the update equation is

$$\tilde{R}_a = R(x_f) + (g_R(\hat{y}) - g_R(y(x_f))). \tag{44}$$

Whichever way, either the Eq. (42) or Eq. (44), they are composed of the following elements, the prior knowledge, which gives the prediction or forecast $R_f = R(x_f)$ for the RV R, and the correction, innovation, or update

$$R_\infty = (\Phi_R(\hat{y}) - \Phi_R(y(x_f))) \approx (g_R(\hat{y}) - g_R(y(x_f))),$$

which is the update difference between the actual observation $\hat{y}$ and the predicted or forecast observation $y(x_f)$.

3.2.1 Getting the Mean Right

The simplest function $R(x)$ to think of is the identity $R(x) := x$. This gives an update—a filter—for the RV x itself. The optimal map will be denoted by Φ_x in this case. From Eq. (42) one has:

$$x_a = x_f + (\Phi_x(\hat{y}) - \Phi_x(y(x_f))) = x_f + x_\infty, \tag{45}$$

and Proposition 12 ensures that the assimilated RV x_a has the correct conditional mean

$$\mathbb{E}\left(x_a|y\right) = \mathbb{E}\left(x_f|\hat{y}\right) = \Phi_x(\hat{y}) =: \bar{x}^{|\hat{y}}. \tag{46}$$

The Eq. (45) is the basis for many filtering algorithms, and many variations on the basic prescription are possible. Often they will be such that the property according to Proposition 12, the correct conditional mean, is only approximately satisfied. This is due to the fact that for one the Eq. (45) is an equation for RVs, which in their entirety cannot be easily handled, they are typically infinite dimensional objects and thus have to be discretised for numerical purposes.

It was also already pointed out that the optimal map Φ_x is not easy to compute, and thus approximations are used, $\Phi_x \approx g_x$, the simplest one being where $g_x = G_x \in \mathscr{L}(\mathcal{Y}, \mathcal{X})$ is taken as a linear map, leading to linear filters [3, 11]. The well-known Kalman filter (KF) [17] and its many variants and extensions—e.g. extended KF, Gauss-Markov-Kalman filter, square root KF, etc.—and simplifications—e.g. 3DVar, 4DVar, Kriging, Gaussian process emulation (GPE)—arise in this way (e.g. [1, 6, 28–30, 34, 35, 37, 41]).

As the conditional expectation of x_a in Eq. (45) is Eq. (46) $\mathbb{E}\left(x_a|\hat{y}\right) = \Phi_x(\hat{y}) = \bar{x}^{|\hat{y}}$, the zero-mean part of x_a is $\tilde{x}_a = x_f - \Phi_x(y(x_f))$. The posterior variance of the RV x_a is thus

$$C_{x_a x_a|\hat{y}} = \mathbb{E}\left(\tilde{x}_a \otimes \tilde{x}_a|\hat{y}\right) = \mathbb{E}\left((x_f - \Phi_x(y(x_f))) \otimes (x_f - \Phi_x(y(x_f)))|\hat{y}\right), \quad (47)$$

and it has been noted many times that this does not depend on the observation $\hat{y}$. Still, one may note (e.g. [41])

Proposition 13 *Assume that x_f is a Gaussian RV, that the observation $y(x_f) = \hat{h}(x_f, v)$—absorbing the scaling ε into v—is affine in x_f and in the uncorrelated Gaussian observational noise v, i.e. $v \perp x_f$ and $C_{vx_f} = 0$. Then the optimal map $\Phi_x = K_x \in \mathscr{L}(\mathcal{Y}, \mathcal{X})$ is linear, and the updated or assimilated RV x_a from Eq. (45) is also Gaussian, and has the* correct *posterior distribution, characterised by the mean Eq. (46), $\bar{x}^{|\hat{y}}$, and the covariance Eq. (47). Setting $w = \hat{h}(x_f, v) := H(x_f) + v$ with $H \in \mathscr{L}(\mathcal{X}, \mathcal{Y})$, one obtains from Eq. (47)*

$$\begin{aligned} C_{x_a x_a|\hat{y}} = \mathbb{E}\left(\tilde{x}_a \otimes \tilde{x}_a|\hat{y}\right) &= C_{x_f x_f} - K_x C_{wx_f} - C_{wx_f}^T K_x^T + K_x C_{ww} K_x^T \\ &= (I - K_x H) C_{x_f x_f} (I - K_x H)^T + K_x C_{vv} K_x^T \end{aligned} \quad (48)$$

for the covariance, and for the mean

$$\bar{x}^{|\hat{y}} = \mathbb{E}\left(x_a|\hat{y}\right) = K_x \hat{y}. \quad (49)$$

Proof As this is a well known result, we only show the connection of Eq. (48) with Eq. (47). Note that

$$\begin{aligned} \tilde{x}_a &= x_f - \Phi_x(y(x_f)) = x_f - K_x(\hat{h}(x_f, v)) \\ &= x_f - K_x(w) = (I - K_x H) x_f - K_x v. \end{aligned}$$

This gives the Eq. (48), and the Eq. (49) follows directly from Eq. (46).

This means that in the purely linear Gaussian case described in Proposition 13 a RV with the correct posterior distribution is given simply by the process of projection.

In the context of the dynamical system Eq. (4) resp. Eq. (5), where the measurement is denoted by y_{n+1}, the update for the tracking equation is

$$z_{n+1} = \hat{f}(z_n, 0, n) + (\Phi_x(y_{n+1}) - \Phi_x(\hat{h}(\hat{f}(z_n, 0, n), 0))) \tag{50}$$

for the case Eq. (4), resp. for Eq. (5)

$$z_{n+1} = f(z_n) + (\Phi_x(y_{n+1}) - \Phi_x(h(f(z_n)))). \tag{51}$$

Again the assimilated or updated state estimate $x_a := z_{n+1}$ is composed of two components, the prediction or forecast $x_f := \hat{f}(z_n, 0, n)$ resp. $x_f := f(z_n)$, and the correction, innovation, or update

$$x_\infty := (\Phi_x(y_{n+1}) - \Phi_x(\hat{h}(\hat{f}(z_n, 0, n), 0)))$$

resp. $x_\infty := (\Phi_x(y_{n+1}) - \Phi_x(h(f(z_n))))$, which takes into account the difference resulting from the actual measurement $\hat{y} := y_{n+1}$ and the forecast measurement $\hat{h}(\hat{f}(z_n, 0, n), 0)$ resp. $h(f(z_n))$.

If the optimal map Φ_x is not easy to compute, one may want to replace it by an approximation as in Eq. (44), say g, so that for example the Eq. (51) would read

$$z_{n+1} = f(z_n) + (g(y_{n+1}) - g(h(f(z_n)))) = (f - g \circ h \circ f)(z_n) + g(y_{n+1}), \tag{52}$$

where one may hope to show that if the map $(f - g \circ h \circ f)$ is a contraction, the difference $x_n - z_n$ will decrease as $n \to \infty$ [19]. Many variations on this theme exist, especially in the case where both the observation map h and the update operator g are linear [38, 42].

3.2.2 Getting Also the Covariance Right

An approximation to a RV which has the required posterior distribution was constructed in Sect. 3.2.1, where at least the mean was correct. One may now go a step further and also get the correct posterior covariance. As a starting point take the *assimilated* RV x_a from Eq. (45) that has the correct conditional mean $\bar{x}^{|\hat{y}}$ from Eq. (46), but the covariance, from Eq. (47), is $C_{x_a x_a|\hat{y}} = \mathbb{E}\left(\tilde{x}_a \otimes \tilde{x}_a|\hat{y}\right)$. To get the covariance and the mean right, we compute what the correct posterior covariance should be, by computing the optimal map for $R(x) := x \otimes x$. This gives for the posterior correlation

$$\hat{C}_p := \mathbb{E}\left(R(x_f)|\hat{y}\right) = \mathbb{E}\left((x_f \otimes x_f)|\hat{y}\right) = \Phi_{x\otimes x}(\hat{y}), \tag{53}$$

so that the posterior covariance is

$$C_p := \hat{C}_p - \bar{x}^{|\hat{y}} \otimes \bar{x}^{|\hat{y}} = \Phi_{x\otimes x}(\hat{y}) - \Phi_x(\hat{y}) \otimes \Phi_x(\hat{y}). \tag{54}$$

Proposition 14 *A new RV x_c with the correct posterior covariance Eq. (54) is built from $x_a = \bar{x}^{|\hat{y}} + \tilde{x}_a$ in Eq. (45) by taking*

$$x_c := \bar{x}^{|\hat{y}} + C_p^{1/2} C_{x_a x_a|\hat{y}}^{-1/2} \tilde{x}_a. \tag{55}$$

Proof As $\mathbb{E}\left(x_c|\hat{y}\right) = \bar{x}^{|\hat{y}}$, one has

$$\begin{aligned} C_{x_c x_c} &= \mathbb{E}\left((C_p^{1/2}\, C_{x_a x_a|\hat{y}}^{-1/2}\, \tilde{x}_a) \otimes (C_p^{1/2}\, C_{x_a x_a|\hat{y}}^{-1/2}\, \tilde{x}_a)|\hat{y}\right) \\ &= C_p^{1/2}\, C_{x_a x_a|\hat{y}}^{-1/2}\, \mathbb{E}\left(\tilde{x}_a \otimes \tilde{x}_a|\hat{y}\right)\, C_{x_a x_a|\hat{y}}^{-1/2}\, C_p^{1/2} \\ &= C_p^{1/2}\, C_{x_a x_a|\hat{y}}^{-1/2}\, C_{x_a x_a|\hat{y}}\, C_{x_a x_a|\hat{y}}^{-1/2}\, C_p^{1/2} = C_p^{1/2}\, C_p^{1/2} = C_p, \end{aligned}$$

proving that the RV x_c in Eq. (55) has the correct posterior covariance.

Having achieved a RV which has the correct posterior mean and covariance, it is conceivable to continue in this fashion, building RVs which match the posterior better and better. A similar idea, but from a different starting point, is used in [27, 31]. In the future, it is planned to combine these two approaches.

3.3 Approximations

In any actual inverse computations several kinds of approximations are usually necessary. Should one pursue an approach of sampling form the posterior distribution Eq. (11) in Sect. 2.2 for example, then typically a sampling and a quantisation or binning approximation is performed, often together with some kernel-estimate of the density. All of these processes usually introduce approximation errors. Here we want to use methods based on the conditional expectation, which were detailed in Sect. 2.3.

Looking for example at Theorems 10 and 11 in Sect. 2.3.2, one has to work in the usually infinite dimensional space $\mathscr{R} = \mathcal{R} \otimes \mathcal{S}$ from Eq. (22) and its subspace $\mathscr{R}_\infty = \mathcal{R} \otimes \mathcal{S}_\infty$ from Eq. (23), to minimise the functional in Eq. (27) to find the optimal map representing the conditional expectation for a desired function $R(x)$, Eqs. (26) and (29), Φ in the space $L_0(\mathcal{Y}, \mathcal{R})$. Then one has to construct a RV whose distribution my be characterised by the conditional expectation, to represent the posterior measure. Approximations in this latter respect were discussed in Sect. 3.2. The space $\mathscr{R}_\infty$ is computationally accessible via $L_0(\mathcal{Y}, \mathcal{R})$, which has to be approximated by some finite dimensional subspace. This will be discussed in this Sect. 3.3. Furthermore, the component spaces of $\mathscr{R} = \mathcal{R} \otimes \mathcal{S}$ are also typically infinite dimensional, and have in actual computations to be replaced by finite dimensional subspaces. This topic will be sketched in Sect. 4.

Computationally we will not be able to deal with the *whole* space $\mathscr{R}_\infty$, so we look at the effect of approximations. Assume that $L_0(\mathcal{Y}, \mathcal{R})$ in Eq. (29) or Eq. (30) is approximated by subspaces $L_{0,m} \subset L_0(\mathcal{Y}, \mathcal{R})$ with $\mathscr{L}(\mathcal{Y}, \mathcal{R}) \subseteq L_{0,m}$, where $m \in \mathbb{N}$ is a parameter describing the level of approximation and $L_{0,m} \subset L_{0,k}$ if $m < k$, such that the subspaces

$$\mathscr{R}_m = \{\varphi(y) \mid \varphi \in L_{0,m};\ \varphi(\hat{h}(x, \varepsilon v)) \in \mathscr{R}\} \subseteq \mathscr{R}_\infty \tag{56}$$

are linearly closed and their union is dense

$$\overline{\bigcup_m \mathscr{R}_m} = \mathscr{R}_\infty, \tag{57}$$

a consistency condition.

To obtain results for the situation where the projection $P_{\mathscr{R}_\infty}$ is replaced by the orthogonal projection $P_{\mathscr{R}_m}$ of $\mathscr{R}$ onto $\mathscr{R}_m$, all that is necessary is to reformulate the Theorems 10 and 11.

Theorem 15 *The orthogonal projection $P_{\mathscr{R}_m}$ of the RV $R(x) \in \mathscr{R}$ is characterised by:*

$$P_{\mathscr{R}_m}(R(x)) := \arg\min_{\tilde{R}\in\mathscr{R}_m} \frac{1}{2}\|R(x) - \tilde{R}\|^2_{\mathscr{R}}, \tag{58}$$

The Eq. (56) implies the existence of a optimal map $\varPhi_m \in L_{0,m}(\mathcal{Y}, \mathcal{R})$ such that

$$P_{\mathscr{R}_m}(R(x)) = \varPhi_m(\hat{h}(x, \varepsilon v)). \tag{59}$$

Taking the vanishing of the first variation/Gâteaux derivative of the loss-function as a necessary condition for a minimum leads to a simple geometrical interpretation: the difference between the original vector-valued RV $R(x)$ and its projection has to be perpendicular to the subspace $\mathscr{R}_m$: $\forall \tilde{R} \in \mathscr{R}_m$:

$$\langle\!\langle R(x) - P_{\mathscr{R}_m}(R(x)), \tilde{R}\rangle\!\rangle_{\mathscr{R}} = 0, \quad \textit{i.e. } R(x) - P_{\mathscr{R}_m}(R(x)) \in \mathscr{R}_m^\perp. \tag{60}$$

Rephrasing Eq. (58) with account to Eq. (60) leads for the optimal map $\varPhi_m \in L_{0,n}(\mathcal{Y}, \mathcal{R})$ to

$$P_{\mathscr{R}_m}(R(x)) = \varPhi_m(\hat{h}(x, \varepsilon v)) := \arg\min_{\varphi\in L_{0,m}(\mathcal{Y},\mathcal{R})} \|R(x) - \varphi(\hat{h}(x, \varepsilon v))\|^2_{\mathscr{R}}, \tag{61}$$

and the orthogonality condition of Eq. (60) leads to

$$\forall\varphi \in L_{0,m}(\mathcal{Y}, \mathcal{R}) : \ \langle\!\langle R(x) - \varPhi_m(\hat{h}(x, \varepsilon v)), \varphi(\hat{h}(x, \varepsilon v))\rangle\!\rangle_{\mathscr{R}} = 0. \tag{62}$$

In addition, as $\mathscr{R}_m$ is linearly closed, one obtains the useful statement

$$\forall\tilde{R} \in \mathscr{R}_m : \ \mathscr{L}(\mathcal{R}) \ni \mathbb{E}\left((R(x) - P_{\mathscr{R}_m}(R(x))) \otimes \tilde{R}\right) = 0. \tag{63}$$

or rephrased $\forall\varphi \in L_{0,m}(\mathcal{Y}, \mathcal{R})$:

$$\mathscr{L}(\mathcal{R}) \ni \mathbb{E}\left((R(x) - \varPhi_m(\hat{h}(x, \varepsilon v))) \otimes \varphi(\hat{h}(x, \varepsilon v))\right) = 0. \tag{64}$$

There is a unique minimiser $P_{\mathscr{R}_m}(R(x)) \in \mathscr{R}_m$ to the problem in Eq. (58), which satisfies the orthogonality *condition Eq. (60), which is equivalent to the the* strong orthogonality condition *Eq. (63). The minimiser is unique as an element of $\mathscr{R}_m$, but the mapping $\Phi_m \in L_{0,m}(\mathcal{Y}, \mathcal{R})$ in Eq. (61) may not necessarily be. It also holds that*

$$\|P_{\mathscr{R}_m}(R(x))\|^2_{\mathscr{R}} = \|R(x)\|^2_{\mathscr{R}} - \|R(x) - P_{\mathscr{R}_m}(R(x))\|^2_{\mathscr{R}}. \tag{65}$$

Additionally, one has

$$\|P_{\mathscr{R}_m}(R(x))\|^2_{\mathscr{R}} \leq \|P_{\mathscr{R}_\infty}(R(x))\|^2_{\mathscr{R}}. \tag{66}$$

Proof It is all already contained in Theorems 10 and 11 when applied to $\mathscr{R}_m$. The stability condition Eq. (66) is due to the simple fact that $\mathscr{R}_m \subseteq \mathscr{R}_\infty$.

From the consistency condition, the stability Eq. (66) as shown in Theorem 15, and *Céa's* lemma, one immediately obtains:

Theorem 16 *For all RVs $R(x) \in \mathscr{R}$, the sequence $R_m := P_{\mathscr{R}_m}(R(x))$ converges to $R_\infty := P_{\mathscr{R}_\infty}(R(x))$:*

$$\lim_{m \to \infty} \|R_\infty - R_m\|^2_{\mathscr{R}} = 0. \tag{67}$$

Proof Well-posedness is a direct consequence of Theorem 11. As the $P_{\mathscr{R}_m}$ are orthogonal projections onto the subspaces $\mathscr{R}_m$, their norms are hence all equal to unity—a stability condition, as shown in Eq. (66). Application of Céa's lemma then directly yields Eq. (67).

3.3.1 Approximation by Polynomials

Here we choose the subspaces of polynomials up to degree m for the purpose of approximation, i.e.

$$\mathscr{R}_m := \overline{\text{span}}\{\varphi \mid \varphi(\hat{h}(x, \varepsilon v)) \in \mathscr{R}, \quad \varphi \in \mathcal{P}_m(\mathcal{Y}, \mathcal{X})\},$$

where $\mathcal{P}_m(\mathcal{Y}, \mathcal{X}) \subset L_0(\mathcal{Y}, \mathcal{X})$ are the polynomials of degree at most m on $\mathcal{Y}$ with values in $\mathcal{X}$. We may write $\psi_m \in \mathcal{P}_m(\mathcal{Y}, \mathcal{X})$ as

$$\psi_m(y) := {}^0H + {}^1H\,y + \cdots + {}^kH\,y^{\vee k} + \cdots + {}^mH\,y^{\vee m}, \tag{68}$$

where ${}^kH \in \mathscr{L}^k_s(\mathcal{Y}, \mathcal{R})$ is symmetric and k-linear; and $y^{\vee k} := \overbrace{y \vee \ldots \vee y}^{k} :=$ Sym $(y^{\otimes k})$ is the symmetric tensor product of the y's taken k times with itself. Let us remark here that the form of Eq. (68), given in monomials, is numerically not a good form—except for very low m—and straightforward use in computations is not recommended. The relation Eq. (68) could be re-written in some orthogonal polynomials—or in fact any other system of multi-variate functions; this generalisation will be considered in Sect. 3.3.3. For the sake of conceptual simplicity, we stay with Eq. (68) and then have that for any RV $R(x) \in \mathscr{R}$

$$\Phi_{R,m}(R(x)) := \psi_{R,m}(y) := {}^0H + \cdots + \cdots + {}^mH\, y^{\vee m} =: \Psi_{R,m}({}^0H, \ldots, {}^mH) \tag{69}$$

the optimal map in Eq. (59) from Theorem 15—where we have added an index R to indicate that it depends on the RV $R(x)$, but for simplicity omitted this index on the coefficient maps ${}^kH$—is a function $\Psi_{R,m}$ of the coefficient maps ${}^kH$. The stationarity or orthogonality condition Eq. (64) can then be written in terms of the ${}^kH$. We need the following abbreviations for any $k, \ell \in \mathbb{N}_0$ and $p \in \mathscr{R}$, $v \in \mathscr{Y}$:

$$\langle p \otimes v^{\vee k}\rangle := \mathbb{E}\left(p \otimes v^{\vee k}\right) = \int_\Omega p(\omega) \otimes v(\omega)^{\vee k}\, \mathbb{P}(\mathrm{d}\omega)$$

and

$${}^kH\langle y^{\vee(\ell+k)}\rangle := \langle y^{\vee \ell} \vee ({}^kH\, y^{\vee k})\rangle = \mathbb{E}\left(y^{\vee \ell} \vee ({}^kH\, y^{\vee k})\right).$$

We may then characterise the ${}^kH$ in the following way:

Theorem 17 *With $\Psi_{R,m}$ from Eq. (69), the stationarity condition Eq. (64) becomes, by the chain rule, for any $m \in \mathbb{N}_0$*

$$\forall \ell = 0, \ldots, m: \quad \sum_{k=0}^{m} {}^kH\langle y^{\vee(\ell+k)}\rangle = \langle R(x) \otimes y^{\vee \ell}\rangle. \tag{70}$$

The Hankel operator matrix $(\langle y^{\vee(\ell+k)}\rangle)_{\ell,k}$ in the linear equations (70) is symmetric and positive semi-definite, hence the system Eq. (70) has a solution, unique in case the operator matrix is actually definite.

Proof The relation Eq. (70) is the result of straightforward application of the chain rule to the Eq. (64).

The symmetry of the operator matrix is obvious—the $\langle y^{\vee k}\rangle$ are the coefficients—and positive semi-definiteness follows easily from the fact that it is the gradient of the functional in Eq. (61), which is convex.

Observe that the operator matrix is independent of the RV $R(x)$ for which the computation is performed. Only the right hand side is influenced by $R(x)$.

The system of operator equations Eq. (70) may be written in more detailed form as:

$$\begin{array}{lllll}
\ell = 0: \; {}^0H & \cdots + {}^kH\langle y^{\vee k}\rangle & \cdots + {}^mH\langle y^{\vee m}\rangle = & & \langle R(x)\rangle, \\
\ell = 1: \; {}^0H\langle y\rangle & \cdots + {}^kH\langle y^{\vee(1+k)}\rangle & \cdots + {}^mH\langle y^{\vee(1+m)}\rangle = & & \langle R(x) \otimes y\rangle, \\
\vdots & \quad \ldots & \quad \vdots & & \vdots \\
\ell = m: \; {}^0H\langle y^{\vee m}\rangle & \cdots + {}^kH\langle y^{\vee(k+m)}\rangle & \cdots + {}^mH\langle y^{\vee 2m}\rangle = & & \langle R(x) \otimes y^{\vee m}\rangle.
\end{array}$$

Using '*symbolic index*' notation a la Penrose—the reader may just think of indices in a finite dimensional space with orthonormal basis—the system Eq. (70) can be given yet another form: denote in symbolic index notation $R(x) = (R^\imath)$, $y = (y^\jmath)$, and ${}^kH = ({}^kH^\imath_{\jmath_1\cdots\jmath_k})$, then Eq. (70) becomes, with the use of the Einstein convention of summation (a tensor contraction) over repeated indices, and with the symmetry explicitly indicated:

$$\begin{aligned} &\forall \ell = 0, \dots, m;\ \jmath_1 \le \dots \le \jmath_\ell \le \dots \le \jmath_{\ell+k} \le \dots \le \jmath_{\ell+m} : \\ &\quad \langle y^{\jmath_1} \cdots y^{\jmath_\ell} \rangle ({}^0H^\imath) + \dots + \langle y^{\jmath_1} \cdots y^{\jmath_{\ell+1}} \cdots y^{\jmath_{\ell+k}} \rangle ({}^kH^\imath_{\jmath_{\ell+1}\cdots\jmath_{\ell+k}}) \\ &\qquad + \dots + \langle y^{\jmath_1} \cdots y^{\jmath_{\ell+1}} \cdots y^{\jmath_{\ell+m}} \rangle ({}^mH^\imath_{\jmath_{\ell+1}\cdots\jmath_{\ell+m}}) = \langle R^\imath y^{\jmath_1} \cdots y^{\jmath_\ell} \rangle. \end{aligned} \tag{71}$$

We see in this representation that the matrix does *not* depend on $\imath$—it is identically *block diagonal* after appropriate reordering, which makes the solution of Eq. (70) or Eq. (71) much easier.

Some special cases are: for $m = 0$—*constant* functions. One does not use any information from the measurement—and from Eq. (70) or Eq. (71) one has

$$\Phi_{R,0}(R(x)) = \psi_{R,0}(y) = {}^0H = \langle R \rangle = \mathbb{E}(R) = \bar{R}.$$

Without any information, the conditional expectation is equal to the unconditional expectation. The update corresponding to Eq. (42)—actually Eq. (44) as we are approximating the map Φ_R by $g_R = \Phi_{R,0}$—then becomes $R_a \approx R_{a,0} = R(x_f) = R_f$, as $R_\infty = 0$ in this case; the assimilated quantity stays equal to the forecast. This was to be expected, is not of much practical use, but is a consistency check.

The case $m = 1$ in Eq. (70) or Eq. (71) is more interesting, allowing up to *linear* terms:

$$\begin{aligned} {}^0H \quad\ + {}^1H\langle y \rangle \quad\ &= \langle R(x) \rangle = \bar{R} \\ {}^0H\langle y \rangle + {}^1H\langle y \vee y \rangle &= \langle R(x) \otimes y \rangle. \end{aligned}$$

Remembering that $C_{Ry} = \langle R(x) \otimes y \rangle - \langle R \rangle \otimes \langle y \rangle$ and analogous for C_{yy}, one obtains by tensor multiplication with $\langle R(x) \rangle$ and symbolic Gaussian elimination

$$\begin{aligned} {}^0H &= \langle R \rangle - {}^1H\langle y \rangle = \bar{R} - {}^1H\bar{y} \\ {}^1H(\langle y \vee y \rangle - \langle y \rangle \vee \langle y \rangle) = {}^1HC_{yy} &= \langle R(x) \otimes y \rangle - \langle R \rangle \otimes \langle y \rangle = C_{Ry}. \end{aligned}$$

This gives

$${}^1H = C_{Ry}C_{yy}^{-1} =: K \tag{72}$$

$${}^0H = \bar{R} - K\bar{y}. \tag{73}$$

where K in Eq. (72) is the well-known *Kalman* gain operator [17], so that finally

$$\Phi_{R,1}(R(x)) = \psi_{R,1}(y) = {}^0H + {}^1Hy = \bar{R} + C_{Ry}C_{yy}^{-1}(y - \bar{y}) = \bar{R} + K(y - \bar{y}). \tag{74}$$

The update corresponding to Eq. (42)—again actually Eq. (44) as we are approximating the map Φ_R by $g_R = \Phi_{R,1}$—then becomes

$$\begin{aligned} R_a \approx R_{a,1} &= R(x_f) + \big((\bar{R} + K(\hat{y} - \bar{y})) - (\bar{R} + K(y(x_f) - \bar{y}))\big) \\ &= R_f + K(\hat{y} - y(x_f)) = R_f + R_{\infty,1}. \end{aligned} \tag{75}$$

This may be called a *linear* Bayesian update (LBU), and is similar to the 'Bayes linear' approach [11]. It is important to see Eq. (75) as a symbolic expression, especially the inverse C_{yy}^{-1} indicated in Eq. (74) should not really be computed, especially when C_{yy} is ill-conditioned or close to singular. The inverse can in that case be replaced by the *pseudo-inverse*, or rather the computation of K, which is in linear algebra terms a *least-squares* approximation, should be done with orthogonal transformations and not by elimination. We will not dwell on these well-known matters here. It is also obvious that the constant term in Eq. (74)—or even Eq. (69) for that matter—is of no consequence for the update filter, as it cancels out.

The case $m = 2$ can still be solved symbolically, the system to be solved is from Eq. (70) or Eq. (71):

$$\begin{aligned} {}^0H \quad &+ {}^1H\langle y\rangle \quad + {}^2H\langle y^{\vee 2}\rangle = \langle R\rangle \\ {}^0H\langle y\rangle \quad &+ {}^1H\langle y^{\vee 2}\rangle + {}^2H\langle y^{\vee 3}\rangle = \langle R \otimes y\rangle \\ {}^0H\langle y^{\vee 2}\rangle &+ {}^1H\langle y^{\vee 3}\rangle + {}^2H\langle y^{\vee 4}\rangle = \langle R \otimes y^{\vee 2}\rangle. \end{aligned}$$

After some symbolic elimination steps one obtains

$$\begin{aligned} {}^0H + {}^1H\langle y\rangle + {}^2H\langle y^{\vee 2}\rangle &= \bar{R} \\ 0 \quad + {}^1H \quad + {}^2H\,\boldsymbol{F} &= K \\ 0 \quad + 0 \quad + {}^2H\,\boldsymbol{G} &= E, \end{aligned}$$

with the Kalman gain operator $K \in (\mathcal{R} \otimes \mathcal{Y})^*$ from Eq. (72), the third order tensors $\boldsymbol{F} \in (\mathcal{Y}^{\otimes 3})^*$ given in Eq. (76), and $E \in (\mathcal{R} \otimes \mathcal{Y}^{\otimes 2})^*$ given in Eq. (77), and the fourth order tensor $\boldsymbol{G} \in (\mathcal{Y}^{\otimes 4})^*$ given in Eq. (78):

$$\boldsymbol{F} = \big(\langle y^{\vee 3}\rangle - \langle y^{\vee 2}\rangle \vee \langle y\rangle\big)\, C_{yy}^{-1}, \tag{76}$$

$$E = \langle R \otimes y^{\vee 2}\rangle - \bar{R} \otimes \langle y^{\vee 2}\rangle - K\left(\langle y^{\vee 3}\rangle - \langle y\rangle \vee \langle y^{\vee 2}\rangle\right) \tag{77}$$

$$\boldsymbol{G} = \left(\langle y^{\vee 4}\rangle - \langle y^{\vee 2}\rangle^{\vee 2}\right) - \boldsymbol{F} \cdot \left(\langle y^{\vee 3}\rangle - \langle y\rangle \vee \langle y^{\vee 2}\rangle\right), \tag{78}$$

where the single central dot '·' denotes as usual a contraction over the appropriate indices, and a colon ':' a double contraction. From this one easily obtains the solution

$${}^2H = E : \boldsymbol{G}^{-1} \tag{79}$$

$${}^1H = K - {}^2H\,\boldsymbol{F} \tag{80}$$

$$^0H = \bar{R} - (K - {}^1H)\bar{y} - {}^2H\langle y^{\vee 2}\rangle = \bar{R} - {}^2H(\boldsymbol{F}\cdot\bar{y} + \langle y^{\vee 2}\rangle). \tag{81}$$

The update corresponding to Eq. (42)—again actually Eq. (44) as we are approximating the map Φ_R now by $g_R = \Phi_{R,2}$—then becomes

$$\begin{aligned} R_a \approx R_{a,2} &= R(x_f) + \left(({}^2H\,\hat{y}^{\vee 2} + {}^1H\,\hat{y}) - ({}^2H\,y(x_f)^{\vee 2} + {}^1H\,y(x_f))\right) \\ &= R_f + \left(\mathsf{E} : \boldsymbol{G}^{-1} : \left(\hat{y}^{\vee 2} - y(x_f)^{\vee 2}\right) + (K - \mathsf{E} : \boldsymbol{G}^{-1} : \boldsymbol{F})(\hat{y} - y(x_f))\right) \\ &= R_f + R_{\infty,2}. \end{aligned} \tag{82}$$

This may be called a *quadratic* Bayesian update (QBU), and it is clearly an extension of Eq. (75).

3.3.2 The Gauss-Markov-Kalman Filter

The $m = 1$ version of Theorem 17 is well-known for the special case $R(x) := x$, and we rephrase this generalisation of the well-known *Gauss-Markov* theorem from [20, Chap. 4.6, Theorem 3]:

Proposition 18 *The update $x_{a,1}$, minimising $\|x_f - \cdot\|^2_{\mathscr{X}}$ over all elements generated by affine mappings (the up to $m = 1$ case of Theorem 17) of the measurement $\hat{y}$ with predicted measurement $y(x_f)$ is given*

$$x_{a,1} = x_f + K(\hat{y} - y(x_f)), \tag{83}$$

where the operator K is the Kalman gain from Eqs. (72) and (75).

The Eq. (83) is reminiscent—actually an extension—not only of the well-known *Gauss-Markov* theorem [20], but also of the *Kalman* filter [17, 30], so that we propose to call Eq. (83) the **Gauss-Markov-Kalman** (GMK) filter (GMKF).

We point out that $x_{a,1}$, x_f, and $y(x_f)$ are RVs, i.e. Eq. (83) is an equation in $\mathscr{X} = \mathcal{X} \otimes \mathcal{S}$ between RVs, whereas the traditional Kalman filter is an equation in $\mathcal{X}$. If the mean is taken in Eq. (83), one obtains the familiar Kalman filter formula [17] for the update of the mean, and one may show [28] that Eq. (83) also contains the Kalman update for the covariance by computing Eq. (47) for this case, which gives the familiar result of Kalman, i.e. the Kalman filter is a low-order part of Eq. (83).

The computational strategy for a typical filter is now to replace and approximate the—only abstractly given—computation of x_a Eq. (45) by the practically possible calculation of $x_{a,m}$ as in Eq. (69). This means that we approximate x_a by $x_{a,m}$ by using $\mathscr{X}_m \subseteq \mathscr{X}_\infty$, and rely on Theorem 16. This corresponds to some loss of information from the measurement as one uses a smaller subspace for the projection, but yields a manageable computation. If the assumptions of Theorem 16 are satisfied, then one can expect for m large enough that the terms in Eq. (69) converge to zero, thus providing an error indicator on when a sufficient accuracy has been reached.

3.3.3 Approximation by General Functions

The derivation in Sect. 3.3.1 was for the special case where polynomials are used to find a subspace $L_{0,m}(\mathcal{Y}, \mathcal{X})$ for the approximation. It had the advantage of showing the connection to the 'Bayes linear' approach [11], to the Gauss-Markov theorem [20], and to the *Kalman* filter [17, 30], giving in Eq. (83) of Proposition 18 the *Gauss-Markov-Kalman* filter (GMKF).

But for a more general approach not limited to polynomials, we proceed similarly as in Eq. (56), but now concretely assume a set of linearly independent functions, not necessarily orthonormal,

$$\mathcal{B} := \{\psi_\alpha \mid \alpha \in \mathcal{A},\ \psi_\alpha \in L_0(\mathcal{Y});\ \psi_\alpha(\hat{h}(x, \varepsilon v)) \in \mathcal{S}\} \subseteq \mathcal{S}_\infty \tag{84}$$

where $\mathcal{A}$ is some countable index set. Assume now that

$$\mathscr{S}_\infty = \overline{\text{span}}\ \mathcal{B},$$

i.e. $\mathcal{B}$ is a Hilbert basis of $\mathcal{S}_\infty$, again a consistency condition.

Denote by $\mathcal{A}_k$ a finite part of $\mathcal{A}$ of cardinality k, such that $\mathcal{A}_k \subset \mathcal{A}_\ell$ for $k < \ell$ and $\bigcup_k \mathcal{A}_k = \mathcal{A}$, and set

$$\mathscr{R}_k := \mathcal{R} \otimes \mathcal{S}_k \subseteq \mathscr{R}_\infty, \tag{85}$$

where the finite dimensional and hence closed subspaces $\mathcal{S}_k$ are given by

$$\mathcal{S}_k := \text{span}\{\psi_\alpha \mid \alpha \in \mathcal{A}_k,\ \psi_\alpha \in \mathcal{B}\} \subseteq \mathcal{S}. \tag{86}$$

Observe that the spaces $\mathscr{R}_k$ from Eq. (85) are linearly closed according to Proposition 4.

Theorems 15 and 16 apply in this case. For a RV $R(x) \in \mathscr{R}$ we make the following 'ansatz' for the optimal map $\Phi_{R,k}$ such that $P_{\mathscr{R}_k}(R(x)) = \Phi_{R,k}(\hat{h}(x, \varepsilon v))$:

$$\Phi_{R,k}(y) = \sum_{\alpha \in \mathcal{A}_k} v_\alpha \psi_\alpha(y), \tag{87}$$

with as yet unknown coefficients $v_\alpha \in \mathcal{R}$. This is a normal *Galerkin*-ansatz, and Eq. (64) from Theorem 15 can be used to determine these coefficients.

Take $\mathcal{Z}_k := \mathbb{R}^{\mathcal{A}_k}$ with canonical basis $\{\boldsymbol{e}_\alpha \mid \alpha \in \mathcal{A}_k\}$, and let

$$\boldsymbol{G}_k := (\langle \psi_\alpha(y(x)), \psi_\beta(y(x)) \rangle_{\mathcal{S}})_{\alpha,\beta \in \mathcal{A}_k} \in \mathscr{L}(\mathcal{Z}_k)$$

be the symmetric positive definite Gram matrix of the basis of $\mathcal{S}_k$; also set

$$\boldsymbol{v} := \sum_{\alpha \in \mathcal{A}_k} \boldsymbol{e}_\alpha \otimes v_\alpha \in \mathcal{Z}_k \otimes \mathcal{R},$$

$$\boldsymbol{r} := \sum_{\alpha \in \mathcal{A}_k} \boldsymbol{e}_\alpha \otimes \mathbb{E}\left(\psi_\alpha(y(x))R(x)\right) \in \mathcal{Z}_k \otimes \mathcal{R}.$$

Theorem 19 *For any $k \in \mathbb{N}$, the coefficients $\{v_\alpha\}_{\alpha \in \mathcal{A}_k}$ of the optimal map $\Phi_{R,k}$ in Eq. (87) are given by the unique solution of the Galerkin equation*

$$(\boldsymbol{G}_k \otimes I_{\mathcal{R}})\boldsymbol{v} = \boldsymbol{r}. \tag{88}$$

It has the formal solution

$$\boldsymbol{v} = (\boldsymbol{G}_k \otimes I_{\mathcal{R}})^{-1}\boldsymbol{r} = (\boldsymbol{G}_k^{-1} \otimes I_{\mathcal{R}})\boldsymbol{r} \in \mathcal{Z}_k \otimes \mathcal{R}.$$

Proof The Galerkin Eq. (88) is a simple consequence of Eq. (64) from Theorem 15. As the Gram matrix $\boldsymbol{G}_k$ and the identity $I_{\mathcal{R}}$ on $\mathcal{R}$ are positive definite, so is the tensor operator $(\boldsymbol{G}_k \otimes I_{\mathcal{R}})$, with inverse $(\boldsymbol{G}_k^{-1} \otimes I_{\mathcal{R}})$.

As in Eq. (71), the block structure of the equations is clearly visible. Hence, to solve Eq. (88), one only has to deal with the 'small' matrix $\boldsymbol{G}_n$.

The update corresponding to Eq. (42)—again actually Eq. (44) as we are approximating the map Φ_R now by a new map $g_R = \Phi_{R,k}$—then becomes

$$R_a \approx R_{a,k} = R(x_f) + \left(\Phi_{R,k}(\hat{y}) - \Phi_{R,k}(y(x_f))\right) = R_f + R_{\infty,k}. \tag{89}$$

This may be called a 'general Bayesian update'. Applying Eq. (89) now again to the special case $R(x) := x$, one obtains a possibly nonlinear filter based on the basis $\mathcal{B}$:

$$x_a \approx x_{a,k} = x_f + \left(\Phi_{x,k}(\hat{y}) - \Phi_{x,k}(y(x_f))\right) = x_f + x_{\infty,k}. \tag{90}$$

In case the $\mathcal{Y}^* \subseteq \text{span}\{\psi_\alpha\}_{\alpha \in \mathcal{A}_k}$, i.e. the basis generates all the linear functions on $\mathcal{Y}$, this is a true extension of the Kalman filter.

4 Numerical Realisation

In the instances where we want to employ the theory detailed in the previous Sects. 2 and 3, the spaces $\mathcal{U}$ and $\mathcal{Q}$ and hence $\mathcal{X}$ are usually infinite dimensional, as is the space $\mathcal{S} = L_2(\Omega)$. For an actual computation they all have to be discretised or approximated by finite dimensional subspaces.

In our examples we will chose finite element discretisations for $\mathcal{U}$, $\mathcal{Q}$, and hence $\mathcal{X}$, and corresponding subspaces. Hence let $\mathcal{X}_M := \text{span}\,\{\varrho_m \; : m = 1, \dots, M\} \subset \mathcal{X}$ be an M-dimensional subspace with basis $\{\varrho_m\}_{m=1}^M$. An element of $\mathcal{X}_M$ will be represented by the vector $\boldsymbol{x} = [x^1, \dots, x^M]^T \in \mathbb{R}^M$ such that $\sum_{m=1}^M x^m \varrho_m \in \mathcal{X}_M$. To avoid a profusion of notations, the corresponding random vector in $\mathbb{R}^M \otimes \mathcal{S}$—a mapping $\Omega \to \mathbb{R}^M \cong \mathcal{X}_M$—will also be denoted by $\boldsymbol{x}$, as the meaning will be clear from the context.

The norm $\|\boldsymbol{x}\|_M$ one has to take on $\mathbb{R}^M$ results from the inner product $\langle \boldsymbol{x}_1|\boldsymbol{x}_2\rangle_M := \boldsymbol{x}_1^T \boldsymbol{Q}\boldsymbol{x}_2$ with $\boldsymbol{Q} = (\langle \varrho_m|\varrho_n\rangle_{\mathcal{X}})$, the Gram matrix of the basis. We will later choose an orthonormal basis, so that $\boldsymbol{Q} = \boldsymbol{I}$ is the identity matrix. Similarly, on $\mathscr{X}_M = \mathbb{R}^M \otimes \mathcal{S}$ the inner product is $\langle\!\langle \boldsymbol{x}_1|\boldsymbol{x}_2\rangle\!\rangle_{\mathscr{X}_M} := \mathbb{E}\left(\langle \boldsymbol{x}_1|\boldsymbol{x}_2\rangle_M\right)$.

The space of possible measurements $\mathcal{Y}$ can usually be taken to be finite dimensional, otherwise we take similarly as before a R-dimensional subspace $\mathcal{Y}_R$, whose elements are similarly represented by a vector of coefficients $\boldsymbol{y} \in \mathbb{R}^R$. For the discretised version of the RV $y(x_f) = y(\hat{h}(x_f, \varepsilon v))$ we will often use the shorthand $\boldsymbol{y}_f := \boldsymbol{y}(\boldsymbol{x}_f) = \boldsymbol{y}(\hat{h}(\boldsymbol{x}_f, \varepsilon \boldsymbol{v}))$.

As some of the most efficient ways of doing the update are linear filters based on the general idea of orthogonal decomposition—Eq. (42) in Sect. 3.2—applied to the mean—Eq. (45) in Sect. 3.2.1—but in the modified form Eq. (44) where g is a linear map, and especially the optimal linear map of the Gauss-Markov-Kalman (GMK) filter Eq. (83), we start from Proposition 18 in Sect. 3.3.2. For other approximations the finite dimensional discretisation would be largely analogous.

On $\mathbb{R}^M$, representing $\mathcal{X}_M$, the Kalman gain operator in Proposition 18 in Eq. (83) becomes a matrix $\boldsymbol{K} \in \mathbb{R}^{M\times R}$. Then the update corresponding to Eq. (83) is

$$\boldsymbol{x}_a = \boldsymbol{x}_f + \boldsymbol{K}(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{x}_f)), \text{ with } \boldsymbol{K} = \boldsymbol{C}_{xy}\,\boldsymbol{C}_{yy}^{-1}. \tag{91}$$

Here the covariances are $\boldsymbol{C}_{xy} := \mathbb{E}\left(\tilde{\boldsymbol{x}}_f \; \tilde{\boldsymbol{y}}(\boldsymbol{x}_f)\right)$, and similarly for $\boldsymbol{C}_{yy}$. Often the measurement error v in the measurement model $\tilde{h}(x_f, \varepsilon v) = h(x_f) + \varepsilon S_y(x_f)v$ is independent of $\boldsymbol{x}$—actually *uncorrelated* would be sufficient, i.e. $\boldsymbol{C}_{xv} = \boldsymbol{0}$—hence, assuming that S_y does not depend on x, $\boldsymbol{C}_{xx} = \boldsymbol{C}_{hh} + \varepsilon^2 \boldsymbol{S}_y \boldsymbol{C}_{vv} \boldsymbol{S}_y^T$ and $\boldsymbol{C}_{xy} = \boldsymbol{C}_{xh}$, where $h = h(x_f)$.

It is important to emphasise that the theory presented in the foregoing Sects. 2 and 3 is independent of any discretisation of the underlying spaces. But one usually can still not numerically compute with objects like $\boldsymbol{x} \in \mathscr{X}_M = \mathbb{R}^M \otimes \mathcal{S}$, as $\mathcal{S} = L_2(\Omega)$ is normally an infinite dimensional space, and has to be discretised. One well-known possibility are samples, i.e. the RV $\boldsymbol{x}(\omega)$ is represented by its value at certain points ω_z, and the points usually come from some quadrature rule. The well-known Monte Carlo (MC) method uses random samples, the quasi-Monte Carlo (QMC) method uses low discrepancy samples, and other rules like sparse grids (Smolyak rule) are possible. Using MC samples in the context of the linear update Eq. (83) is known as the *Ensemble Kalman Filter* (EnKF), see [34] for a general overview in this context, and [6, 7] for a thorough description and analysis. This method is conceptually fairly simple and is currently a favourite for problems where the computation of the predicted measurement $\boldsymbol{y}(\boldsymbol{x}_f(\omega_z))$ is difficult or expensive. It needs far fewer samples for meaningful results than MCMC, but on the other hand it uses the linear approximation inherent in Eq. (91).

Here we want to use so-called *functional* or *spectral* approximations, so similarly as for $\mathcal{X}_M$, we pick a finite set of linearly independent vectors in $\mathcal{S}$. As $\mathcal{S} = L_2(\Omega)$, these abstract vectors are in fact RVs with finite variance. Here we will use the best known example, namely *Wiener*'s *polynomial chaos* expansion (PCE) as basis [10, 14, 15, 22, 24, 43], this allows us to use Eq. (91) without sampling, see [26, 28, 29, 34, 35], and also [1, 37].

The PCE is an expansion in multivariate *Hermite polynomials* [10, 14, 15, 22, 24]; we denote by $H_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \prod_{k\in\mathbb{N}} h_{\alpha_k}(\theta_k) \in \mathcal{S}$ the multivariate polynomial in standard and independent Gaussian RVs $\boldsymbol{\theta}(\omega) = (\theta_1(\omega), \dots, \theta_k(\omega), \dots)_{k\in\mathbb{N}}$, where h_j is the usual uni-variate Hermite polynomial, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k, \dots)_{k\in\mathbb{N}} \in \mathcal{N} := \mathbb{N}_0^{(\mathbb{N})}$ is a multi-index of generally infinite length but with only finitely many entries non-zero. As $h_0 \equiv 1$, the infinite product is effectively finite and always well-defined.

The *Cameron-Martin* theorem assures us [14, 15, 22] that the set of these polynomials is dense in $\mathcal{S} = L_2(\Omega)$, and in fact $\{H_{\boldsymbol{\alpha}}/\sqrt{(\boldsymbol{\alpha}!)}\}_{\boldsymbol{\alpha}\in\mathcal{N}}$ is a complete orthonormal system (CONS), where $\boldsymbol{\alpha}! := \prod_{k\in\mathbb{N}}(\alpha_k!)$ is the product of the individual factorials, also well-defined as except for finitely many k one has $\alpha_k! = 0! = 1$. So one may write $\boldsymbol{x}(\omega) = \sum_{\boldsymbol{\alpha}\in\mathcal{N}} \boldsymbol{x}^{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}}(\boldsymbol{\theta}(\omega))$ with $\boldsymbol{x}^{\boldsymbol{\alpha}} \in \mathbb{R}^M$, and similarly for $\boldsymbol{y}$ and all other RVs. In this way the RVs are expressed as functions of other, known RVs $\boldsymbol{\theta}$—hence the name *functional* approximation—and not through samples.

The space $\mathcal{S}$ may now be discretised by taking a finite subset $\mathcal{J} \subset \mathcal{N}$ of size $J = |\mathcal{J}|$, and setting $\mathcal{S}_J = \operatorname{span}\{H_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{J}\} \subset \mathcal{S}$. The orthogonal projection P_J onto $\mathcal{S}_J$ is then simply

$$P_J : \mathcal{X}_M \otimes \mathcal{S} \ni \sum_{\boldsymbol{\alpha}\in\mathcal{N}} \boldsymbol{x}^{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}} \mapsto \sum_{\boldsymbol{\alpha}\in\mathcal{J}} \boldsymbol{x}^{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}} \in \mathcal{X}_M \otimes \mathcal{S}_J. \tag{92}$$

Taking Eq. (91), one may rewrite it as

$$\begin{aligned} \boldsymbol{x}_a &= \boldsymbol{x}_f + \boldsymbol{K}(\hat{\boldsymbol{y}} - \boldsymbol{y}_f) & (93)\\ &= \sum_{\boldsymbol{\alpha}\in\mathcal{N}} \boldsymbol{x}_a^{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha}\in\mathcal{N}} \left(\boldsymbol{x}_f^{\boldsymbol{\alpha}} + \boldsymbol{K}\left(\hat{\boldsymbol{y}}^{\boldsymbol{\alpha}} - \boldsymbol{y}_f^{\boldsymbol{\alpha}}\right)\right) H_{\boldsymbol{\alpha}}(\boldsymbol{\theta}). & (94)\end{aligned}$$

Observe, that as the measurement or observation $\hat{\boldsymbol{y}}$ is a constant, one has in Eq. (94) that only $\hat{\boldsymbol{y}}^0 = \hat{\boldsymbol{y}}$, all other coefficients $\hat{\boldsymbol{y}}^{\boldsymbol{\alpha}} = \boldsymbol{0}$ for $\boldsymbol{\alpha} \neq \boldsymbol{0}$.

Projecting both sides of Eq. (94) onto $\mathcal{X}_M \otimes \mathcal{S}_J$ is very simple and results in

$$\sum_{\boldsymbol{\alpha}\in\mathcal{J}} \boldsymbol{q}_a^{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathcal{J}} \left(\boldsymbol{q}_f^{\boldsymbol{\alpha}} + \boldsymbol{K}\left(z^{\boldsymbol{\alpha}} - \boldsymbol{y}_f^{\boldsymbol{\alpha}}\right)\right) H_{\boldsymbol{\alpha}}. \tag{95}$$

Obviously the projection P_J commutes with the Kalman operator K and hence with its finite dimensional analogue $\boldsymbol{K}$. One may actually concisely write Eq. (95) as

$$P_J \boldsymbol{x}_a = P_J \boldsymbol{x}_f + P_J \boldsymbol{K}(\hat{\boldsymbol{y}} - \boldsymbol{y}_f) = P_J \boldsymbol{x}_f + \boldsymbol{K}(P_J \hat{\boldsymbol{y}} - P_J \boldsymbol{y}_f). \tag{96}$$

Elements of the discretised space $\mathscr{X}_{M,J} = \mathcal{X}_M \otimes \mathcal{S}_J \subset \mathscr{X}$ thus may be written fully expanded as $\sum_{m=1}^{M} \sum_{\boldsymbol{\alpha}\in\mathcal{J}} x^{\boldsymbol{\alpha},m} \varrho_m H_{\boldsymbol{\alpha}}$. The tensor representation is $\boldsymbol{x} :=$

$\sum_{\alpha \in \mathcal{J}} \boldsymbol{x}^{\alpha} \otimes \boldsymbol{e}^{\alpha}$, where the $\{\boldsymbol{e}^{\alpha}\}$ are the canonical basis in $\mathbb{R}^J$, and may be used to express Eq. (95) or Eq. (96) succinctly as

$$\mathsf{x}_a = \mathsf{x}_f + \mathsf{K}(\hat{\mathsf{y}} - \mathsf{y}_f), \tag{97}$$

again an equation between the tensor representations of some RVs, where $\mathsf{K} = \boldsymbol{K} \otimes \boldsymbol{I}$, with $\boldsymbol{K}$ from Eq. (91). Hence the update equation is naturally in a tensorised form. This is how the update can finally be computed in the PCE representation without any sampling [26, 28, 34, 35]. Analogous statements hold for the forms of the update Eq. (68) with higher order terms $n > 1$, and do not have to be repeated here. Let us remark that these updates go very seamlessly with very efficient methods for sparse or low-rank approximation of tensors, c.f. the monograph [12] and the literature therein. These methods are PCE-forms of the Bayesian update, and in particular the Eq. (97), because of its formal affinity to the Kalman filter (KF), may be called the polynomial chaos expansion based Kalman filter (PCEKF).

It remains to say how to compute the terms ${}^{k}H$ in the update equation Eq. (68)—or rather the terms in the defining Eq. (70) in Theorem 17—in this approach. Given the PCEs of the RVs, this is actually quite simple as any moment can be computed directly from the PCE [24, 28, 35]. A typical term $\langle y^{\vee k}\rangle = \langle \mathrm{Sym}(y^{\otimes k})\rangle = \mathrm{Sym}(\langle y^{\otimes k}\rangle)$ in the operator matrix Eq. (70), where $\boldsymbol{y} = \sum_{\alpha} \boldsymbol{y}^{\alpha} H_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$, may be computed through

$$\begin{aligned}\langle \boldsymbol{y}^{\otimes k}\rangle &= \mathbb{E}\left(\bigotimes_{i=1}^{k} \sum_{\boldsymbol{\alpha}_i} \left(\boldsymbol{y}^{\boldsymbol{\alpha}_i} H_{\boldsymbol{\alpha}_i}\right)\right) \\ &= \mathbb{E}\left(\sum_{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k} \bigotimes_{i=1}^{k} \boldsymbol{y}^{\boldsymbol{\alpha}_i} \prod_{i=1}^{k} H_{\boldsymbol{\alpha}_i}\right) = \sum_{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k} \bigotimes_{i=1}^{k} \boldsymbol{y}^{\boldsymbol{\alpha}_i}\, \mathbb{E}\left(\prod_{i=1}^{k} H_{\boldsymbol{\alpha}_i}\right)\end{aligned} \tag{98}$$

As here the $H_{\boldsymbol{\alpha}}$ are *polynomials*, the last expectation in Eq. (98) is finally over products of powers of pairwise independent normalised Gaussian variables, which actually may be done analytically [14, 15, 22]. But some simplifications come from remembering that $\boldsymbol{y}^0 = \mathbb{E}(\boldsymbol{y}) = \bar{\boldsymbol{y}}$, $H_{\boldsymbol{0}} \equiv 1$, the orthogonality relation $\langle H_{\boldsymbol{\alpha}} | H_{\boldsymbol{\beta}}\rangle = \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}\, \boldsymbol{\alpha}!$, and that the Hermite polynomials are an *algebra*. Hence $H_{\boldsymbol{\alpha}} H_{\boldsymbol{\beta}} = \sum_{\boldsymbol{\gamma}} c^{\boldsymbol{\gamma}}_{\boldsymbol{\alpha},\boldsymbol{\beta}} H_{\boldsymbol{\gamma}}$, where the *structure* coefficients $c^{\boldsymbol{\gamma}}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ are known analytically [22, 24, 28, 35].

Similarly, for a RV $R = R(x)$, for a typical right-hand-side term $\langle R(x) \otimes y^{\vee k}\rangle = \langle R \otimes \mathrm{Sym}(y^{\otimes k})\rangle$ in Eq. (70) with $\boldsymbol{R} = \sum_{\beta} \boldsymbol{R}^{\boldsymbol{\beta}} H_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ one has

$$\langle R \otimes \mathrm{Sym}(y^{\otimes k})\rangle = \sum_{\boldsymbol{\beta},\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k} \boldsymbol{R} \otimes \mathrm{Sym}\left(\bigotimes_{i=1}^{k} \boldsymbol{y}^{\boldsymbol{\alpha}_i}\right) \mathbb{E}\left(H_{\boldsymbol{\beta}} \prod_{i=1}^{k} H_{\boldsymbol{\alpha}_i}\right). \tag{99}$$

As these relations may seem a bit involved—they are actually just an intricate combination of *known* terms—we show here how simple they become for the case of the covariance needed in the linear update formula Eq. (83) or rather Eq. (91):

$$\boldsymbol{C}_{yy} = \sum_{\alpha\in\mathcal{N},\alpha\neq 0} (\boldsymbol{\alpha}!)\ \boldsymbol{y}^{\alpha}\otimes\boldsymbol{y}^{\alpha} \approx \sum_{\alpha\in\mathcal{J},\alpha\neq 0} (\boldsymbol{\alpha}!)\ \boldsymbol{y}^{\alpha}\otimes\boldsymbol{y}^{\alpha}, \tag{100}$$

$$\boldsymbol{C}_{xy} = \sum_{\alpha\in\mathcal{N},\alpha\neq 0} (\boldsymbol{\alpha}!)\ \boldsymbol{x}^{\alpha}\otimes\boldsymbol{y}^{\alpha} \approx \sum_{\alpha\in\mathcal{J},\alpha\neq 0} (\boldsymbol{\alpha}!)\ \boldsymbol{x}^{\alpha}\otimes\boldsymbol{y}^{\alpha}. \tag{101}$$

Looking for example at Eq. (91) and our setup as explained in Sect. 1, we see that the coefficients of $\boldsymbol{y}(\boldsymbol{x}_f) = \sum_\alpha \boldsymbol{y}_f^\alpha H_\alpha$ have to be computed from those of $\boldsymbol{x}_f = \sum_\beta \boldsymbol{x}_f^\beta H_\beta$. This propagation of uncertainty through the system is known as *uncertainty quantification* (UQ), e.g. [24] and the references therein. For the sake of brevity, we will not touch further on this subject, which nevertheless is the bedrock on which the whole computational procedure is built.

We next concentrate in Sect. 5 on examples of updating with ψ_m for the case $m = 1$ in Eq. (68), whereas in Sect. 6 an example for the case $m = 2$ in Eq. (68) will be shown.

5 The Linear Bayesian Update

All the examples in this Sect. 5 have been computed with the case $m = 1$ of up to linear terms in Eq. (68), i.e. this is the LBU with PCEKF. As the traditional Kalman filter is highly geared towards Gaussian distributions [17], and also its Monte Carlo variant EnKF which was mentioned in Sect. 4 tilts towards Gaussianity, we start with a case—already described in [28]—where the quantity to be identified has a strongly non-Gaussian distribution, shown in black—the 'truth'—in Fig. 1. The operator describing the system is the identity—we compute the quantity directly, but there is a Gaussian measurement error. The 'truth' was represented as a 12th degree PCE. We use the methods as described in Sect. 4, and here in particular the Eqs. (91) and (97), the PCEKF.

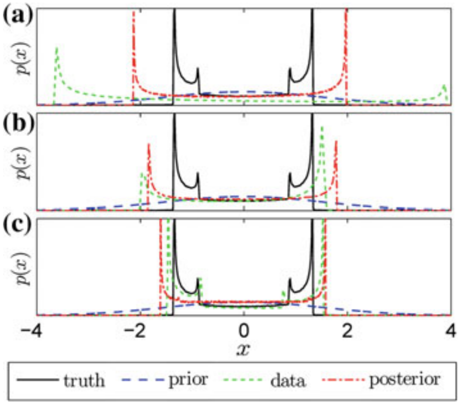


Fig. 1 Pdfs for linear Bayesian update (LBU), from [28]

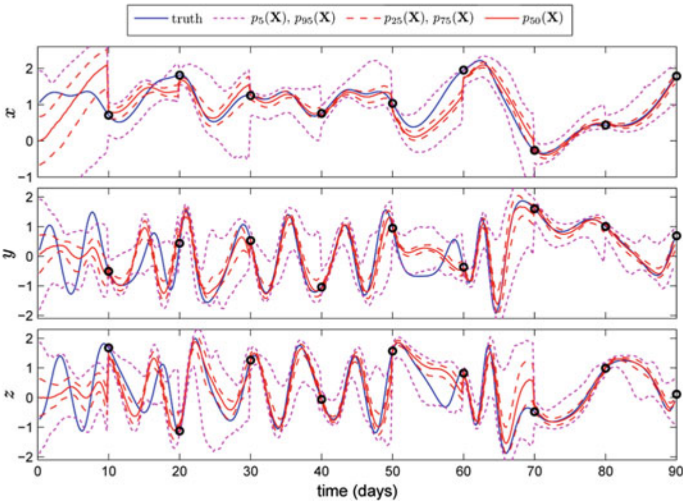


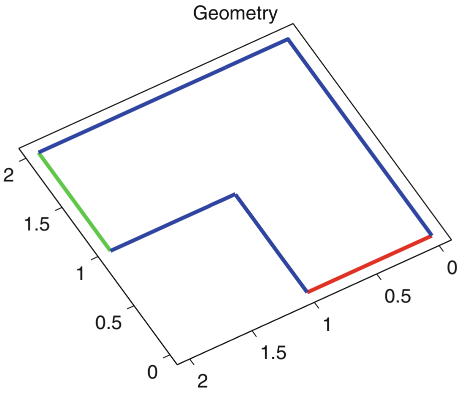
Fig. 2 Time evolution of Lorenz-84 state and uncertainty with the LBU, from [28]

The update is repeated several times (here ten times) with new measurements—see Fig. 1. The task is here to identify the distribution labelled as 'truth' with ten updates of N samples (where $N = 10, 100, 1000$ was used), and we start with a very broad Gaussian prior (in blue). Here we see the ability of the polynomial based LBU, the PCEKF, to identify highly non-Gaussian distributions, the posterior is shown in red and the pdf estimated from the samples in green; for further details see [28].

The next example is also from [28], where the system is the well-known Lorenz-84 chaotic model, a system of three nonlinear ordinary differential equations operating in the chaotic regime. This is truly an example along the description of Eqs. (5) and (9) in Sect. 2.1. Remember that this was originally a model to describe the evolution of some amplitudes of a spherical harmonic expansion of variables describing world climate. As the original scaling of the variables has been kept, the time axis in Fig. 2 is in *days*. Every ten days a noisy measurement is performed and the state description is updated. In between the state description evolves according to the chaotic dynamic of the system. One may observe from Fig. 2 how the uncertainty—the width of the distribution as given by the quantile lines—shrinks every time a measurement is performed, and then increases again due to the chaotic and hence noisy dynamics. Of course, we did not really measure world climate, but rather simulated the 'truth' as well, i.e. a *virtual* experiment, like the others to follow. More details may be found in [28] and the references therein.

From [35] we take the example shown in Fig. 3, a linear stationary diffusion equation on an L-shaped plane domain as alluded to in Sect. 1. The diffusion coefficient κ in Eq. (2) is to be identified. As argued in [34], it is better to work with $q = \log \kappa$ as the diffusion coefficient has to be positive, but the results are shown in terms of κ.

Fig. 3 Diffusion domain, from [35]



One possible realisation of the diffusion coefficient is shown in Fig. 4. More realistically, one should assume that κ is a symmetric positive definite tensor field, unless one knows that the diffusion is *isotropic*. Also in this case one should do the updating on the logarithm. For the sake of simplicity we stay with the scalar case, as there is no principal novelty in the non-isotropic case. The virtual experiments use different right-hand-sides f in Eq. (2), and the measurement is the observation of the solution u averaged over little patches.

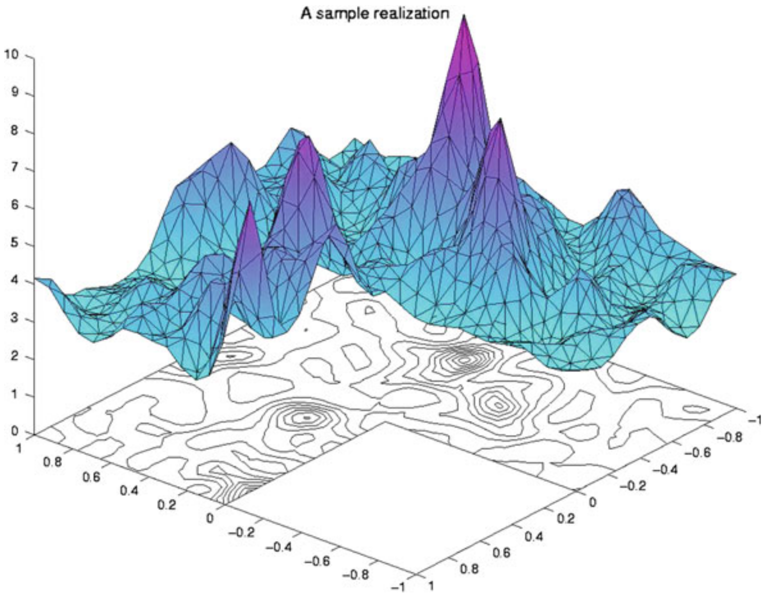


Fig. 4 Conductivity field, from [35]

Fig. 5 Convergence of identification, from [35]

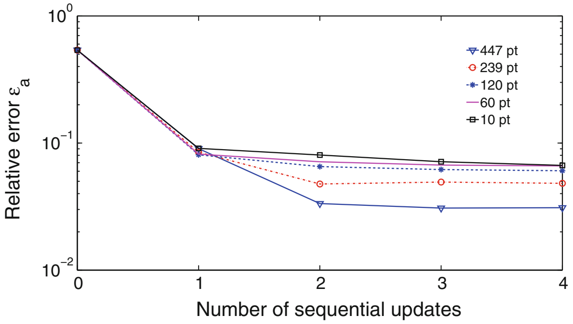
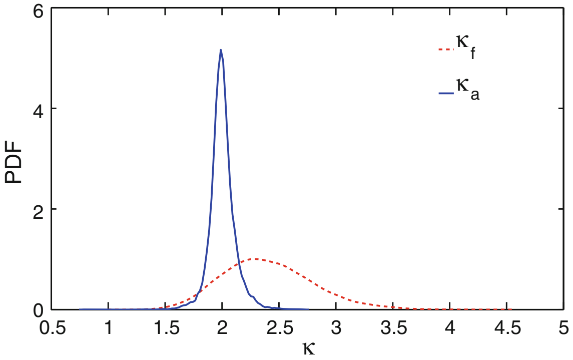


Fig. 6 Prior and posterior, from [35]



In Fig. 5 one may observe the decrease of the error with successive updates, but due to measurement error and insufficient information from just a few patches, the curves level off, leaving some residual uncertainty. The pdfs of the diffusion coefficient at some point in the domain before and after the updating is shown in Fig. 6, the 'true' value at that point was $\kappa = 2$. Further details can be found in [35].

6 The Nonlinear Bayesian Update

In this section we want to show a computation with the case $m = 2$ of up to quadratic terms in ψ_m in Eq. (68). We go back to the example of the chaotic Lorentz-84 [28] model already shown in Sect. 5, from Eqs. (5) and (9) in Sect. 2.1. This kind of experiment has several advantages but at the same time also challenges for identification procedures: it has only a three-dimensional state space, these are the uncertain 'parameters', i.e. $\boldsymbol{x} = (x_1, x_2, x_3) = (x, y, z) \in \mathcal{X} = \mathbb{R}^3$, the corresponding operator A resp. f in the abstract Eq. (1) resp. Eq. (5) is sufficiently nonlinear to make the problem difficult, and adding to this we operate the equation in its chaotic regime, so that new uncertainty from the numerical computation is added between measurements.

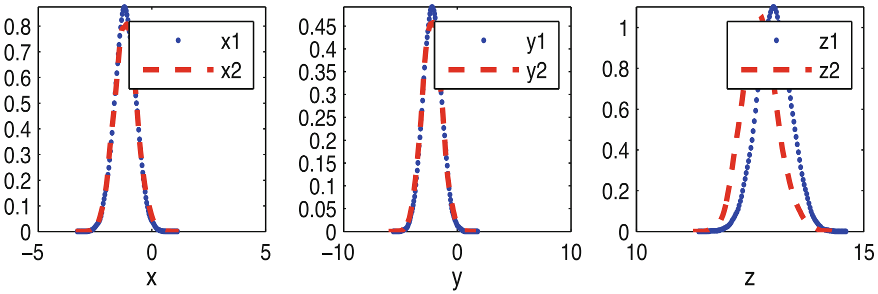


Fig. 7 Linear measurement: Comparison posterior for LBU ($m = 1$) and QBU ($m = 2$) after one update

As a first set of experiments we take the measurement operator to be linear in $\boldsymbol{x}$, i.e. we can observe the *whole* state directly. At the moment we consider updates after each day—whereas in Sect. 5 the updates were performed every 10 days. The update is done once with the linear Bayesian update (LBU), and again with a *quadratic* nonlinear BU (QBU) with $m = 2$. The results for the posterior pdfs are given in Fig. 7, where the linear update is dotted in blue, and the full red line is the quadratic QBU; there is hardly any difference between the two, most probably indicating that the LBU is already very accurate.

As the differences between LBU and QBU were small—we take this as an indication that the LBU is not too inaccurate an approximation to the conditional expectation—we change the experiment and take a nonlinear measurement function, which is now cubic: $h(\boldsymbol{x}) = (x^3, y^3, z^3)$. We now observe larger differences between LBU and QBU.

These differences in posterior pdfs after one update may be gleaned from Fig. 8, and they are indeed larger than in the linear case Fig. 7, due to the strongly nonlinear measurement operator, showing that the QBU may provide much more accurate tracking of the state, especially for non-linear observation operators.

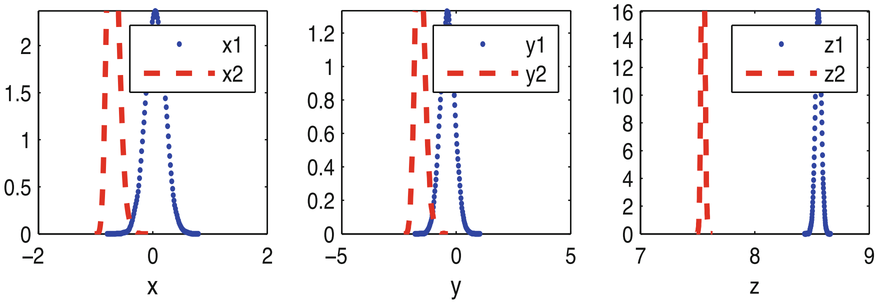


Fig. 8 Cubic measurement: Comparison posterior for LBU ($m = 1$) and QBU ($m = 2$) after one update

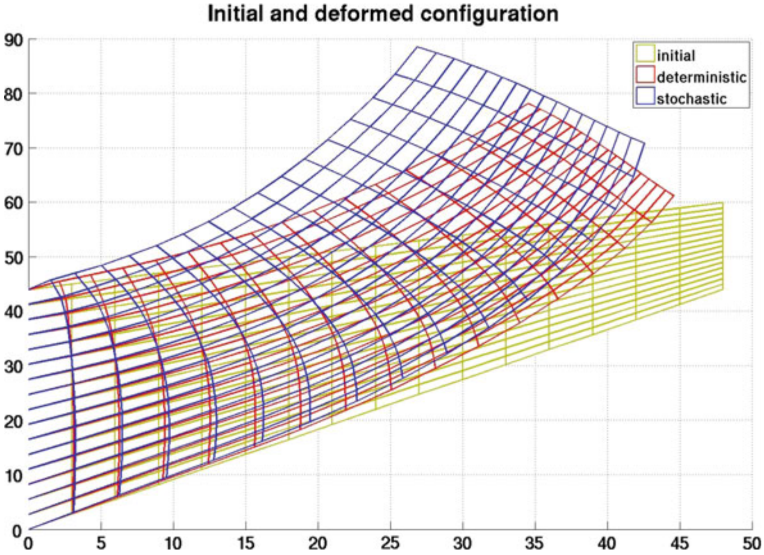
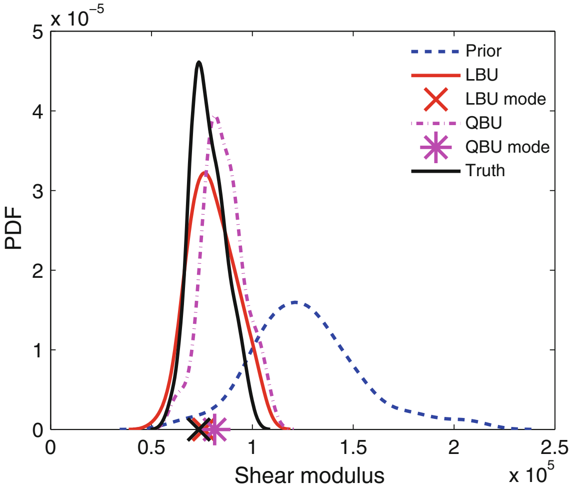


Fig. 9 Deformations, from [34, 36]

As a last example we take a strongly nonlinear and also non-smooth situation, namely elasto-plasticity with linear hardening and large deformations and a *Kirchhoff-St. Venant* elastic material law [34, 36]. This example is known as *Cook's membrane*, and is shown in Fig. 9 with the undeformed mesh (initial), the deformed one obtained by computing with average values of the elasticity and plasticity material constants (deterministic), and finally the average result from a stochastic forward calculation of the probabilistic model (stochastic), which is described by a variational inequality [36].

The shear modulus G, a random field and not a deterministic value in this case, has to be identified, which is made more difficult by the non-smooth non-linearity. In Fig. 10 one may see the 'true' distribution at one point in the domain in an unbroken black line, with the mode—the maximum of the pdf—marked by a black cross on the abscissa, whereas the prior is shown in a dotted blue line. The pdf of the LBU is shown in an unbroken red line, with its mode marked by a red cross, and the pdf of the QBU is shown in a broken purple line with its mode marked by an asterisk. Again we see a difference between the LBU and the QBU. But here a curious thing happens; the mode of the LBU-posterior is actually closer to the mode of the 'truth' than the mode of the QBU-posterior. This means that somehow the QBU takes the prior more into account than the LBU, which is a kind of overshooting which has been observed at other occasions. On the other hand the pdf of the QBU is narrower—has less uncertainty—than the pdf of the LBU.

Fig. 10 LBU and QBU for the shear modulus



7 Conclusion

The connection between inverse problems and uncertainty quantification was shown. An abstract model of a system was introduced, together with a measurement operator, which provides a possibility to predict—in a probabilistic sense—a measurement. The framework chosen is that of Bayesian analysis, where uncertain quantities are modelled as random variables. New information leads to an update of the probabilistic description via Bayes's rule.

After elaborating on the—often not well-known—connection between conditional probabilities as in Bayes's rule and conditional expectation, we set out to compute and—necessarily—approximate the conditional expectation. As a polynomial approximation was chosen, there is the choice up to which degree one should go. The case with up to linear terms—the linear Bayesian update (LBU)—is best known and intimately connected with the well-known Kalman filter. We call this update the Gauss-Markov-Kalman filter. In addition, we show how to compute approximations of higher order, in particular the quadratic Bayesian update (QBU).

There are several possibilities on how one may choose a numerical realisation of these theoretical concepts, and we decided on functional or spectral approximations. It turns out that this approach goes very well with recent very efficient approximation methods building on separated or so-called low-rank tensor approximations.

Starting with the linear Bayesian update, a series of examples of increasing complexity is shown. The method works well in all cases. Some examples are then chosen to show the nonlinear or rather quadratic Bayesian update, where we go up to quadratic terms. A series of experiments is chosen with different measurement operators, which have quite a marked influence on whether the linear and quadratic update are close to each other.

Acknowledgments This work was partially supported by the German research foundation "Deutsche Forschungsgemeinschaft" (DFG).

References

1. Blanchard, E.D., Sandu, A., Sandu, C.: A polynomial chaos-based Kalman filter approach for parameter estimation of mechanical systems. Journal of Dynamic Systems, Measurement, and Control **132**(6), 061404 (2010). doi:10.1115/1.4002481
2. Bobrowski, A.: Functional Analysis for Probability and Stochastic Processes. Cambridge University Press, Cambridge (2005)
3. Bosq, D.: Linear Processes in Function Spaces. Theory and Applications., *Lecture Notes in Statistics*, vol. 149. Springer, Berlin (2000). Contains definition of strong or L-orthogonality for vector valued random variables
4. Bosq, D.: General linear processes in Hilbert spaces and prediction. Journal of Statistical Planning and Inference **137**, 879–894 (2007). doi:10.1016/j.jspi.2006.06.014
5. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Kluwer, Dordrecht (2000)
6. Evensen, G.: Data Assimilation — The Ensemble Kalman Filter. Springer, Berlin (2009)
7. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. IEEE Control Systems Magazine **29**, 82–104 (2009). doi:10.1109/MCS.2009.932223
8. Galvis, J., Sarkis, M.: Regularity results for the ordinary product stochastic pressure equation. SIAM Journal on Mathematical Analysis **44**, 2637–2665 (2012). doi:10.1137/110826904
9. Gamerman, D., Lopes, H.F.: Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall, Boca Raton, FL (2006)
10. Ghanem, R., Spanos, P.D.: Stochastic finite elements—A spectral approach. Springer, Berlin (1991)
11. Goldstein, M., Wooff, D.: Bayes Linear Statistics—Theory and Methods. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester (2007)
12. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. Springer, Berlin (2012)
13. Hida, T., Kuo, H.H., Potthoff, J., Streit, L.: White Noise—An Infinite Dimensional Calculus. Kluwer, Dordrecht (1999)
14. Holden, H., Øksendal, B., Ubøe, J., Zhang, T.S.: Stochastic Partial Differential Equations. Birkhäuser, Basel (1996)
15. Janson, S.: Gaussian Hilbert spaces. Cambridge Tracts in Mathematics, 129. Cambridge University Press, Cambridge (1997)
16. Jaynes, E.T.: Probability Theory, The Logic of Science. Cambridge University Press, Cambridge (2003)
17. Kálmán, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME—J. of Basic Engineering (Series D) **82**, 35–45 (1960)
18. Kučerová, A., Matthies, H.G.: Uncertainty updating in the description of heterogeneous materials. Technische Mechanik **30**(1–3), 211–226 (2010)
19. Law, K.H.J., Litvinenko, A., Matthies, H.G.: Nonlinear evolution, observation, and update (2015)
20. Luenberger, D.G.: Optimization by Vector Space Methods. John Wiley & Sons, Chichester (1969)
21. Madras, N.: Lectures on Monte Carlo Methods. American Mathematical Society, Providence, RI (2002)
22. Malliavin, P.: Stochastic Analysis. Springer, Berlin (1997)
23. Marzouk, Y.M., Najm, H.N., Rahn, L.A.: Stochastic spectral methods for efficient Bayesian solution of inverse problems. Journal of Computational Physics **224**(2), 560–586 (2007). doi:10.1016/j.jcp.2006.10.010

24. Matthies, H.G.: Uncertainty quantification with stochastic finite elements. In: E. Stein, R. de Borst, T.J.R. Hughes (eds.) Encyclopaedia of Computational Mechanics. John Wiley & Sons, Chichester (2007). doi:10.1002/0470091355.ecm071
25. Matthies, H.G., Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. Computer Methods in Applied Mechanics and Engineering **194**(12–16), 1295–1331 (2005)
26. Matthies, H.G., Litvinenko, A., Pajonk, O., Rosić, B.V., Zander, E.: Parametric and uncertainty computations with tensor product representations. In: A. Dienstfrey, R. Boisvert (eds.) Uncertainty Quantification in Scientific Computing, *IFIP Advances in Information and Communication Technology*, vol. 377, pp. 139–150. Springer, Berlin (2012). doi:10.1007/978-3-642-32677-6
27. Moselhy, T.A., Marzouk, Y.M.: Bayesian inference with optimal maps. Journal of Computational Physics **231**, 7815–7850 (2012). doi:10.1016/j.jcp.2012.07.022
28. Pajonk, O., Rosić, B.V., Litvinenko, A., Matthies, H.G.: A deterministic filter for non-Gaussian Bayesian estimation — applications to dynamical system estimation with noisy measurements. Physica D **241**, 775–788 (2012). doi:10.1016/j.physd.2012.01.001
29. Pajonk, O., Rosić, B.V., Matthies, H.G.: Sampling-free linear Bayesian updating of model state and parameters using a square root approach. Computers and Geosciences **55**, 70–83 (2013). doi:10.1016/j.cageo.2012.05.017
30. Papoulis, A.: Probability, Random Variables, and Stochastic Processes, third edn. McGraw-Hill Series in Electrical Engineering. McGraw-Hill, New York (1991)
31. Parno, M., Moselhy, T., Marzouk, Y.: A multiscale strategy for Bayesian inference using transport maps. arXiv:1507.07024v1 [stat:CO] (2015)
32. Rao, M.M.: Conditional Measures and Applications. CRC Press, Boca Raton, FL (2005)
33. Roman, L., Sarkis, M.: Stochastic Galerkin method for elliptic SPDEs: A white noise approach. Discrete Cont. Dyn. Syst. Ser. B **6**, 941–955 (2006)
34. Rosić, B.V., Kučerová, A., Sýkora, J., Pajonk, O., Litvinenko, A., Matthies, H.G.: Parameter identification in a probabilistic setting. Engineering Structures **50**, 179–196 (2013). doi:10.1016/j.engstruct.2012.12.029
35. Rosić, B.V., Litvinenko, A., Pajonk, O., Matthies, H.G.: Sampling-free linear Bayesian update of polynomial chaos representations. Journal of Computational Physics **231**, 5761–5787 (2012). doi:10.1016/j.jcp.2012.04.044
36. Rosić, B.V., Matthies, H.G.: Identification of properties of stochastic elastoplastic systems. In: M. Papadrakakis, G. Stefanou, V. Papadopoulos (eds.) Computational Methods in Stochastic Dynamics, *Computational Methods in Applied Sciences*, vol. 26, pp. 237–253. Springer, Berlin (2013). doi:10.1007/978-94-007-5134-7
37. Saad, G., Ghanem, R.: Characterization of reservoir simulation models using a polynomial chaos-based ensemble Kalman filter. Water Resources Research **45**, W04,417 (2009). doi:10.1029/2008WR007148
38. Sanz-Alonso, D., Stuart, A.M.: Long-time asymptotics of the filtering distribution for partially observed chaotic dynamical systems. arXiv:1411.6510v1 [math.DS] (2014)
39. Segal, I.E., Kunze, R.A.: Integrals and Operators. Springer, Berlin (1978)
40. Stuart, A.M.: Inverse problems: A Bayesian perspective. Acta Numerica **19**, 451–559 (2010). doi:10.1017/S0962492910000061
41. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia, PA (2004)
42. Tarn, T.J., Rasis, Y.: Observers for nonlinear stochastic systems. IEEE Transactions on Automatic Control **21**, 441–448 (1976)
43. Wiener, N.: The homogeneous chaos. American Journal of Mathematics **60**(4), 897–936 (1938)
44. Xiu, D., Karniadakis, G.E.: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM Journal of Scientific Computing **24**, 619–644 (2002)

Heterogeneous Materials Models, Coupled Mechanics-Probability Problems and Energetically Optimal Model Reduction

Rainer Niekamp, Martin Krosche and Adnan Ibrahimbegović

Abstract The main scientific goal of this chapter is to provide the sound theoretical basis and an efficient multi-scale computational algorithm for the description of irreversible behaviour of heterogeneous materials, through coupling of nonlinear mechanics and probability. Although we focus upon concrete as the most frequently used construction material, the proposed methodology applies to a number of other composite materials of interest for practical applications. The first main challenge pertains to providing the sound formulation for coupled nonlinear mechanics-probability computations for the prediction of failure and durability of a massive concrete structure. We show that the multi-scale interpretation of damage mechanisms can provide the most meaningful probability density description by identifying a limited number of the sources of uncertainty for material parameters governing the failure phenomena, which can be described in terms of random fields with a probability description computed from finer scales. The second main challenge is in providing an efficient solution procedure to this coupled mechanics-probability problem, which is done by using the Spectral Stochastic Finite Element Method (SSFEM). In particular, we seek to circumvent the main disadvantage of the classical SSFEM regarding the Curse of Dimension, where the coupled mechanics-probability problem dimension grows with a number of random fields. The proposed techniques include Low-Rank approaches and solution space reductions, with a rank-one update scheme based upon the variational formulation of the problem. However, since the resulting Low-Rank representation is not necessarily an optimal one with respect to

R. Niekamp (✉) · M. Krosche
TU-Braunschweig, Institute for Scientific Computing, Braunschweig, Germany
e-mail: r.niekamp@tu-bs.de

M. Krosche
e-mail: martin.krosche@tu-bs.de

A. Ibrahimbegović
Sorbonne Universités/Université Techanologie Compiègne,
Chair for Computational Mechanics, Paris, France
e-mail: adnan.ibrahimbegovic@utc.fr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_11

the minimal energy at the given rank, we further extend this scheme to provide an optional solution space adaptation and the possibility to compute the energetically optimal reduced basis.

1 Introduction

One of the main issues for concrete structures is their integrity and durability under long term operational loading and degradation, with crack growth under mechanical loading or temperature change. Our objective is to develop methods capable of a detailed representation of crack propagation leading to localised failure phenomena, which can significantly reduce the structure durability. The second main issue stems from the fact that the localized failure is very sensitive to material heterogeneities, and that one needs to quantify the corresponding uncertainty propagation. Special attention is given to massive structures with irreplaceable components, which require the most detailed description of localised failure with non-negligible contributions from consideration of both fracture process zone mechanisms and macro cracks. This calls for the development of a heterogeneous multi-scale method capable of dealing with evolution and interaction of failure mechanisms defined at different scales through a consistent mechanical and stochastic coupling. In particular, we need: (i) the meso-scale distinguishing between aggregate and cement paste; (ii) the macro-scale of concrete with refined damage and/or plasticity criteria capable of better using information from finer scales than standard phenomenological models and (iii) the efficient computational procedure for uncertainty propagation by solving the corresponding coupled mechanics-probability problem.

The computational modelling of localised mechanical failure mechanisms of heterogeneous materials, focusing on the technologically important example of concrete, has received much attention in recent literature. While at the structural scale one may view concrete as a homogeneous medium (e.g. [4, 17]), this is not possible at the scale of localised failure, where concrete is heterogeneous over many scales. To be able to computationally model the behaviour of such a material, a multi-scale (e.g. [9]) approach is in order. As the details on the small scales are uncertain, they will be modelled probabilistically, i.e. as a random medium (see e.g. [11]). One cah choose to start with an existing probabilistic model on the micro-scale (1 μm–1 mm) to define the bonding of the cement paste with the aggregate on the meso-scale (e.g. [18]) (1 mm–10 cm) including the inherent uncertainty, by mapping the uncertainty of the micro-model through uncertainty quantification techniques to the material properties of the meso-scale. Added to this will be the geometric randomness of the placement, size, and shape of the aggregate (see e.g. [3]). Here is where the localised failure mechanisms will be modelled computationally. As macroscopic cracks develop from fracture process zones (FPZ), this has effects on the macro-scale behaviour (10 cm–1 m), where a non-local interaction occurs with the reinforcing steel. From here one may reach the structural scale of e.g. a whole bridge or dam (1–100 m). The scale-transition to span is fully comparable to nano-mechanics challenge, but with a shifts.

Thus, at each scale transition there is a simplification (model reduction) of both the mechanical behaviour and the stochastic description (reduction in the number of random variables used) in order to make the whole model computationally feasible. While the mechanical conditions at a scale transition are to match displacements and forces (stresses) as well as energy and dissipation, the probabilistic scale transition can use Bayesian identification methods to achieve the best match (e.g. [16]). The uncertainty from a heterogeneous material may be viewed as an epistemic uncertainty, and some researchers favour deterministic methods for their description. Here we stay with the probabilistic modelling, as this allows the incorporation of new knowledge via Bayesian methods resulting with updating of probability distribution at coarse scales.

The multi-scale coupling of highly nonlinear failure mechanisms and probability has not yet been fully accomplished successfully within a predictive model, and is thus the main goal of this work. We propose an original multi-scale probability approach for dealing with failure mechanics, with the fine and coarse scale properties defined in terms of random fields whose probability distribution is obtained as the result of uncertainty propagation from the finer scales (see [6]). The final outcome is a stochastic model for localised failure behaviour which is also able to reflect and quantify the inherent uncertainties. The model of this kind can better perform in structural scale testing and validation against existing test results. Moreover, the data from both material and structural scale experiments can be used in a Bayesian identification process at various scales.

The main challenge in probabilistic multi-scale analysis is how to account for the uncertainty propagation through two or more different spatial scales. There is a large and growing body of literature on how to provide uncertainty propagation at one scale by means of Monte Carlo or stochastic spectral approximations (collocation, projection, Galerkin), the vast majority of publications is restricted to linear or mildly nonlinear problems, or limited to random parameters rather than random fields when quantifying uncertainties in the inelastic behaviour of materials.

Highly nonlinear mechanical processes such as plasticity and softening have just started being addressed with properties as random fields, and there are advances in adding stochastics in a multi-scale framework (e.g. [6]). The computational cost of multi-scale probabilistic nonlinear analysis is enormous, and ways for reduced order models (ROM) have been always important. One possibility are POD or PGD approaches or truncated iterations, which are all a form of low-rank tensor approximations. A more physically inspired way is in [6], which retains the interpretation of a reduced model enforcing its compatibility with standard computational formats of inelastic models (e.g. [8]). The main challenge that remains is how to combine probabilistic approaches with a multi-scale method with more general evolution equations for internal variables and still be able to retain the computational efficiency (see e.g. [6]). Especially as regards to cracking, the identification of suitable probability distributions is necessary for the description of large deviations which are far from equilibrium.

2 Theoretical Formulation Heterogeneous Multiscale Method

2.1 State-of-the-art Developments

Even in a deterministic setting, propagating information from fine to coarse scales in nonlinear multi-scale models for failure mechanics of concrete remains a formidable challenge. Two families of methods have emerged. The first method is based on classical homogenisation procedures, but applied to nonlinear analysis where the homogenised stress or homogenised tangent modulus (or rather energy) are constructed from finer (micro) scales by imposing either the average stress or average strain from the coarser (macro) scale. The finite element implementation is reduced to computing the homogenised stress and modulus at the level of each Gauss integration point. The second method is used for the case where scale separation does not hold (due to an insufficiently large difference between macro and micro scales), which does not allow classical homogenisation. For such a case, the class ical homogenisation results would obtain the apparent properties only and not sharp bounds. The corresponding replacement of homogenization for such a case has been proposed in [9], and its finite element formulation and software implementation have been studied in [13], respectively. The main idea of this method is to use a mesh in an element, and to store the results of finer scale computations at the level of the particular finite element in terms of its internal force vector or its tangent stiffness matrix. The subsequent developments of this method (e.g. [7]) have shown the benefit of using discrete elements for representing the fine scales and corresponding inelastic failure mechanisms, allowing to describe the material heterogeneities (e.g. [3]). Several other recent works (e.g. [18]) have pointed out the influence of material heterogeneities on concrete failure modes, and thus the superiority of the multi-scale approach providing much more predictive interpretation from classical phenomenological models (e.g. [17]). The goal of this work is to carry on further with these developments based upon the synergy of two domains in order to provide a sound probability distribution for these heterogeneities and quantify the corresponding uncertainty propagation, thus seeking to increase the predictive capabilities of our models.

Modelling of localised failure mechanisms for concrete leading to softening response represent a very significant challenge for ensuring the convergence of finite element computations. The standard finite element models cannot deliver a mesh invariant response, and a number of remedies have proposed over last 20 years (e.g. see [8]) for a summary. The X-FEM method, which is dominant nowadays, is related to representing the crack-induced displacement discontinuity in the finite element mesh by using a Heaviside function. Our contribution to that domain [8] was to further generalise this model for the case where the fracture process zone is not negligible, where we combine the standard plasticity or damage (representing micro-cracks) with a displacement discontinuity (representing macro-cracks). The method is known as Embedded Discontinuity or ED-FEM, since the discontinuity does not have to be connected from element to element. We have shown more recently that the damage model of this kind can be applied to capture micro-cracks and macro-cracks in concrete, and that its parameters can be identified from heterogeneous stress fields such as in a 3-point bending test. More recently we have further combined the ED-FEM model for concrete with X-FEM model for bond-slip [5]. This is a multi-scale approach where the macro-scale of concrete is combined with non-local scale for bond-slip connecting all the slips along a particular reinforcement bar. In this manner we provide a sound interpretation of crack spacing and opening in reinforced concrete, with respect to previous unsuccessful attempts based upon phenomenological models of reinforced concrete. There still remains the challenge to further extend such a multi-scale approach, in connecting more than two-scales but rather spanning in the most meaningful manner all the scales from micro (cement RVE), over meso (concrete RVE) to macro (reinforced concrete X-RVE).

2.2 Meso-Scale Model of Material Heterogeneities With deterministic Material Parameters

At meso-scale, we consider concrete a heterogeneous material built of two different phases and we assume that each of these phases is described by the inclusions positions and shapes (see Fig. 1). These two phases introduce two types of discontinuities (see [7]), namely a discontinuity of the strain field and a discontinuity of the displacement field, both of them lying at the same position (prescribed by the known physical interface between the two phases). Meshing is one of the major issue in modelling heterogeneous materials. Namely, trying to provide exact representation of different phases and their complex shapes might frequently lead to a quite high number of degrees-of-freedom and also quite distorted meshes. Moreover, the meshing process itself might consist in a complex and time-consuming algorithm. We show here how

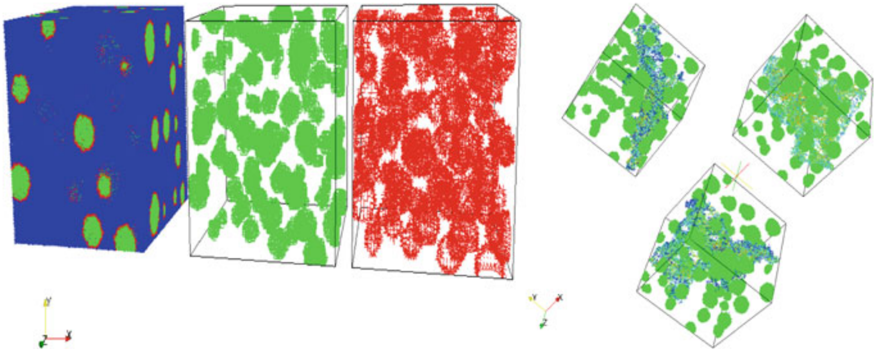


Fig. 1 (i) Three phase representation of concrete meso-structure: aggregate (*green*), cement matrix (*blue*) and interface (*red*); (ii) Representation of failure mechanisms in uni-axial tension tests indicating broken links

to employ so-called structured mesh, which employs regular element shaped in order to simplify the meshing process for two-phase heterogeneous materials. Hence, the structured meshes are not constraint by the physical interfaces between the different phases, which can ‘cross’ any of these regular elements. The key ingredients to provide such mesh are field discontinuities introduced inside the elements in which the physical interfaces are present. These kinematics enhancements might be developed within the framework of the Incompatible Modes Method (see [8]), and require a dedicated solution algorithm which is illustrated next. Using the strain discontinuity permits the proper strain representation of two different sets of elastic properties corresponding to each phase. Using the displacement discontinuity leads to the possibility to model de-bonding or any failure mechanism at the interface. For the latter, two failure mechanisms are considered: one corresponding to the opening of the crack in the normal direction and the second one to the sliding in the tangent direction. Both of these discontinuities are introduced by using the Incompatible Modes Method. The key advantage of this method is to lead to a constant number of global degrees-of-freedom in a structured mesh.

For clarity we address a 2D case, where both of those kinematics enhancements are added on top of the standard CST element (Fig. 2). Hence this element is divided into two parts by introducing an interface whose position is obtained by the intersection of the chosen structured mesh with the inclusions placed within the structure. The domain Ω^e of the standard 3-node constant stress triangle (CST) element is thus divided into two sub-domains Ω^{e^-} and Ω^{e^+}. One of the most important and well-known features of strong (displacement field) discontinuities models is their capability to be independent from the mesh, even in localized failure. This ability is due to the fact that the dissipation process occurs on a line (i.e. the interface) and not in the whole volume. However, different elastic-plastic or elastic-damage behavior laws, with positive hardening, might be chosen for each the two sub-domains split by the interface, with different elastic properties (see Fig. 1). Often the material parameters of interface are given the same properties as the cement paste, but this is not the only possible choice.

It is worth to note that the strain field discontinuity is always present, due to the different elastic constants between the two phases. On the other hand, the displacement field discontinuity needs to be activate only for representing a localized failure mechanism between the two phases that activates according to some chosen failure criterion.

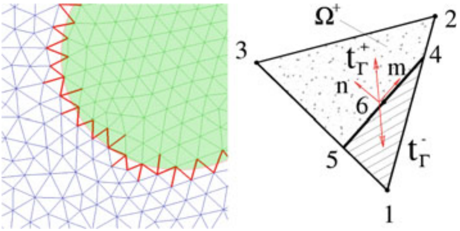


Fig. 2 (i) Structured mesh representation of two-phase interface; (ii) Two phase 3 node triangular element, interface position and element sub-domains with phase I and II

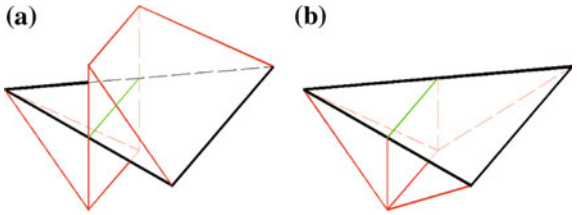


Fig. 3 Incompatible modes corresponding to displacements (**a**) and strain (**b**) discontinuities for CST element

Introducing those discontinuities requires to enhance the kinematics of the element by using two incompatible modes. Thus, the displacements field might be written as follows:

$$\mathbf{u}^h(\underline{x},t) = \sum_{a=1}^{3} N_a(\underline{x})\mathbf{d}_a(t) + \mathbf{M}_{\mathbf{I}}^{\alpha}(\underline{x})\alpha_{\mathbf{I}}(t) + \mathbf{M}_{\mathbf{I}}^{\beta}(\underline{x})\beta_{\mathbf{I}}(t) + \mathbf{M}_{\mathbf{II}}(\underline{x})\alpha_{\mathbf{II}}(t) \tag{1}$$

This expression contains four terms: the first one provides a constant strain field inside the element (as the classical CST element does). The second and third terms both represent jumps in the displacements field, in the normal and the tangential directions. Finally, the last part provides the strain field discontinuity. All the strain and displacement enhancements are limited to a single element only; the latter provides much better basis for constructing robust operator split analysis from the X-FEM method. The shape functions $\mathbf{M}_{\mathbf{I}}(\underline{x})$ for the first incompatible mode (see Fig. 3a) corresponding to the displacements field discontinuity for both normal and tangent directions might be written as:

$$\mathbf{M}_{\mathbf{I}}(\underline{x}) = H_{\Gamma_S}(\underline{x}) - \sum_{a\in\Omega^+} N_a(\underline{x}) \tag{2}$$

where N_a represents the normal CST shape functions element and H_{Γ_S} the Heaviside function placed at the interface position. The shape function $\mathbf{M}_{\mathbf{II}}(\underline{x})$ which provides the jump in the strain field is shown on Fig. 3b.

Considering the displacement interpolation (1), the strain field might be written as:

$$\varepsilon^h(\underline{x},t) = \mathbf{Bd} + \mathbf{G}_{\mathbf{II}}\alpha_{\mathbf{II}} + (\mathbf{n}^T \otimes \mathbf{n})\mathbf{G}_{\mathbf{I}_r}^{\alpha}\alpha_{\mathbf{I}} + \frac{1}{2}\Big[\mathbf{n}^T \otimes \mathbf{m} + \mathbf{m}^T \otimes \mathbf{n}\Big]\mathbf{G}_{\mathbf{I}_r}^{\beta}\beta_{\mathbf{I}} \tag{3}$$

where $\mathbf{B}(\underline{x})$ are the well known CST element strain-displacement matrix (e.g. see [20]) and $\mathbf{G}_{\mathbf{I}_r}(\underline{x})$ contains the derivatives of the first incompatible mode. Finally, in (4), $\mathbf{G}_{\mathbf{II}}$ is the matrix containing the derivatives of the second shape function $\mathbf{M}_{\mathbf{II}}(\underline{x})$

Deriving from the Incompatible Modes Method for the two kind of discontinuities added on the top of the classical CST element (strain field and displacement field), the total system to be solve consists of four equilibrium equations, with (4a) as the global equilibrium equation and (4b–d) are corresponding to the local ones.

$$\begin{cases} A_{e=1}^{nel}\left[f^{int} - f^{ext} = 0\right] \\ \mathbf{h}_{\mathbf{I}}^{\alpha,e} = 0 \\ \mathbf{h}_{\mathbf{I}}^{\beta,e} = 0 \\ \mathbf{h}_{\mathbf{II}}^{e} = 0 \end{cases} \Longrightarrow \begin{cases} \int_{\Omega^e} \mathbf{B}^T \sigma d\Omega - \int_{\Omega^e} \mathbf{N}^T b d\Omega = 0 \\ \int_{\Omega^e} \mathbf{G}_{\mathbf{I}_v}^{\alpha,T} \sigma d\Omega = 0 \\ \int_{\Omega^e} \mathbf{G}_{\mathbf{I}_v}^{\beta,T} \sigma d\Omega = 0 \\ \int_{\Omega^e} \mathbf{G}_{\mathbf{II}}^{T} \sigma d\Omega = 0 \end{cases} \tag{4}$$

It is worth to remind that Eq. (4b, c) have to be solved only in case of activation of the displacement discontinuity in the normal or the tangent direction. The consistent linearization of this set of equations leads to the linear system, in the matrix form:

$$\begin{bmatrix} \mathbf{K}^e & \mathbf{F}_{\mathbf{I}_r}^{\alpha,e} & \mathbf{F}_{\mathbf{I}_r}^{\beta,e} & \mathbf{F}_{\mathbf{II}}^{e} \\ \mathbf{F}_{\mathbf{I}_r}^{\alpha,e^T} & \mathbf{H}_{\mathbf{I}}^{\alpha,e} & \mathbf{F}_{H}^{e} & \mathbf{F}_{S}^{\alpha,e} \\ \mathbf{F}_{\mathbf{I}_r}^{\beta,e^T} & \mathbf{F}_{H}^{e^T} & \mathbf{H}_{\mathbf{I}}^{\beta,e} & \mathbf{F}_{S}^{\beta,e} \\ \mathbf{F}_{\mathbf{II}}^{e,T} & \mathbf{F}_{S}^{\alpha,e^T} & \mathbf{F}_{S}^{\beta,e^T} & \mathbf{H}_{\mathbf{II}}^{e} \end{bmatrix}_{n+1}^{(k)} \begin{pmatrix} \Delta d \\ \Delta\alpha_{\mathbf{I}} \\ \Delta\beta_{\mathbf{I}} \\ \Delta\alpha_{\mathbf{II}} \end{pmatrix}_{n+1}^{(k+1)} = \begin{pmatrix} -r \\ -\mathbf{h}_{\mathbf{I}}^{\alpha,e} \\ -\mathbf{h}_{\mathbf{I}}^{\beta,e} \\ -\mathbf{h}_{\mathbf{II}}^{e} \end{pmatrix}_{n+1}^{(k)} \tag{5}$$

The expanded form for each block can be found in [7].

The operator split strategy consists in first solving the local equations of system (4) (namely equations (4b–d)) at each numerical integration point and for fixed global degrees-of-freedom values. The second step is then to carry out static condensations. This leads to the effective stiffness matrix and and thus the last step is to solve the global system of equations (4) to obtain the updated value of the displacement field $d_{n+1}^{(k+1)} = d_{n+1}^{(k)} + \Delta d_{n+1}^{(k+1)}$.

$$\widehat{\mathbf{K}}_{n+1}^{(k)} \cdot \Delta d_{n+1}^{(k+1)} = -r_{n+1}^{(k)} \tag{6}$$

One of the key point to note is that the total number of global unknowns remains the same as with the standard CST element which is the major advantage of Incompatible Modes Method. Simple illustrative examples dealing with the use of structured meshes might be found in [3]).

Here we aim to make a comparison between structured and unstructured meshes in order to assess the capability for both cases to get very close results. For this we consider a porous material made of a perfectly plastic matrix with circular voids of different sizes. The first case (Fig. 4a) presents an exact mesh obtained by using the software GMSH. Obviously in this case each element contains only one phase (namely the matrix or the "voids"). Moreover several elements are strongly distorted and they exhibit quite different sizes. For these two reasons the stiffness matrix is poorly conditioned. The second case (Fig. 4b) relies on a structured mesh which is based on a regular grid. In this case, the elements needs to represent two phases to model the inclusions and we adopt the strategy presented at the beginning of this Section. Figure 4 shows the axial displacement contour plot (with an amplification factor of 100) for both unstructured and structured meshes. Figure 5 plots the corresponding macroscopic axial reactions displacement curve.

We show that both cases provide the results in a very close agreement, but with a gain of computing time by a factor of 20 in favor of the structured mesh strategy.

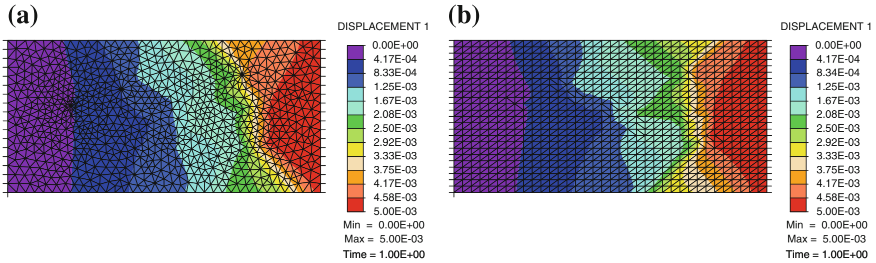
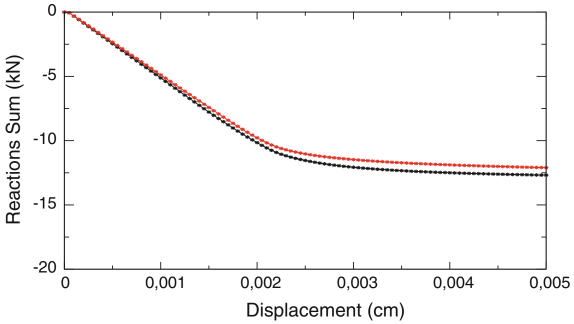


Fig. 4 Longitudinal displacement contour plot corresponding to max.load for adaptive mesh (**a**) and regular mesh (**b**)

Fig. 5 Reactions sum versus displacement curve (*black* unstructured mesh, *red* structured mesh)



This advantage is mainly due to the tangent matrix optimal conditioning. Combined to a meshing process which is much easier, the structured mesh way appears to be a good and accurate method to model heterogeneous material, especially in the context of many realizations that have to be analyzed. This last point is one of the key issues considering probabilistic aspects for heterogeneous materials.

2.3 Probability Aspects of Inelastic Localized Failure for Heterogenous Materials

At finer scale than the macroscopic one, cement-based materials obviously appear to be heterogeneous. As an example, at this meso-scale mortars are made of three phases: two solid ones (the grains and the cement paste) and voids. It is well-known from experimental data that macroscopic properties of such materials are strongly linked to the (at least) meso-scale constituents. Moreover, considering a constant porosity, the voids shapes and positions also have a major influence on the macroscopic properties, especially for small specimens. This key point is linked to the statistical RVE size whose size has to be determined along a prescribed macroscopic error tolerance.

To further illustrate these ides, we consider herein a porous material, typical of mortars at a meso-scale level. At this scale we assume that such material is characterized by a two-phase micro-structure with a stiff phase and a soft phase. The former will be referred as the "matrix" and the latter is supposed to represent the voids or inclusions. Depending on the number of inclusions, their sizes and positions, the non-linear macroscopic response of such a material will vary. In other words, the macroscopic properties, such as Young's modulus or the yield stress, will be influenced by the meso-scale geometry. Our goals here are: first to determine the statistical RVE size corresponding to such a geometry (morphological RVE); second to carry out numerically the variations of the macroscopic characteristics upon the inclusion sizes and positions. The key point for this study is that the variability introduced into the model is restricted to the specimen geometry only, whereas the mechanical characteristics of the two phases are assumed to be deterministic. To be more precise, the matrix phase is supposed to be accurately modelled by an elastic-perfectly plastic model based upon the Drucker-Prager criterion. The voids are represented by a simple linear isotropic elasticity model with very small Young's modulus value. In the following sections we first begin to describe the Gibbs point process, leading to the realizations of the meso-structures. We also show an example of one typical mesh obtained and the corresponding macroscopic response to a tension test. We focus on the notion of Statistical RVE leading to a volume element large enough to assure that its macroscopic properties are assumed to be deterministic, up to a certain tolerance.

We further present the methodology leading to such a RVE definition and discuss the corresponding procedure, with the size corresponding to a rectangular domain ($3.6 \times 1.8\,\text{cm}^2$). The meso-structure geometry of such domain is here supposed to be accurately modelled by a Gibbs point process. Such point process is built on a two steps scheme. The first one is the determination of the inclusions number according to a Poisson law. The second step consists in the determination of the inclusion centers coordinates as well as the radius for each inclusion. While such a Gibbs process already naturally leads to a set of non-intersecting inclusions, we applied an even more restrictive criterion, by choosing the minimal distance between the inclusions (here equal to 2 mm). Moreover, in order to be consistent with the mesh size and the model features, the inclusions radius are bounded between 1 and 3 mm. Figure 6 shows a particular realization of the meso-structure and the corresponding

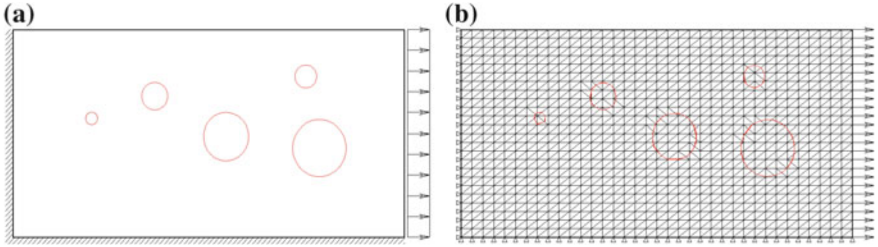


Fig. 6 Meso-structure geometry (**a**) and corresponding structured mesh (**b**)

structured mesh. We can notice that each inclusion is correctly modelled by a set of discontinuities without any major distortion.

Since the material parameters are chosen to be deterministic, the statistics of the macroscopic response depends on the meso-structure geometry only, defined by the voids volume fraction and consequently the voids radius and centers' positions. Thus the macroscopic problem is stochastic and requires stochastic integration method which is presented in the next section. Two approaches can be drawn to find a probability distribution describing a random phenomena. The first one, so-called frequentist approach, is based on statistical tests, like the χ^2 test for the Gaussian probability law. Results of these tests are error margins that evaluate how the outcomes of the given random phenomena fit with respect to a given probability law. The second, so-called Bayesian approach, is trying to use all the available information along with the maximum entropy theory in order to provide the most general probability law for a given state of information; thus, to fully describe this probability law, the statistical moments of different orders have to be computed. In this work, the second approach is chosen. The macroscopic material properties we tend to characterize are all defined on the positive real line. Moreover we assume that they can be given a mean value and a finite standard deviation. On the basic of such information, the maximum entropy theory leads to the most general probability law for this case in terms of the log-normal distribution, which is fully described by its computed mean value and standard deviation.

Thus, the final product of the proposed approach will be a coupled mechanics-probability problem, which is posed at the coarse scale. The fine scale computational results are here used to build the corresponding probability distribution of the material parameters for the coarse scale model defined as random fields. Needless to say, for a massive structures the size of such coupled mechanics-probability model is still very large. The solution are presently available only for 1D case [6], representative of the simple tension test. Thus, we need to provide an efficient solution procedure with suitable reduction method for more general case, which is discussed next.

3 Stochastic Fields Coupling

The classical form of the SSFEM uses only two parameters to describe the solution space: the dimension and the order of the spanning basis functions. For instance properties of the numerical model itself are not taken into account. By using such a classical choice the size of the solution space grows exponentially when increasing the two mentioned parameters. This undesired property is known as the *Curse of Dimension* in the literature.

Low-Rank representations and adaptive solution space techniques have been developed recently for a practical applicability of the SSFEM. In [2] a Low-Rank approximation of the numerical problem to be considered is done on a geometrically coarse mesh to derive afterwards few basis functions to represent the solution efficiently. These basis functions are transferred to the geometrically fine discretised numerical problem to be solved actually. A spectral decomposition of the stochastic solution field is derived from the expectation of the minimal total potential energy principle in [14, 15]. This approach is referred to as the *Generalized Spectral Decomposition* (*GSD*) method. In this context three different numerical algorithms are presented to solve the resulting numerical system. In [12, 19] a Low-Rank representation of the solution field is directly substituted in the fully discretised stationary diffusion equation with stochastic parameter. Higher rank matrices arise due to the composition of the system matrix itself. This rank increase is reduced by applying a truncated *Singular Value Decomposition* (*SVD*). This is done so to say on the fly meaning inside each iteration step of the numerical scheme to solve the problem. A solution space reduction in a preprocessing step is proposed in [1]. Here the geometrical degrees of freedom are virtually reduced to one. The resulting purely stochastic model—named zero-dimensional model—is used to extract the best basis functions (from a predefined set of basis functions) for the actual numerical problem to be solved.

In general the idea of a Low-Rank approximation is orthogonal to a solution space reduction. In contrast to [14, 15] we derive directly a rank-one update scheme from the minimal total potential energy principle. The resulting algorithm is comparable to a special form of the so called restarting algorithm of the GSD, see [14, 15]. However our algorithm in its standard form does not necessarily produce an optimal Low-Rank representation concerning the minimal energy at the given rank. We extend the algorithm to an optional adaptive solution space technique and a rank reduction technique to get an optimal Low-Rank decomposition.

3.1 Heterogeneous Multiscale Model

We consider here a stochastic heterogeneous multiscale model of concrete with uncertain material parameters including microcracks. In order to describe the mechanism of degradation of concrete more than spatial scale is needed. This requires the development of the Heterogeneous Multiscale Method, capable of dealing with different scales simultaneously within a single computation in order to deliver reliable predictions. The source of damage has to be described here on a mesoscale where mirocracks can be resolved and their the spatial distribution analysed. A propagation to the macroscale of this uncertainty is then needed for the stochastic simulation of a massive structure.

In order to give the ideas in a simple notation we consider the linear stationary groundwater flow equation (syntactically equal to the stationary diffusion equation)

$$\nabla_x \kappa(\boldsymbol{x}, \omega) \nabla_{\boldsymbol{x}} u(\boldsymbol{x}, \omega) = f(\boldsymbol{x}) \tag{7}$$

with the uncertain hydraulic head u and the uncertain hydraulic conductivity κ, each of both represented by a stochastic field defined onto the geometrical variable $\boldsymbol{x}$ and the elementary event ω. The source term f and the boundary conditions (not specified explicitly here) are considered to be deterministic. κ is assumed to be non-Gaussian distributed and is available in a truncated *Karhunen-Loève Expansion* (*KLE*), a spectral decomposition of the form $\kappa(\boldsymbol{x}, \omega) \approx \bar{\kappa}(\boldsymbol{x}) + \sum_{i=1}^{m} \sqrt{\lambda_i}\, \kappa_i(\boldsymbol{x})\, \xi_i(\omega)$. $\bar{\kappa}$ is the expected value of κ. λ_i's and κ_i's are the i'th largest eigenvalues and corresponding eigenfunctions of the Fredholm integral of the 2nd kind with the covariance function of κ as kernel. The random variables $\{\xi_i(\omega)\}_{i\in\{1,\dots,m\}}$ are uncorrelated, centered and of unit variance. These variables are discretised each by a truncated so called *Polynomial Chaos Expansion* (*PCE*) in orthogonal Hermite polynomials $\{\psi_\gamma\}_{\gamma\in\mathscr{J}}$ defined onto a Gaussian distributed random vector $\boldsymbol{\theta}$: $\xi_i(\omega) \approx \sum_{\alpha\in\mathscr{J}} \xi_{i,\alpha}\, \psi_\alpha(\theta(\omega))$.

Deriving the weak form of Eq. 7 and using truncated PCEs for the test and the ansatz functions as well as an analogous geometrical discretisation leads to the fully discretised system to be solved. It is given by

$$\sum_{i=0}^{l} \boldsymbol{K}_i\, \boldsymbol{U}\, \boldsymbol{\Delta}_i = \boldsymbol{F} \qquad \Longleftrightarrow \qquad \mathscr{A}(\boldsymbol{U}) = \boldsymbol{F} \tag{8}$$

with $[\boldsymbol{\Delta}_i]_{\alpha,\beta} := \sum_{\gamma\in\mathscr{J}} \sqrt{\lambda_i}\; \xi_{i,\gamma}\; [\boldsymbol{\Delta}_\gamma]_{\alpha,\beta}$ and $[\boldsymbol{\Delta}_\gamma]_{\alpha,\beta} := \langle \psi_\alpha \psi_\beta \psi_\gamma \rangle$. The triple product $\langle \psi_\alpha \psi_\beta \psi_\gamma \rangle$ marks the expected value of the product of the ansatz function ψ_α, the test function ψ_β and the basis function ψ_γ of the discretised parameter κ. $\boldsymbol{K}_i$ is the stiffness matrix, in which the material parameter is declared as the ith eigenvector (FEM discretised eigenfunction) of the KLE of κ. We refer to [10] for more details.

3.2 Variational Low-Rank Approach with Successive Rank-1 Update (VLR-SR1U)

We derive a Low-Rank approach with successive rank-1 update based on a minimisation of the expectation of the total potential energy of a system discretised by the SSFEM. The demonstrative system is the one in Eq. 8. The corresponding scheme and its algorithm are referred as *Variational Low-Rank Approach with Successive Rank-1 Update* (*VLR-SR1U*). It comes out, that the proposed algorithm is equivalent to the restarting algorithm of the *Generalized Spectral Decomposition* (*GSD*) method in [14, 15] with the configuration of stepwise rank-1 updates and without an orthonormalisation and global updating. As known the resultant Low-Rank approximations are suboptimal and may be optimised in additional steps.

For linear operators—like in the considered case of the stationary groundwater flow problem—we introduce a special handling for more efficiency. This does not lead to ill-posed realisations of the material parameter, which may happen otherwise. Furthermore we present a stand-alone optimisation algorithm *VLR-OPT* used to find the optimal solution of an approximation of fixed rank. It is derived by applying VLR-SR1U onto projections of the original system. The algorithm is combined in different manner with the VLR-SR1U scheme to solve the original system. Additionally a stand-alone algorithm is proposed to estimate, whether the introduction of new stochastic basis functions may improve the representation of a current approximation of the solution essentially. The algorithm considers the residual in a solution space artificially enlarged by the new stochastic basis functions. We denote the algorithm as *Residual-Based Solution Space Estimator* (*RBSSE*). RBSSE is combined with the VLR-SR1U scheme for an adaptive construction of the solution space.

The basic VLR-SR1U algorithm, the optimisation algorithm VLR-OPT and the estimator algorithm RBBSE are discussed in Sects. 3.3, 3.4 and 3.5. The last two algorithms are also combined with the first one.

3.3 Basic VLR-SR1U

We derive the VLR-SR1U scheme demonstratively for the system in Eq. 8. The corresponding minimisation problem is given by

$$\mathscr{E}(\boldsymbol{U}) \;:=\; \frac{1}{2}\mathscr{A}(\boldsymbol{U}) \;:\; \boldsymbol{U} \;-\; \boldsymbol{F} \;:\; \boldsymbol{U} \;\longrightarrow\; \min. \tag{9}$$

This formulation is also identified as the variational formulation in the literature. The inner product $A : B := \sum_{i,j} A_{ij} B_{ij}$ with $A, B \in \mathbb{R}^{n_1 \times n_2}$ is known as the *Frobenius* one. The Low-Rank approach is introduced by the ansatz

$$\boldsymbol{U} = \boldsymbol{U}^{-} + \boldsymbol{g}\boldsymbol{h}^{T}. \tag{10}$$

It marks a rank-1 update of a given solution $\boldsymbol{U}^{-}$ by the tensor product of the geometrical update vector $\boldsymbol{g}$ and the stochastic update vector $\boldsymbol{h}^{T}$. The differentiation of the operator $\mathscr{E}$ with respect to $\boldsymbol{U}$ leads to

$$\frac{\partial\mathscr{E}(\boldsymbol{U})}{\partial\boldsymbol{U}}(\delta\boldsymbol{U}) \;=\; \underbrace{(\mathscr{A}(\boldsymbol{U}) \;-\; \boldsymbol{F})}_{=:R(\boldsymbol{U})} \;:\; \delta\boldsymbol{U}. \tag{11}$$

Then $\boldsymbol{U}$ can be perturbed by varying once $\boldsymbol{g}$ and once $\boldsymbol{h}$:

$$\begin{aligned} \frac{\partial\boldsymbol{U}}{\partial\boldsymbol{g}}(\delta\boldsymbol{g}) \;&=\; \delta\boldsymbol{g}\boldsymbol{h}^{T} \\ \frac{\partial\boldsymbol{U}}{\partial\boldsymbol{h}}(\delta\boldsymbol{h}) \;&=\; \boldsymbol{g}\delta\boldsymbol{h}^{T}. \end{aligned}$$

The derivatives of $\mathscr{E}$ with respect to $\boldsymbol{g}$ and $\boldsymbol{h}$ are given by

$$\begin{aligned} \frac{\partial \mathscr{E}(\boldsymbol{U})}{\partial \boldsymbol{g}}(\delta \boldsymbol{g}) &= R_U : (\delta \boldsymbol{g} \boldsymbol{h}^T) = \delta \boldsymbol{g}^T R_U \boldsymbol{h} \\ \frac{\partial \mathscr{E}(\boldsymbol{U})}{\partial \boldsymbol{h}}(\delta \boldsymbol{h}) &= R_U : (\boldsymbol{g} \delta \boldsymbol{h}^T) = \boldsymbol{g}^T R_U \delta \boldsymbol{h} \end{aligned} \tag{12}$$

with $R_U := R(\boldsymbol{U})$. Now the minimisation problem can be rewritten as $\frac{\partial \mathscr{E}(\boldsymbol{U})}{\partial \boldsymbol{g}}(\delta \boldsymbol{g}) \equiv 0$ and $\frac{\partial \mathscr{E}(\boldsymbol{U})}{\partial \boldsymbol{h}}(\delta \boldsymbol{h}) \equiv 0$. Due to a possible arbitrary choice of $\delta \boldsymbol{g}$ and $\delta \boldsymbol{h}$ it follows:

$$\begin{aligned} R_U \boldsymbol{h} &= \boldsymbol{0} \\ \boldsymbol{g}^T R_U &= \boldsymbol{0}. \end{aligned} \tag{13}$$

To obtain an iterative scheme the solution $\boldsymbol{U}$ can be substituted by the ansatz $\boldsymbol{U}_{r+1} = \boldsymbol{U}_r + \boldsymbol{g}_{r+1} \boldsymbol{h}_{r+1}^T$ with rank r. Then Eq. 13 ends up in the coupled system to be solved:

$$\begin{aligned} \mathscr{G}\, \boldsymbol{g}_{r+1} &= \mathfrak{f} \quad \text{with} \quad \mathscr{G} := \sum_{i=0}^{l} \underbrace{\boldsymbol{h}_{r+1}^T \boldsymbol{\Delta}_i \boldsymbol{h}_{r+1}}_{=:d_i} \boldsymbol{K}_i, \quad \mathfrak{f} := (\boldsymbol{F} - \mathscr{A}(\boldsymbol{U}_r))\, \boldsymbol{h}_{r+1}, \\ \mathscr{H}\, \boldsymbol{h}_{r+1} &= \bar{\mathfrak{f}} \quad \text{with} \quad \mathscr{H} := \sum_{i=0}^{l} \underbrace{\boldsymbol{g}_{r+1}^T \boldsymbol{K}_i \boldsymbol{g}_{r+1}}_{=:k_i} \boldsymbol{\Delta}_i, \quad \bar{\mathfrak{f}} := (\boldsymbol{F} - \mathscr{A}(\boldsymbol{U}_r))^T\, \boldsymbol{g}_{r+1}. \end{aligned} \tag{14}$$

Surely $\boldsymbol{U}_r$ is given in its Low-Rank representation.

When the stiffness matrix $\boldsymbol{K}_i = \boldsymbol{K}(\kappa_i)$ is linear with respect to its material κ_i, then matrix $\mathscr{G}$ can be expressed as

$$\mathscr{G} = \sum_{i=0}^{l} d_i \boldsymbol{K}(\kappa_i) = \boldsymbol{K}(\Sigma_{i=0}^{l} d_i \kappa_i).$$

This reduces the computational costs. Furthermore the sum of the material realisations is positive. In contrast a single material realisation is in general indefinite. The latter could become problematic, when deterministic solvers are involved, which permits only positive material parameters to confirm the requirement well-posedness.

The Algorithm 1 describes the basic VLR-SR1U scheme. The coupled system in 14 is numerically solved by an alternating iterative process. The initial guess is chosen randomly. The increment of the rank is done inside the first loop located in code lines 3–22. The alternating iterative process is realised by the second loop located in the code lines 10–15. The current rank-1 vectors are normalised to keep them in bounds.

Algorithm 1 VLR-SR1U

```
$\boldsymbol{H} \leftarrow \emptyset$,    $\boldsymbol{G} \leftarrow \emptyset$
while not accurate enough and max. number of iterations not reached do
  $\boldsymbol{h} \leftarrow$ rand,    $\boldsymbol{g} \leftarrow \boldsymbol{0}$
  while not accurate enough and max. number of iterations not reached do
    $\boldsymbol{h} \leftarrow$ normalise $\boldsymbol{h}$
    $\boldsymbol{g} \leftarrow$ solve  $\mathscr{G}\,\boldsymbol{g} = \mathfrak{f}$
    $\boldsymbol{g} \leftarrow$ normalise $\boldsymbol{g}$
    $\boldsymbol{h} \leftarrow$ solve  $\mathscr{H}\,\boldsymbol{h} = \bar{\mathfrak{f}}$
  end while
  $\boldsymbol{G} \leftarrow [\boldsymbol{G}, \boldsymbol{g}]$
  $\boldsymbol{H} \leftarrow [\boldsymbol{H}, \boldsymbol{h}]$
end while
```

3.4 VLR-OPT: Optimisation of Given Low-Rank Approximation

In this section a stand-alone optimisation algorithm VLR-OPT is introduced to optimise a given Low Rank approximation. It is combined to VLR-SR1U in two distinct configurations to optimise the suboptimal solution of VLR-SR1U: VLR-SR1U-OPT(1) and VLR-SR1U-OPT(2). In the first configuration VLR-OPT is applied on the fly i.e. during each rank update of VLR-SR1U. The second configuration applies VLR-OPT in a post-processing step to optimise the Low Rank approximation at the final rank.

VLR-OPT uses rank-1 update solvers for a subspace iteration. At this the rank itself is not touched. The initial idea of VLR-OPT is to apply the algorithm of VLR-SR1U to get rank-1 updates for $\boldsymbol{G}$ and $\boldsymbol{H}$:

$$\boldsymbol{G} = \boldsymbol{G}^{-} + \boldsymbol{g}\boldsymbol{v}^{T} \ \text{ and } \ \boldsymbol{H} = \boldsymbol{H}^{-} + \boldsymbol{h}\boldsymbol{w}^{T}. \tag{15}$$

Therefore the original Eq. 8 is projected onto $\boldsymbol{H}$:

$$\mathscr{A}(\boldsymbol{U})\boldsymbol{H} \;=\; \boldsymbol{F}\boldsymbol{H} \iff \mathscr{A}_{\boldsymbol{H}}(\boldsymbol{G}) \;=\; \boldsymbol{F}_{\boldsymbol{H}}.$$

Analogously to Sect. 3.3 the application of the mentioned algorithm onto this projection leads to the following coupled system to be solved:

$$\begin{aligned}
\mathscr{G}_{\boldsymbol{H}}\, \boldsymbol{g}_{r+1} = \mathfrak{f}_{\boldsymbol{H}} \ \text{ with } \ \mathscr{G}_{\boldsymbol{H}} &:= \sum_{i=0}^{l} \underbrace{\boldsymbol{v}_{r+1}^{T}\boldsymbol{H}^{T}\boldsymbol{\Delta}_{i}\boldsymbol{H}\boldsymbol{v}_{r+1}}_{=:\bar{d}_i\in\mathbb{R}} \boldsymbol{K}_{i},\ \mathfrak{f}_{\boldsymbol{H}} := \big(\boldsymbol{F}_{\boldsymbol{H}} - \mathscr{A}_{\boldsymbol{H}}(\boldsymbol{G}^{-})\big)\,\boldsymbol{v}_{r+1},\\
\mathscr{V}\, \boldsymbol{v}_{r+1} = \bar{\mathfrak{f}}_{\boldsymbol{H}} \ \text{ with } \ \mathscr{V} &:= \sum_{i=0}^{l} \underbrace{\boldsymbol{g}_{r+1}^{T}\boldsymbol{K}_{i}\boldsymbol{g}_{r+1}}_{=:k_i\in\mathbb{R}} \boldsymbol{H}^{T}\boldsymbol{\Delta}_{i}\boldsymbol{H},\ \bar{\mathfrak{f}}_{\boldsymbol{H}} := \big(\boldsymbol{F}_{\boldsymbol{H}} - \mathscr{A}_{\boldsymbol{H}}(\boldsymbol{G}^{-})\big)^{T}\,\boldsymbol{g}_{r+1}.
\end{aligned} \tag{16}$$

The projection of the original equation onto $\boldsymbol{G}$ is considered likewise:

$$\boldsymbol{G}^T \mathscr{A}(\boldsymbol{U}) \;=\; \boldsymbol{G}^T \boldsymbol{F} \iff \mathscr{A}_{\boldsymbol{G}}(\boldsymbol{H}) \;=\; \boldsymbol{F}_{\boldsymbol{G}}.$$

The corresponding coupled system is given by

$$\begin{aligned}
\mathscr{H}_{\boldsymbol{G}}\, \boldsymbol{h}_{r+1} = \mathfrak{f}_{\boldsymbol{G}} \;\text{ with }\; \mathscr{H}_{\boldsymbol{G}} &:= \sum_{i=0}^{l} \underbrace{\boldsymbol{w}_{r+1}^T \boldsymbol{G}^T \boldsymbol{K}_i \boldsymbol{G} \boldsymbol{w}_{r+1}}_{=:\bar{k}_i \in \mathbb{R}} \boldsymbol{\Delta}_i, \;\; \mathfrak{f}_{\boldsymbol{H}} := \left(\boldsymbol{F}_{\boldsymbol{G}}^T - \mathscr{A}_{\boldsymbol{G}}^T(\boldsymbol{H}^-)\right) \boldsymbol{w}_{r+1}, \\
\mathscr{W}\, \boldsymbol{w}_{r+1} = \bar{\mathfrak{f}}_{\boldsymbol{G}} \;\text{ with }\; \mathscr{W} &:= \sum_{i=0}^{l} \underbrace{\boldsymbol{h}_{r+1}^T \boldsymbol{\Delta}_i \boldsymbol{h}_{r+1}}_{=:d_i \in \mathbb{R}} \boldsymbol{G}^T \boldsymbol{K}_i \boldsymbol{G}, \;\; \bar{\mathfrak{f}}_{\boldsymbol{G}} := \left(\boldsymbol{F}_{\boldsymbol{G}}^T - \mathscr{A}_{\boldsymbol{G}}^T(\boldsymbol{H}^-)\right)^T \boldsymbol{h}_{r+1}.
\end{aligned} \tag{17}$$

These coupled systems are solved alternating to obtain rank-1 updates for $\boldsymbol{G}$ and $\boldsymbol{H}$. The rank-1 updates are tensorially multiplied and added to the current solution. In this manner all ranks are updated and the rank itself is maintained.

VLR-OPT is described by Algorithm 2. The input is given by the Low Rank approximation $\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{G}}\tilde{\boldsymbol{H}}^T$. The energy minimisations located in code lines 3 and 5 denote to solve the coupled systems 16 and 17. These two lines indicate the rank-1 update solvers described in Algorithms 3 and 4. The tensor product and the additions located in lines 4 and 6 are performed to update the rank vectors globally.

Algorithm 2 VLR-OPT

```
1: $\boldsymbol{G} \leftarrow \tilde{\boldsymbol{G}}, \quad \boldsymbol{H} \leftarrow \tilde{\boldsymbol{H}}$
2: while not accurate enough and max. number of iterations not reached do
3:     $\boldsymbol{g}, \boldsymbol{v} \leftarrow$ minimise $\mathscr{E}((\boldsymbol{G} + \boldsymbol{g}\boldsymbol{v}^T)\boldsymbol{H}^T)$
4:     $\boldsymbol{G} \leftarrow \boldsymbol{G} + \boldsymbol{g}\boldsymbol{v}^T$
5:     $\boldsymbol{h}, \boldsymbol{w} \leftarrow$ minimise $\mathscr{E}(\boldsymbol{G}(\boldsymbol{H} + \boldsymbol{h}\boldsymbol{w}^T)^T)$
6:     $\boldsymbol{H} \leftarrow \boldsymbol{H} + \boldsymbol{h}\boldsymbol{w}^T$
7: end while
```

Algorithm 3 minimise $\mathscr{E}((\boldsymbol{G} + \boldsymbol{g}\boldsymbol{v}^T)\boldsymbol{H}^T)$

```
$\boldsymbol{g} \leftarrow$ rand, $\quad \boldsymbol{v} \leftarrow \boldsymbol{0}$
while not accurate enough and max. number of iterations not reached do
    $\boldsymbol{g} \leftarrow$ normalise $\boldsymbol{g}$
    $\boldsymbol{v} \leftarrow$ solve $\mathscr{V}\, \boldsymbol{v} = \bar{\mathfrak{f}}_{\boldsymbol{H}}$
    $\boldsymbol{v} \leftarrow$ normalise $\boldsymbol{v}$
    $\boldsymbol{g} \leftarrow$ solve $\mathscr{G}_{\boldsymbol{H}}\, \boldsymbol{g} = \mathfrak{f}_{\boldsymbol{H}}$
end while
```

Algorithm 4 minimise $\mathscr{E}(\boldsymbol{G}(\boldsymbol{H} + \boldsymbol{h}\boldsymbol{w}^T)^T)$

```
$\boldsymbol{h} \leftarrow$ rand,    $\boldsymbol{w} \leftarrow \boldsymbol{0}$
while not accurate enough and max. number of iterations not reached do
    $\boldsymbol{h} \leftarrow$ normalise $\boldsymbol{h}$
    $\boldsymbol{w} \leftarrow$ solve  $\mathscr{W}\ \boldsymbol{w} = \bar{\mathfrak{f}}_{\boldsymbol{G}}$
    $\boldsymbol{w} \leftarrow$ normalise $\boldsymbol{w}$
    $\boldsymbol{h} \leftarrow$ solve  $\mathscr{H}_{\boldsymbol{G}}\ \boldsymbol{h} = \mathfrak{f}_{\boldsymbol{G}}$
end while
```

3.5 RBSSE: Adaptive Construction of the Stochastic Solution Space

The Low Rank approach reveals a possibility to reduce the computational costs and memory usage. Advantages of the same kinds can be additionally obtained, if a small solution space can be found representing the solution well. Therefore the solution space is adaptively constructed in this work. Relevant stochastic basis functions are selected from a pool of stochastic basis functions by a stand-alone residual-based indicator already introduced as the RBSSE. It is embedded inside the VLR-SR1U scheme with or without the optimisation mentioned in Sect. 3.4. A mutual error control between the rank update of VLR-SR1U and the solution space adaption by RBSSE may be utilised.

RBSSE uses the residual of the discretised problem with respect to the current approximation of the solution. It is explained in the following by means of the fully discretised groundwater flow problem (see Eq. 8). The corresponding residual is defined as

$$\boldsymbol{R} \ := \ \boldsymbol{F} \ - \ \sum_{i=0}^{l} \boldsymbol{K}_i \ \boldsymbol{U} \ \boldsymbol{\Delta}_i . \tag{18}$$

The set of the current stochastic basis functions $\{\psi_\alpha\}_{\alpha \in \mathscr{I}^c}$ $(|\{\psi_\alpha\}_{\alpha \in \mathscr{I}^c}| = N_s$ is now extended by additional ones $\{\psi_{\alpha^+}\}_{\alpha^+ \in \mathscr{I}^{c+}}$ $(|\{\psi_{\alpha^+}\}_{\alpha^+ \in \mathscr{I}^{c+}}| = N_s^+)$. The influence of these new stochastic basis functions onto the quality of the solution is estimated by extending the residual artificially. Therefore each stochastic matrix $\boldsymbol{\Delta}_i$ results in:

$$\boldsymbol{\Delta}_i^{\oplus} := \left(\begin{array}{c|c} \boldsymbol{\Delta}_i & \boldsymbol{\Delta}_i^{\triangleleft} \\ \hline \boldsymbol{\Delta}_i^{\triangleright} & \boldsymbol{\Delta}_i^{+} \end{array} \right) .$$

Matrix $\boldsymbol{\Delta}_i^+$ captures only the additional basis functions. The matrices $\boldsymbol{\Delta}_i^{\triangleleft}$ and $\boldsymbol{\Delta}_i^{\triangleright} = (\boldsymbol{\Delta}_i^{\triangleleft})^T$ describe the coupling between the current and the additional basis functions. The new matrices are defined by $[\boldsymbol{\Delta}_i^+]_{\alpha^+,\beta^+} := \sum_{\gamma \in \mathscr{I}^c} \sqrt{\lambda_i}\ \xi_{i,\gamma}\ [\boldsymbol{\Delta}_\gamma]_{\alpha^+,\beta^+}$ and $[\boldsymbol{\Delta}_i^{\triangleleft}]_{\alpha^+,\beta} := \sum_{\gamma \in \mathscr{I}^c} \sqrt{\lambda_i}\ \xi_{i,\gamma}\ [\boldsymbol{\Delta}_\gamma]_{\alpha^+,\beta}$. Matrix $\boldsymbol{\Delta}_i^{\oplus}$ allows to consider the extended residual:

$$\boldsymbol{R}^{\oplus} \ := \ \boldsymbol{F}^{\oplus} \ - \ \sum_{i=0}^{l} \boldsymbol{K}_i \ \boldsymbol{U}^{\oplus} \ \boldsymbol{\Delta}_i^{\oplus} .$$

The coefficients of the extended solution $\boldsymbol{U}^{\oplus}$ associated with the added basis functions are set to zero: $\boldsymbol{U}^{\oplus} := [\boldsymbol{U}, \boldsymbol{0}]$ with a $N_{\boldsymbol{x}}$-by-N_s^+ zero matrix $\boldsymbol{0}$. The extended right-hand side (RHS) $\boldsymbol{F}^{\oplus}$ results from the concatenation of $\boldsymbol{F}$ and $\boldsymbol{F}^{+}$: $\boldsymbol{F}^{\oplus} := [\boldsymbol{F}, \boldsymbol{F}^{+}]$. $\boldsymbol{F}^{+}$ is the RHS corresponding to the added basis functions. The added zero coefficients of the extended solution allows to simplify the computation of $\boldsymbol{R}^{\oplus}$:

$$\boldsymbol{R}^{\oplus} \;:=\; \boldsymbol{F}^{\oplus} \;-\; \sum_{i=0}^{l} \boldsymbol{K}_i \, \boldsymbol{U} \, \boldsymbol{\Delta}_i^{\uplus} \qquad \text{with} \qquad \boldsymbol{\Delta}_i^{\uplus} := [\boldsymbol{\Delta}_i, \boldsymbol{\Delta}_i^{\triangleleft}].$$

$\boldsymbol{R}^{\oplus}$ may be considered as the concatenation of $\boldsymbol{R}$ and $\boldsymbol{R}^{+}$: $\boldsymbol{R}^{\oplus} = [\boldsymbol{R}, \boldsymbol{R}^{+}]$. $\boldsymbol{R}^{+}$ is the residual associated with the additional basis functions: $\boldsymbol{R}^{+} := \boldsymbol{F}^{+} \;-\; \sum_{i=0}^{l} \boldsymbol{K}_i \, \boldsymbol{U} \, \boldsymbol{\Delta}_i^{\triangleleft}$, $\boldsymbol{R}^{+} \in \mathbb{R}^{N_x \times N_s^+}$.

$\boldsymbol{R}^{+}$ is used to obtain a reasonable indicator to rate the basis functions with respect to their ability to describe the solution. For it $\boldsymbol{R}^{+}$ is translated to the corresponding unit of measurement. The best possible translation is specified by the indicator $\mathfrak{i}$:

$$\mathfrak{i} \;:=\; (\boldsymbol{A}^{+})^{-1} \, \boldsymbol{r}^{+} \;\text{ with }\; \boldsymbol{A}^{+} := \sum_{i=0}^{l} \boldsymbol{\Delta}_i^{+} \;\otimes\; \boldsymbol{K}_i . \tag{19}$$

$\boldsymbol{r}^{+} \in \mathbb{R}^{N_x \cdot N_s^+}$ is the proper vectorisation of matrix $\boldsymbol{R}^{+}$. In practice the symmetric matrix $(\boldsymbol{A}^{+})^{-1}$ is not computed. Instead of using $(\boldsymbol{A}^{+})^{-1}$ we use the matrix $\boldsymbol{J} := (diag(\boldsymbol{A}^{+}))^{-1}$ known as the Jacobi pre-conditioner in the context of pre-conditioning. The matrix operator $diag$ maintains only the diagonal entries while the other entries are set to zero. The corresponding indicator is specified by $\mathfrak{i}_{\boldsymbol{J}} \;:=\; \boldsymbol{J} \, \boldsymbol{r}^{+}$. The indicator can be also as a matrix operator $\mathfrak{I}_{\boldsymbol{J}} \;:=\; \hat{\boldsymbol{J}} \odot \boldsymbol{R}^{+}$. $\hat{\boldsymbol{J}}$ is the proper matrix notation of the diagonal of $\boldsymbol{J}$; the operator $\odot$ marks the component-wise matrix product.

For each stochastic basis function the indicator $\mathfrak{I}_{\boldsymbol{J}}$ contains as many values as geometrical degrees of freedom. These values are reduced to one value by the L_2-norm to rate the corresponding stochastic basis function. This is reflected by function $r(\mathfrak{I}_{\boldsymbol{J}}) := [\; ||(\mathfrak{I}_{\boldsymbol{J}})_1||_{L_2} \,, \ldots, \; ||(\mathfrak{I}_{\boldsymbol{J}})_{N_s^+}||_{L_2} \;] \;\in \mathbb{R}^{N_s^+}$; $(\mathfrak{I}_{\boldsymbol{J}})_j$ means the j's column of matrix $\mathfrak{I}_{\boldsymbol{J}}$.

Algorithm 5 describes the RBSSE. The Jacobi pre-conditioner may be replaced by another one. Also the reduction function r may be chosen differently. The list of required input arguments is not complete, but shows the essential arguments. The selection of the relevant additional stochastic basis functions in code line 7 provides another option: e.g. 10 % of the best rated additional basis functions could be selected.

The RBSSE is now applied inside the Algorithm 1 of the VLR-SR1U scheme to obtain an adaptive construction of the solution space. The resulting Algorithm 6 is referred as *VLR-SR1U-ADAPT*. The input arguments are given by an initial set $\mathscr{I}^c$ of already accepted stochastic basis functions and a set $\mathscr{I}^{c^+}$ of potentially relevant stochastic basis functions. $\mathscr{I}^{c^+}$ may be also generated on the fly, that means it may vary in the rank iterations. That makes sense, as otherwise the rating of the set of additional stochastic basis functions may become too memory and time consuming.

Algorithm 5 RBSSE

```
Require: $\mathscr{I}^c$, $\mathscr{I}^{c^+}$, $\boldsymbol{F}^+$
1: $\{\boldsymbol{\Delta}_i^{\triangleleft}\}_{i\in\{0,\dots,l\}} \leftarrow$ construct( $\mathscr{I}^c$, $\mathscr{I}^{c^+}$ )
2: $\{\boldsymbol{\Delta}_i^{+}\}_{i\in\{0,\dots,l\}} \leftarrow$ construct( $\mathscr{I}^{c^+}$ )
3: $\boldsymbol{R}^+ \leftarrow$ compute( $\{\boldsymbol{\Delta}_i^{\triangleleft}\}_{i\in\{0,\dots,l\}}$, $\boldsymbol{F}^+$ )
4: $\hat{\boldsymbol{J}} \leftarrow$ compute( $\{\boldsymbol{\Delta}_i^{+}\}_{i\in\{0,\dots,l\}}$ )
5: $\mathfrak{I}_{\hat{\boldsymbol{J}}} \leftarrow$ compute( $\boldsymbol{J}$, $\boldsymbol{R}^+$ )
6: $r(\mathfrak{I}_{\boldsymbol{J}}) \leftarrow$ reduce( $\mathfrak{I}_{\boldsymbol{J}}$ )
7: $\mathscr{I}^+ \leftarrow$ select( $r(\mathfrak{I}_{\boldsymbol{J}})$ )
8: return $\mathscr{I}^+$
```

The set $\mathscr{I}^+$ contains the current best rated stochastic basis functions, which are added to the current set $\mathscr{I}^c$ in the next rank iteration (see code line 5). The adaptive construction of the solution space is presented in the code lines 4–7. The system update in code line 6 means to extend the entire system to the added stochastic basis functions. The rating and selection of the additional stochastic basis functions happen 19–21. Optionally the VLR-OPT Algorithm 2 can be involved also e.g. in code line 18 for an OPT(1) configuration or after code line 22 for an OPT(2) configuration. A combination is also demonstrated in Sect. 3.6.3.

An interplay between the residuals $\boldsymbol{R}$ and $\boldsymbol{R}^+$ may be observed. $||\boldsymbol{R}|| > ||\boldsymbol{R}^+||$ (with a suitable matrix norm $||\cdot||$) may identify, that more iterations are reasonable to reach more accuracy before adapting the stochastic solution space. $||\boldsymbol{R}|| < ||\boldsymbol{R}^+||$ may identify, that an adaption of the solution space is required to reach more accuracy.

3.6 Numerical Experiments

In the following subsections numerical experiments are presented demonstrating the behaviour of the proposed basic scheme VLR-SR1U, its optimised extensions VLR-SR1U-OPT(1) and VLR-SR1U-OPT(2) as well as its adaptive extension VLR-SR1U-ADAPT. A combination of adaption and optimisation is included also.

The underlying problem to be solved is identified by the stationary groundwater flow equation defined on a rectangular domain with uncertain hydraulic conductivity κ. The left and right boundaries are specified by Dirichlet conditions equal zero and one; the top and bottom boundaries are specified by Neumann conditions equal zero. Sources and sinks are not introduced. The number of geometrical DoFs is $N_{\boldsymbol{x}} = 21 \times 11 = 231$. κ is considered to be lognormal distributed and obtained by an exponentiation of a Gaussian field γ: $\kappa := e^{\gamma}$. γ is defined in five stochastic variables, which marks simultaneously the stochastic dimension. It exhibits a variance of 1.0 and a correlation lengths of 20.0, which is a fifth of the maximal length of the geometrical domain. The representation of κ—used in the following experiments—is described by a KLE capturing approximately 99.99 % of the information content of γ. Unless otherwise stated the maximal polynomial order of the stochastic basis

Algorithm 6 VLR-SR1U-ADAPT

```
Require: $\mathscr{I}^c$, $\mathscr{I}^{c^+}$
$\mathscr{I}^+ \leftarrow \emptyset$
$\boldsymbol{H} \leftarrow \emptyset, \quad \boldsymbol{G} \leftarrow \emptyset$
while not accurate enough and max. number of iterations not reached do
  if $\mathscr{I}^+ \neq \emptyset$ then
    $\mathscr{I}^c \leftarrow [\mathscr{I}^c, \mathscr{I}^+]$
    update system
  end if

  $\boldsymbol{h} \leftarrow$ rand, $\quad \boldsymbol{g} \leftarrow$ init
  while not accurate enough and max. number of iterations not reached do
    $\boldsymbol{h} \leftarrow$ normalise $\boldsymbol{h}$
    $\boldsymbol{g} \leftarrow$ solve $\mathscr{G}\, \boldsymbol{g} = \mathfrak{f}$
    $\boldsymbol{g} \leftarrow$ normalise $\boldsymbol{g}$
    $\boldsymbol{h} \leftarrow$ solve $\mathscr{H}\, \boldsymbol{h} = \bar{\mathfrak{f}}$
  end while
  $\boldsymbol{G} \leftarrow [\boldsymbol{G}, \boldsymbol{g}]$
  $\boldsymbol{H} \leftarrow [\boldsymbol{H}, \boldsymbol{h}]$

  $\boldsymbol{F}^+ \leftarrow$ compute( $\mathscr{I}^{c^+}$ )
  $\mathscr{I}^+ \leftarrow$ RBSSE( $\mathscr{I}^c$, $\mathscr{I}^{c^+}$, $\boldsymbol{F}^+$ )
  $\mathscr{I}^{c^+} \leftarrow \mathscr{I}^{c^+} \backslash \mathscr{I}^+$
end while
```

functions for the solution is fixed to 4. This leads to a full rank of 126 with a standard choice of stochastic basis functions. The discretion in the geometrical and stochastic spaces is obviously coarse. This allows to solve the discretised problem resulting from basic SSFEM—with a standard set of stochastic basis functions—directly. The corresponding solution is exactly the one to which the solution of the VLR-SR1U scheme (and its extensions) should converge. An numerical analysis till to machine precision is therefore possible. The convergence is demonstrated by considering statistic of moments of a scalar-valued purely stochastic function applied onto the solution. This function integrates over the geometrical space. Full rank considerations are done for analytical purposes.

Section 3.6.1 contains numerical results concerning the basic VLR-SR1U scheme. The two VLR-SR1U-OPT configurations are discussed in Sect. 3.6.2. VLR-SR1U-ADAPT is focused in Sect. 3.6.3.

3.6.1 Basic VLR-SR1U

This subsection discusses the numerical results of the basic VLR-SR1U scheme towards the mentioned model problem. The convergences of the relative error, the residual (more precisely its L_2 norm) and the minimisation of the expectation of the energy associated with successive rank-1 updates are presented in Fig. 7. Furthermore the number of required iterations inside a rank-1 update is shown. An

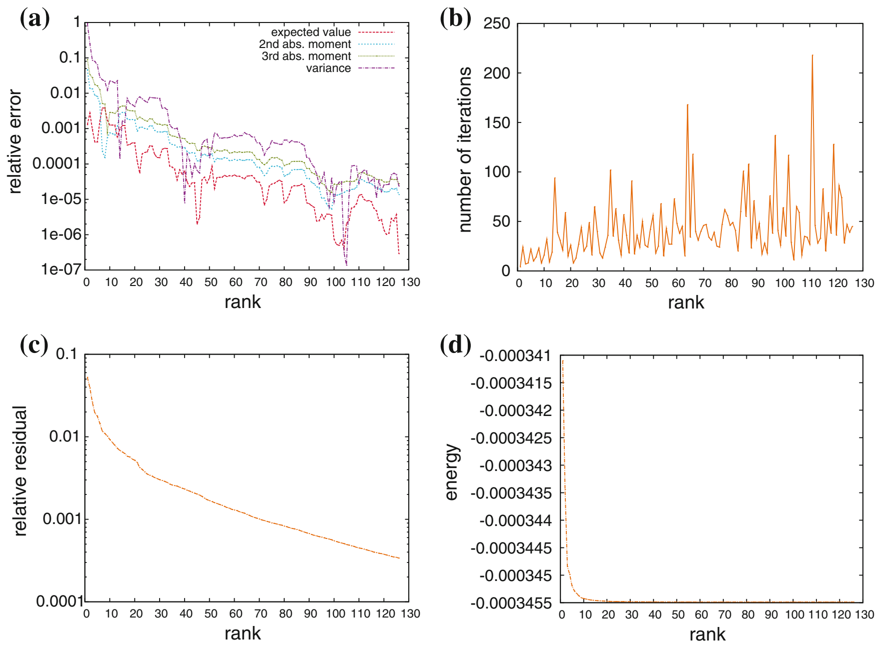


Fig. 7 Convergence plots of VLR-SR1U with tolerance bound 10^{-3}: (**a**) shows the relative error over the current rank. The number of iterations, the residual and the expectation of the energy are plotted in (**b**), (**c**) and (**d**) over the current rank

iteration means here to update once the current stochastic rank-1 vector and once the current geometrical rank-1 update. The corresponding loop starts in code line 10 of the VLR-SR1U Algorithm 1; a maximal number of iterations and a tolerance bound mark the break criterion. The maximal number is chosen, so that the tolerance bound—set to 10^{-3}—is always reached before. The mentioned suboptimality of the basic VLR-SR1U scheme is obvious: errors are approximately between 10−4 and 10−6 (depending on the statistic moment) at full rank.

The influence of the tolerance bound on the quality of the solution shall be considered now. For this the bound is varied from 10^{-1} to 10^{-5}. The interesting quantities are compared in table of Fig. 8. The variation of the tolerance bound does not seem to have an essential effect onto the relative errors. The residual and the energy is only worse for the bound 10^{-1}. The number of iterations per rank-1 update mark the largest difference between the tolerance bounds: an error bound of 10^{-5} performs 9507 iterations till to full rank in contrast to 683 iterations for the error bound 10^{-1}. Consequently a choice may tend to larger error bounds, e.g. 10^{-2}.

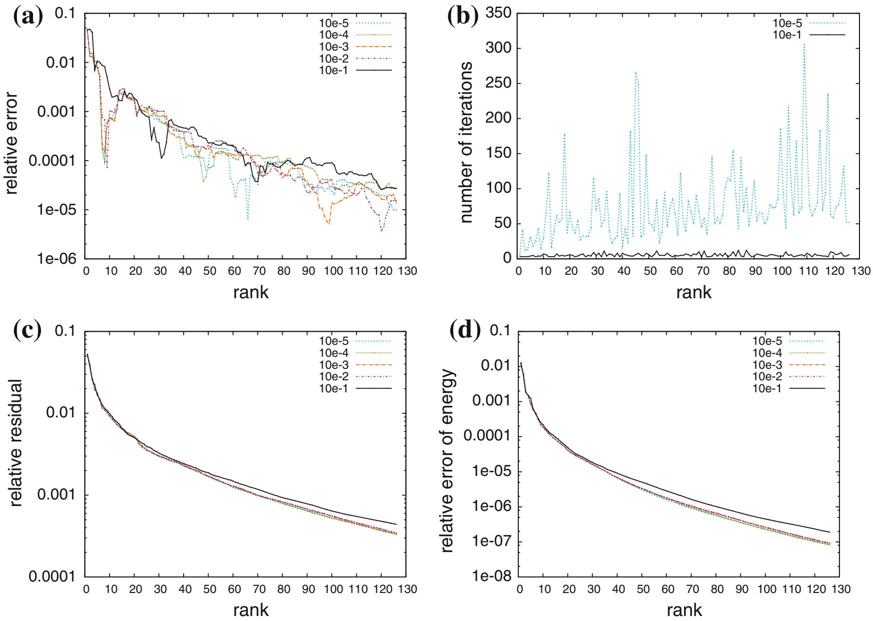


Fig. 8 Convergence plots of VLR-SR1U for different tolerance bounds 10^{-1} to 10^{-5}: (**a**) shows the relative error of the 2nd absolute stochastic moment over the current rank. The number of iterations, the residual and the relative error of the expectation of the energy are plotted in (**b**), (**c**) and (**d**) over the current rank. The reference for the energy is taken from the directly solved SSFEM solution of the identical problem

3.6.2 VLR-SR1U-OPT

As mentioned the basic VLR-SR1U scheme leads to suboptimal Low-Rank approximation. The configuration VLR-SR1U-OPT(1) optimises on the fly—that means during each rank-1 update—by applying the optimisation algorithm VLR-OPT; the configuration VLR-SR1U-OPT(2) applies the optimisation algorithm VLR-OPT at the end of a preferred rank so to speak in a post-processing step. The numerical results for both configurations are summarised in the table of Fig. 9. VLR-SR1U-OPT(1) is compared to the basic VLR-SR1U scheme in figure (a); figure (b) compares VLR-SR1U-OPT(2) to VLR-SR1U-OPT(1). The entire number of iterations—given in figure (b)—is specified in the next passage. The convergence behaviour is discussed afterwards.

One single iteration means in the following to update once the current left rank-1 vector and once the corresponding right rank-1 vector. Inside the basic VLR-SR1U scheme these vectors are the geometrical vector $\boldsymbol{g}$ and the stochastic vector $\boldsymbol{h}$ corresponding to the current rank-1 update. Inside the VLR-OPT scheme these vectors are either the two rank-1 vectors $\boldsymbol{g}$ and $\boldsymbol{v}$ or the two rank-1 vectors $\boldsymbol{h}$ and $\boldsymbol{w}$.

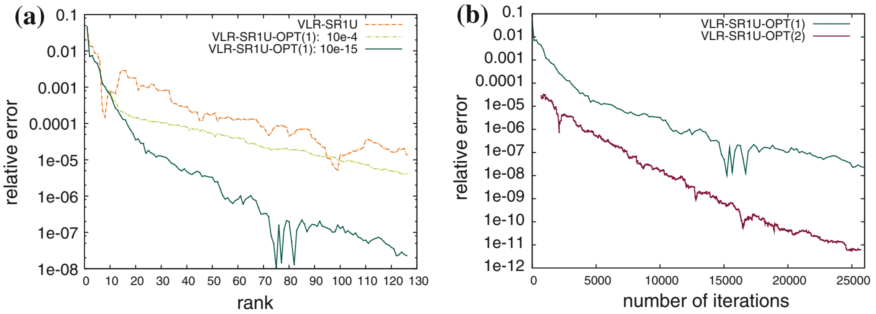


Fig. 9 VLR-SR1U-OPT convergence plots: (**a**) shows the relative error of the 2nd absolute stochastic moment over the current rank. Here VLR-SR1U-OPT(1) is compared to the basic VLR-SR1U scheme. The latter has a tolerance bound of 10^{-3}. VLR-SR1U-OPT(1) uses the following setting per optimisation step: a maximal number of iteration of 10 and the two different tolerance bounds 10^{-4} and 10^{-15}. (**b**) Shows the relative error over the current number of entire iterations. Here VLR-SR1U-OPT(1) with tolerance bound of 10^{-15} is compared to VLR-SR1U-OPT(2)

Whenever one of these three pairs of rank-1 vectors are computed, the iteration counter is incremented by one. This leads to the entire number of iterations.

Figure (a) compares VLR-SR1U-OPT(1) to the basic VLR-SR1U scheme. The break criterion of the basic VLR-SR1U scheme is given by a tolerance bound of 10^{-3} (the maximal number of iterations is never reached). VLR-SR1U-OPT(1) is considered in two different settings. Both use the tolerance bound of 10^{-3} and a maximal number of iterations of 10 for the intrinsic VLR-SR1U calls. The optimisation step is specified differently: the tolerance bound is defined either by 10^{-4} (setting A) or by 10^{-15} (setting B); the maximal number of optimisation steps is 10. VLR-SR1U-OPT(1) with setting A shows better convergence behaviour only till to rank 20 in comparison to the basic VLR-SR1U scheme. Afterwards the optimisation is almost not performed, as the larger tolerance bound of 10^{-4} is already reached after one optimisation step for a rank >20. VLR-SR1U-OPT(1) with setting B enforces almost always to pass all 10 optimisation steps in each rank-1 update. The convergence decreases slightly at higher ranks, as higher ranks may have a higher demand on the optimisation.

Figure (b) compares VLR-SR1U-OPT(2) to VLR-SR1U-OPT(1) with setting A. In the first step VLR-SR1U-OPT(2) applies the basic VLR-SR1U scheme with a large tolerance bound of 10^{-1}. In this way it obtains quickly an approximation of full rank. In the second step this approximation is optimised by VLR-OPT. The performed 1250 optimisation steps provide a precision of approximately 10^{-11}. Consequently VLR-SR1U-OPT(2) is preferable to the VLR-SR1U-OPT(1) concerning the convergence. However the final rank need to be known in contrast to VLR-SR1U-OPT(1).

3.6.3 VLR-SR1U-ADAPT

This subsection discusses numerical results of the VLR-SR1U-ADAPT scheme. The strategy to select new basis functions is here a simple one: the rank-1 solution space is spanned only by the constant stochastic basis function; a new stochastic basis function is added at each rank-1 increment. The new basis function is the one best rated by the RBSSE algorithm. As a consequence the solution space at rank r is spanned by r stochastic basis functions. The convergence is demonstrated in the Fig. 10. Here VLR-SR1U-ADAPT is compared to the basic VLR-SR1U scheme.

While the basic VLR-SR1U scheme uses 56 stochastic basis functions up to order $p = 3$, the VLR-SR1U-ADAPT scheme chooses from a set of 126 stochastic basis functions with maximal order 4. The reference of the solution is obtained by solving the corresponding problem directly with stochastic basis functions of maximal order 6. Figure (a) shows similar convergence rates for the 2nd absolute stochastic moment. The *amount of information of the solution* denotes the number of floating point numbers required to represent a Low-Rank solution. This is defined by a function of rank r:

$$a(r) := \begin{cases} r(N_{\boldsymbol{x}} + N_s) & : \quad \text{fixed solution space} \\ rN_{\boldsymbol{x}} + \sum_{i=1}^{r} i & : \quad \text{solution space successively incremented by 1.} \end{cases} \tag{20}$$

The first case fits to the basic VLR-SR1U scheme; the second case fits to the VLR-SR1U-ADAPT scheme with the mentioned configuration. The rank is considered till 56, which marks the full rank for the approximation of the basic VLR-SR1U scheme. The sets of the stochastic basis functions corresponding to VLR-SR1U and VLR-SR1U-ADAPT at this full rank are compared in Table 1. It can be observed, that sometimes VLR-SR1U-ADAPT preferred to use stochastic basis functions of order 4. An optimisation of both rank 56 approximations by *VLR-SR1U-ADAPT-OPT(2)*

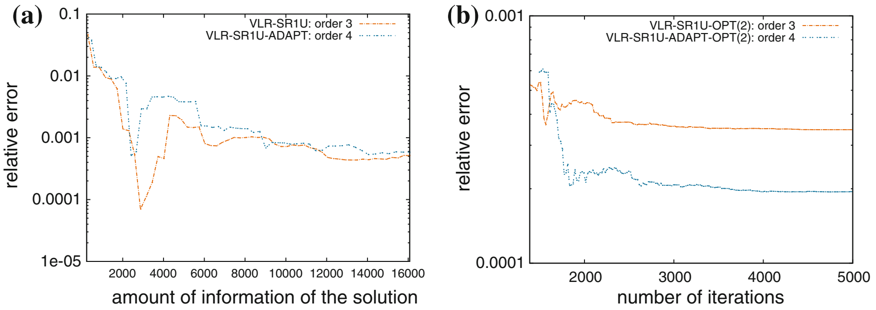


Fig. 10 VLR-SR1U-ADAPT: The relative error over the current amount of information (see Eq. 20) is shown in (**a**) for the 2nd absolute stochastic moment. At this VLR-SR1U with stochastic basis functions of maximal order 3 is compared to VLR-SR1U-ADAPT with stochastic basis functions of maximal order 4. (**b**) Shows the relative error of the 2nd absolute stochastic moment corresponding to an optimisation of the approximation of rank 56 obtained by VLR-SR1U and VLR-SR1U-ADAPT

Table 1 Number of stochastic basis functions per order for VLR-SR1U and VLR-SR1U-ADAPT at rank 56

	Order of basis functions				
	0	1	2	3	4
VLR-SR1U	1	5	15	35	0
VLR-SR1U-ADAPT	1	5	15	26	9

reveals, that the approximation of VLR-SR1U-ADAPT is the more accurate one, see figure (b).

Contrary to the second algorithm the first one does not necessarily provide an optimal decomposition regarding the minimum energy at a given rank. The algorithms are extended by an adaptive technique to construct the solution space a posteriori. In the latter case the residual is used to indicate, which stochastic degrees of freedom are the most promising ones for representing the solution. First numerical experiments show promise. However an adjustment is required.

The numerical behaviour of the VLR-SR1U and its different configurations is discussed in the previous subsection. In summary the following statements can be made. The approximations of the basic VLR-SR1U scheme are suboptimal: at full rank the bounds of the relative errors are between 10^{-4} and 10^{-6} depending on the considered quantity, see Sect. 3.6.1. A single rank vector is on itself optimal, but the solution does not minimise the energy with respect to all matrices of the same rank. This is faced by the scheme VLR-OPT to optimise a given Low-Rank approximation. The VLR-SR1U-OPT(1) scheme uses VLR-OPT during each rank-1 update. In this way the mentioned error bound is dropped down. It may decrease till machine precision, but this depends on the settings of the break criterion for the optimisation loop, see Sect. 3.6.2. VLR-SR1U-OPT(2) applies VLR-OPT only at the final rank: in comparison to VLR-SR1U-OPT-(1) better convergences can be observed, see Sect. 3.6.2. The disadvantage of VLR-SR1U-OPT(2) is the requirement of an a priori specification of the final rank. The VLR-SR1U-ADAPT scheme extends the basic VLR-SR1U scheme to an adaptive construction of the solution space. While the stochastic basis functions of VLR-SR1U were limited by order 3, VLR-SR1U-ADAPT could also select stochastic basis functions of order 4. For an approximation of a fixed rank, it could be observed, that VLR-SR1U-ADAPT chose stochastic basis functions of order 4 by leaving out some of order 3. However VLR-SR1U and VLR-SR1U-ADAPT showed similar convergences. Then the VLR-OPT scheme was applied on both approximations. In this way it could be shown, that the adaptively chosen stochastic basis functions were describing the solution more accurately than the a priorly fixed ones from VLR-SR1U, see Sect. 3.6.3.

References

1. Marcel Bieri and Christoph Schwab. Sparse high order FEM for elliptic sPDEs. *Computer Methods in Applied Mechanics and Engineering*, 198(13–14):1149–1170, March 2009.
2. Alireza Doostan, Roger G. Ghanem, and John Red-Horse. Stochastic model reduction for chaos representations. *Computer Methods in Applied Mechanics and Engineering*, 196(37–40):3951–3966, August 2007.
3. Martin Hautefeuille, Sergy Melnyk, J.B. Colliat, and Adnan Ibrahimbegovic. Probabilistic aspects of localized failure of massive heterogeneous structures. *Int. J. Engineering Computations*, 26:168–194, 2009.
4. Gunter Hofstetter and Herbert Mang. *Computational Mechanics of Reinforced Concrete Structures*. Vieweg, 1995.
5. A. Ibrahimbegovic, A. Boulkertous, L. Davenne, and D. Brancherie. Modelling of reinforced-concrete structures providing crack-spacing based on x-fem, ed-fem and novel operator split solution procedure. *Int. J. Numerical Methods in Engineering*, 83:452–481, 2010.
6. A. Ibrahimbegovic and H.G. Matthies. Probabilistic multiscale analysis of inelastic localized failure in solid mechanics. *Computer Assisted Methods in Engineering and Science*, 19:277–304, 2012.
7. A. Ibrahimbegovic and S. Melnyk. Embedded discontinuity finite element method for modeling of localized failure in heterogeneous materials with structured mesh: an alternative to the extended finite element method. *Computational Mechanics*, 40:149–155, 2007.
8. Adnan Ibrahimbegovic. *Nonlinear Solid Mechanics: Theoretical Formulation and Finite Element Solution Methods*. Springer, 2009.
9. Adnan Ibrahimbegovic and Damijan Markovic. Strong coupling methods in multi-phase and multi-scale modeling of inelastic behave ior of heterogeneous structures. *Computer Methods in Applied Mechanics and Engineering*, 192:3089–3107, 2003.
10. Andreas Keese. *Numerical Solution of Systems with Stochastic Uncertainties - A General Purpose Framework for Stochastic Finite Elements*. PhD thesis, Technische Universität Braunschweig, 2005.
11. Hermann Matthies. Uncertainty quantification with stochastic finite elements. (*eds. E. Stein et al.) Encyclopedia of Computational Mechanics*, pages 1–70, 2007.
12. Hermann G. Matthies and Elmar Zander. Sparse Representations in Stochastic Mechanics. *SEECCM, Rhodos, Greece*, 2009.
13. R. Niekamp, D. Markovic, A. Ibrahimbegovic, H.G. Matthies, and R. L. Taylor. Multi-scale modeling of heterogeneous structures with inelastic constitutive behavior: Part ii software coupling implementation aspects. *Int. J. Engineering Computations*, 26:6–26, 2009.
14. Anthony Nouy. A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 196(45–48):4521–4537, September 2007.
15. Anthony Nouy. Generalized spectral decomposition method for solving stochastic finite element equations: Invariant subspace problem and dedicated algorithms. *Computer Methods in Applied Mechanics and Engineering*, 197(51–52):4718–4736, October 2008.
16. B.V. Rosic, A. Kucerova, J. Skora, O. Pajonk, A. Litvinenko, and H.G. Matthies. Parameter identification in a probabilistic setting. *Engineering Structures*, 50:170–196, 2013.
17. Erwin Stein and Dimtri Tihomirov. Anisotropic damage-plasticity modelling of reinforced concrete. *Proceedings ECCM99-European Conference on Computational Mechanics*, 1:1–19, 1999.
18. Peter Wriggers and S. Moftah. Mesoscale models for concrete: Homogenisation and damage behaviour. *Finite Elements in Analysis and Design*, 42:623–636, 2006.
19. Elmar Zander. Tensor product methods for SPDEs. *Jahresbericht 2009, Institut für Wissenschaftliches Rechnen, Technische Universität Braunschweig*, 2010.
20. Olgierd C. Zienkiewicz and Robert Leroy Taylor. *The Finite Element Method, I, II, III*. Butterworth Heinemann, Oxford, 5 edition, 2001.

Modelling of Internal Fluid Flow in Cracks with Embedded Strong Discontinuities

Mijo Nikolic, Adnan Ibrahimbegovic and Predrag Miscevic

Abstract This chapter presents a discrete approach for modelling failure of heterogeneous rock material with discrete crack propagation and internal fluid flow through the saturated porous medium, where the coupling conditions between the solid and fluid phase obey the Biot's porous media theory. Discrete cracks and localized failure mechanisms are provided through the concept of embedded discontinuity FEM. Furthermore, the basis for presented discrete 2D plane strain model representation of heterogeneous material consisting of material grains, is an assembly of Voronoi cells that are kept together by cohesive links in terms of Timoshenko beams. Embedded discontinuities are built in cohesive links thus providing the discontinuity propagation between the rock grains in mode I and mode II. The model can also take into account the fracture process zone with pre-existing microcracks coalescence prior to the localized failure. Several numerical simulations are given to illustrate presented discrete approach.

1 Introduction

Cracks and other localized failure mechanisms in rocks and other heterogeneous materials represent the dominant failure mechanisms, which occur often in civil engineering practice like in dam failure, foundation collapse, stability of excava-

M. Nikolic
LMT Cachan, École Normale Supérieure de Cachan, 61 Avenue du Président Wilson,
94230 Cachan, France
e-mail: mijo.nikolic@gradst.hr

M. Nikolic · P. Miscevic
University of Split, FCEAG Split, Matice hrvatske 15, 21000 Split, Croatia
e-mail: predrag.miscevic@gradst.hr

A. Ibrahimbegovic (✉)
Laboratoire Roberval de Mecanique, Centre de Recherche Royallieu,
Chair for Computational Mechanics, Sorbonne Universitès/UT Compiègne,
60200 Compiègne, France
e-mail: adnan.ibrahimbegovic@utc.fr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_12

tions, slopes and tunnels, landslides and rock falls. The risk of localized failure should be better understood in order to be prevented. The localized failure in rocks is usually characterized by a sudden and brittle failure without warning in a sense of larger and visible deformations prior to failure. This happens also under the strong influence of material heterogeneities, pre-existing cracks and other defects. Numerical modelling represents a main approach in engineering design and research with the the simulations standing as significant tool for obtaining more insight into the full control of material behaviour.

The fluid flow through deformable porous rock medium additionally modifies its mechanical properties and failure response. Two coupling mechanisms play the key role in the fluid-structure interaction problem of this kind: the first concerns the influence of of pore pressure increase inducing the material dilation, and the second pertains to compressive mechanical stress leading to an increase of a pore pressure and making the material less compliant than in the fully-drained case. This problem has received a great attention in engineering literature. The elastic and (homogenized) plastic hardening response was addressed in pioneering works of Terzaghi and Biot [1, 2] and in more recent contribution [3].

Proper modeling of localized failure demands different approach than continuum approach used in usual engineering tasks, where Finite Element Method (FEM) has been considered as the main tool for solving vast majority of applications [4–6]. In order to provide a reliable predictive model for failure of rocks, the discontinuous solutions should be found, where pre-existing cracks continue to form into new ones during the increased loading leading to failure. The evolution of crack patterns shows that localization is a key factor inducing brittle failure. Thus, the main challenge tackled is to provide enhanced predictive models for localized failure by taking into account the material heterogeneities and pre-existing cracks.

Two notable enhanced methods derived from the standard framework of Finite Element Method (FEM) to deal with localization, i.e. cracks, discontinuities. The first one is the Finite Element Method with Embedded Discontinuities (ED-FEM), representing cracks truly in each element (e.g. see [7–10]). The second one is Extended Finite Element Method (X-FEM) where cracks are represented globally [11–13]. The same methods have been used recently for simulating the localized failure when fluid flows through the porous domain. Namely, X-FEM has been used in simulating hydraulic fracturing of fully-saturated [14] and partially-saturated [15] porous media with cohesive cracks, as well as in saturated shear band formations [16]. The fluid saturated poro-elastic and poro-plastic medium with localized failure zones have been simulated with ED-FEM in [17, 18], while the partially saturated medium can be found in [19]. Another approach for simulating the failure of porous fractured media is with automatic mesh refinement process presented in [20], which was also extended to 3D situation in [21].

This chapter presents an approach for modelling the localized failure in rocks under the influence of heterogeneities and pre-existing defects like found in [22, 23]. The class of discrete lattice models have been chosen for general framework of the numerical model that have been previously used in simulating the progressive failure of concrete and rocks [24]. Namely, the basis of this framework is in representation

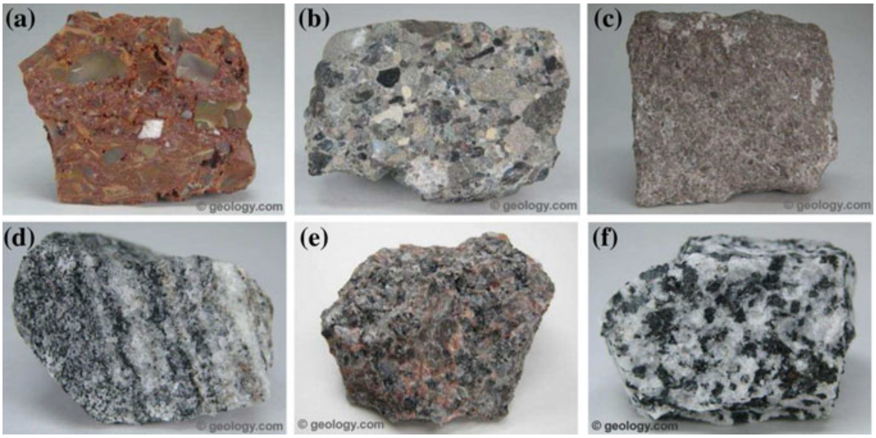


Fig. 1 Grainy structure of different rocks: **a** breccia (sedimentary), **b** conglomerate (sedimentary), **c** limestone (sedimentary), **d** gneiss (metamorphic), **e** granite (igneous), **f** quartz-diorite (igneous). The size of all of the samples is approximately 5 cm. The photographs are taken from http://geology.com/rocks/

of heterogeneous material which is considered as assembly of grains of material held together by cohesive links. This framework corresponds also to the geological formation of rocks, where many different groups of rocks possess a grainy structure which allows the grain recognition even with the bare eye (Fig. 1).

> Rock domain is discretized with the Voronoi cells representing rock grains, while Timoshenko beams act as cohesive links between them (Fig. 2).

Several papers developed discrete lattice models, where the domain is discretized with the Voronoi cells [25, 26].

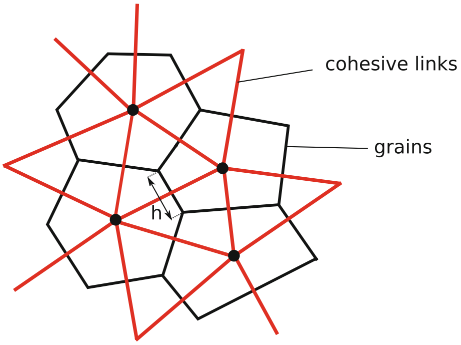


Fig. 2 The basis of the proposed discrete model relies on the lattice of Timoshenko beams which represent the cohesive links keeping the rock grains (Voronoi cells) together

Usually, the discrete lattice models simulate the progressive failure characterized by localization with re-meshing process [27]. Namely, the cohesive links are sequentially removed from the mesh when the discontinuity propagate between the grains. The main difference in the presented model, with respect to latter approach, concerns embedded discontinuities placed within the framework of ED-FEM, where Timoshenko beam elements are equipped with enhanced kinematics capable of capturing the localization effects, like shown in [28–30]. Namely, the embedded discontinuities are placed in the middle of each Timoshenko beam. This corresponds to the Voronoi cell network, where each cohesive link is cut by half by the edge between two neighbouring Voronoi cells.

The embedded discontinuity in the longitudinal local direction of cohesive link (Timoshenko beam) enables the grain dilation due to mode I or tensile failure mode. However, Timoshenko beams also allow to account for pronounced shear effects in both elastic and plastic phase which is used here for representing the failure in mode II (shear sliding along the grains) adding the corresponding displacement or strong discontinuity in the transversal local direction. This leads to localized solutions (i.e. discontinuity propagations) which are enabled like shown in Fig. 3.

Heterogeneities are considered through two different phases representing the initial state; the intact rock material and the initial weaker material that stands for preexisting defects. The macroscopic response of the system is largely influenced by the distribution and position of the phases. The intact rock material is represented by the stronger links, (i.e. Timoshenko beams). Thus, the discontinuity is more likely to propagate through the weaker phase. Failure of the material can occur in both modes separately, as well as in their combination.

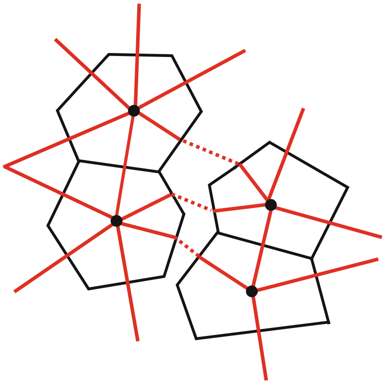


Fig. 3 The strong discontinuity propagation between the Voronoi cells invokes the enhanced kinematics activation

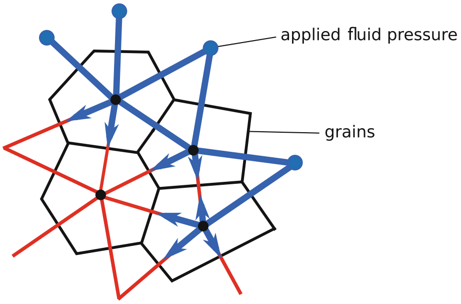


Fig. 4 The fluid flow is dispersed across the lattice network of Timoshenko beams

Fluid flow through the saturated porous domain is governed by a diffusion equation incorporating the Darcy law in terms of continuous pore pressures across the discrete lattice domain (Fig. 4), like shown in [18, 19, 31].

Fluid flow is spread across the lattice of beams, where fluid pressure acts as additional degree of freedom of the beam. The coupled process between the mechanics strain and fluid flow in deformable medium with micro-cracks is governed by Biot's porous media theory [2].

2 Numerical Model Formulations

Rock is considered as porous solid saturated with a fluid. The flow conditions allows that convective terms and gravity acceleration be neglected in this problem. Standard equilibrium equation of saturated two-phase medium is given by relation

$$\nabla \cdot \sigma = 0, \tag{1}$$

where the total stress is

$$\sigma = \sigma_s + \sigma_f = \sigma' - bp \tag{2}$$

and subscripts s and f denote the solid and the fluid part, respectively. The effective stress $\sigma_s = \sigma'$ measures the material properties of the solid skeleton under drained conditions, p is fluid pressure and b is Biot coefficient. Fluid equation is given with

$$\frac{1}{M}\frac{\partial p}{\partial t} + b\nabla \cdot v_s + \nabla \cdot v_f = 0, \tag{3}$$

where vectors v_s and v_f represent the velocities of the solid and the fluid, respectively. The latter is defined by the Darcy law

$$v_f = -k_f \nabla p \tag{4}$$

where k_f is the permeability of the porous medium, M is Biot's modulus defined as

$$\frac{1}{M} = \frac{n_f}{K_f} + \frac{b - n_f}{K_s}, \tag{5}$$

and b is a Biot coefficient defined as

$$b = 1 - \frac{K_t}{K_s}. \tag{6}$$

Here, n_f denotes porosity, K_f is the bulk modulus of the fluid, K_s is a bulk modulus of the solid and K_t is the overall bulk modulus of the porous medium.

The presented model is based on Timoshenko beam elements connecting the grains of material in terms of Voronoi cells. Thus, the weak form of the equilibrium equation (1), in terms of stress resultant (Timoshenko beams) states

$$\int_0^{l_e} \frac{d\mathbf{w}}{dx} \boldsymbol{\sigma} dx = \int_0^{l_e} \mathbf{w}\mathbf{f} dx + \mathbf{w}\mathbf{F}, \tag{7}$$

where $\boldsymbol{\sigma} = [N \quad T \quad M]^T$ represents the stress resultant vector, $\mathbf{f} = [f \quad q \quad m]^T$ is the distributed load vector and $\mathbf{F} = [F \quad Q \quad C]^T$ is the vector of concentrated forces. The right hand side in (7) provides the vector of external forces $\mathbf{F}^{ext}$ with the standard finite element manipulations. The vector $\mathbf{w}$ represents a virtual generalized displacements $V_0 = \{\mathbf{w} : [0, l_e] \mapsto R \ \mid \ [\mathbf{w}]_{\Gamma_u} = 0\}$, which ought to be differentiable and verify $\mathbf{w} \in V_0$.

The constitutive relations for the porous medium (2) are given in terms of total stresses, effective stresses and pore pressures $\boldsymbol{\sigma} = \boldsymbol{\sigma}' - b\mathbf{p}$. The total stress in terms of stress resultants can be decomposed into

$$\begin{bmatrix} N \\ T \\ M \end{bmatrix} = \begin{bmatrix} N' \\ T' \\ M' \end{bmatrix} - b \begin{bmatrix} pA \\ 0 \\ 0 \end{bmatrix}, \tag{8}$$

where the effective stress resultant components can be obtained through the solid's skeleton 'drained' elasticities denoted with $\mathbf{D}_{sk}$

$$\begin{bmatrix} N' \\ T' \\ M' \end{bmatrix} = \underbrace{\begin{bmatrix} EA & 0 & 0 \\ 0 & GA & 0 \\ 0 & 0 & EI \end{bmatrix}}_{\mathbf{D}sk} \begin{bmatrix} \varepsilon \\ \gamma \\ \kappa \end{bmatrix}. \tag{9}$$

Note that E represents Young's modulus, G shear modulus and I moment of inertia of the beam.

The fluid flow equation (3) takes the weak form for the discrete lattice representation of the domain

$$-\int_0^{l_e} \pi M^{-1} \frac{dp}{dt} dx + \int_0^{l_e} \frac{d\pi}{dx} \alpha \mathbf{v}_s dx + \int_0^{l_e} \frac{d\pi}{dx} k_f \frac{dp}{dx} dx = Q^{ext}, \tag{10}$$

where π is the virtual pressure field that obeys the same regularity as the virtual displacement field.

2.1 Enhanced Kinematics

This section provides the enhanced formulation for Timoshenko beam as cohesive link, resulting with embedded discontinuities in local longitudinal direction for mode I failure, and in transversal direction for mode II failure. The localized failure is accompanied by a softening regime in a global macro-response, where the heterogeneous displacement field is used in order to obtain a mesh-independent response. The formulation for fracture process zone with micro-cracks is also presented here through the hardening regime with standard plasticity.

The localization implies heterogeneous displacement field which no longer remains regular, even for smooth stress field. Thus, the displacement field ought to be introduced and written as the sum of a sufficiently smooth, regular part and a discontinuous part. Furthermore, the axial and transversal displacement fields need to be calculated independently.

A cohesive link finite element with two nodes of length l_e and cross section A is considered (Fig. 5). The standard degrees of freedom at each node $i \in [1, 2]$ are axial displacement u_i, transversal displacement v_i and rotation θ_i, accompanied with pressure p_i degree of freedom. The strain measures for standard Timoshenko element are given

$$\begin{aligned} \varepsilon(x) &= \frac{du(x)}{dx} \\ \gamma(x) &= \frac{dv(x)}{dx} - \theta(x) \\ \kappa(x) &= \frac{d\theta(x)}{dx}. \end{aligned} \tag{11}$$

In order to obtain the displacement jumps in the interiors of the cohesive links, the displacement fields need to be enhanced leading to regular and singular parts, where latter can be represented as a product of Heaviside function and displacement jump. The enhanced displacement fields can thus be written as

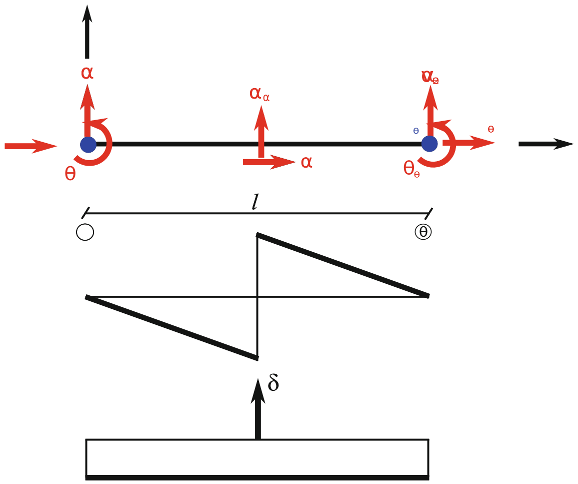


Fig. 5 The enhanced finite element with it's degrees of freedom and discontinuous shape function $M(x)$ and it's derivative $G(x)$

$$
\begin{aligned}
u(x) &= \overline{u}(x) + \alpha_u H_{x_c} \\
v(x) &= \overline{v}(x) + \alpha_v H_{x_c},
\end{aligned}
\tag{12}
$$

where α_u and α_v represent incompatible mode parameters which denote the displacement jumps in axial and transversal direction providing the failure modes I and II. H_{x_c} is the Heaviside function being equal to one if $x > x_c$, and zero otherwise, while x_c is the position of the discontinuity. The presented model assumes the position of discontinuity to be in the middle of the beam. This is the case when each cohesive link is cut in half by the two neighboring Voronoi cells.

The enhanced deformation fields, in terms of regular and singular parts, results from (12) with

$$
\begin{aligned}
\varepsilon(x) &= \overline{\varepsilon}(x) + \alpha^{(u)} \delta_{x_c} \\
\gamma(x) &= \overline{\gamma}(x) + \alpha^{(v)} \delta_{x_c},
\end{aligned}
\tag{13}
$$

where $\overline{\varepsilon}$ and $\overline{\gamma}$ denote regular parts, and Dirac delta δ_{x_c} is the singular part representation of the deformation field. The Dirac delta function δ_{x_c} takes an infinite value at $x = x_c$ and remains equal to zero everywhere else.

For this element, standard linear interpolation functions are used for regular displacement approximation

$$
\mathbf{N} = \left\{ N_1(x) = 1 - \frac{x}{l_e};\;\; N_2(x) = \frac{x}{l_e} \right\}, \tag{14}
$$

along with their derivatives

$$
\mathbf{B} = \left\{ B_1(x) = -\frac{1}{l_e};\;\; B_2(x) = \frac{1}{l_e} \right\}. \tag{15}
$$

Beside standard interpolations, the enhanced interpolation function M is derived in the spirit of ED-FEM (see [6, 22, 23]) and can be used alongside standard interpolation functions to describe the heterogeneous displacement fields with activated discontinuity jump producing embedded discontinuity inside the finite element. The $M(x)$ is defined as

$$M(x) = \begin{cases} -\frac{x}{l_e}; & x \in [0, x_c\rangle \\ 1 - \frac{x}{l_e}; & x \in \langle x_c, l_e] \end{cases}, \tag{16}$$

while $G(x)$ represents the derivative of enhanced function $M(x)$, with respect to local coordinate direction x

$$\begin{aligned} G(x) &= \overline{G} + \delta_{x_c} \\ &= -\frac{1}{l_e} + \delta_{x_c}, \quad x \in [0, l_e]. \end{aligned} \tag{17}$$

Enhanced functions M and G are shown in Fig. 5. This kind of formulation cancels the contribution of incompatible mode parameter on the element boundary leading to possibility of computing the discontinuity parameters locally, while the global equations remain with the nodal displacements as primal unknowns.

Finally, the enhanced finite element displacement interpolations are written in terms of embedded discontinuity

$$\begin{aligned} u(x) &= \sum_{a=1}^{2} N_a(x)u_a + M(x)\alpha_u \\ v(x) &= \sum_{a=1}^{2} N_a(x)v_a + M(x)\alpha_v \\ \theta(x) &= \sum_{a=1}^{2} N_a(x)\theta_a. \end{aligned} \tag{18}$$

The discrete approximation of deformation field can be obtained from the above displacement field (18) resulting with

$$\begin{aligned} \varepsilon(x) &= \sum_{a=1}^{2} B_a(x)u_a + G(x)\alpha_u \\ \gamma(x) &= \sum_{a=1}^{2} (B_a(x)v_a - N_a(x)\theta_a) + G(x)\alpha_v \\ \kappa(x) &= \sum_{a=1}^{2} B_a(x)\theta_a, \end{aligned} \tag{19}$$

The fluid flow is enabled by adding the pressure degree of freedom on top of standard Timoshenko degrees of freedom leading to enhanced element, not only in terms of added pressures, but also in localized discontinuity contributions. The enhanced finite element with all degrees of freedom is shown in Fig. 5.

The pressure field is interpolated with the linear shape functions as well $\{N_1^p(x) = 1 - \frac{x}{l_e},\ \ N_2^p(x) = \frac{x}{l_e}\}$. The corresponding derivatives are $\{B_1^p(x) = -\frac{1}{l_e},\ \ B_2^p(x) = \frac{1}{l_e}\}$. However, the pressure interpolation functions are denoted with the superscript p for clearer presentation. Since the fluid flow problem is transient, the time parameter t is introduced and the discretization field for pressure follows

$$p(x,t) = \sum_{a=1}^{2} N_a^p(x) p_a(t). \tag{20}$$

The discretization of the pressure gradient is

$$\frac{\partial p}{\partial x}(x,t) = \sum_{a=1}^{2} B_a^p(x) p_a(t), \tag{21}$$

while its time derivative

$$\frac{\partial p}{\partial t}(x,t) = \sum_{a=1}^{2} N_a^p(x) \dot{p}_a(t). \tag{22}$$

The generalized nodal pressure field can be denoted with $\mathbf{p} = (p_1, p_2)^T$.

2.2 The Enhanced Weak Form

The generalized virtual deformations are interpolated in the same way as the real ones

$$\delta\boldsymbol{\varepsilon} = \mathbf{B}\delta\mathbf{d} + G\delta\boldsymbol{\alpha}, \tag{23}$$

with δ standing for prefix indicating the corresponding virtual field or variation. Such interpolated fields produce the internal force vector and the finite element residual vector due to discontinuity

$$\begin{aligned} \mathbf{F}^{int} &= \int_0^{l_e} \mathbf{B}^T \boldsymbol{\sigma} dx, \\ \mathbf{h}^{(e)} &= \int_0^{l_e} (\overline{G} + \delta_{x_c}) \boldsymbol{\sigma} dx. \end{aligned} \tag{24}$$

From the condition of residual equation being equal to zero, the internal forces at the discontinuity ought to be calculated

$$\begin{aligned}\mathbf{h}^{(e)} &= \int_0^{l_e} (\overline{G} + \delta_{x_c})\boldsymbol{\sigma}\, dx \\ &= \int_0^{l_e} \overline{G}\boldsymbol{\sigma}\, dx + \mathbf{t}.\end{aligned} \tag{25}$$

Vector **t** represents the internal forces at discontinuity, which are in relation with the forces from the bulk

$$\mathbf{t} = -\int_0^{l_e} \overline{G}\boldsymbol{\sigma}\, dx, \quad \mathbf{t} = (t_u, t_v, 0)^T \tag{26}$$

2.3 Constitutive Model

It has been observed that representative behaviour of rock material, including the post-peak behaviour, can be separated into five different stages based upon stress-strain characteristics. These stages can be defined as: crack closure, linear elastic deformation, crack initiation and stable crack growth, critical energy release and unstable crack growth, failure and post-peak behaviour. Figure 6 shows typical stress-strain curve of the brittle rock under the compression test and its failure stages.

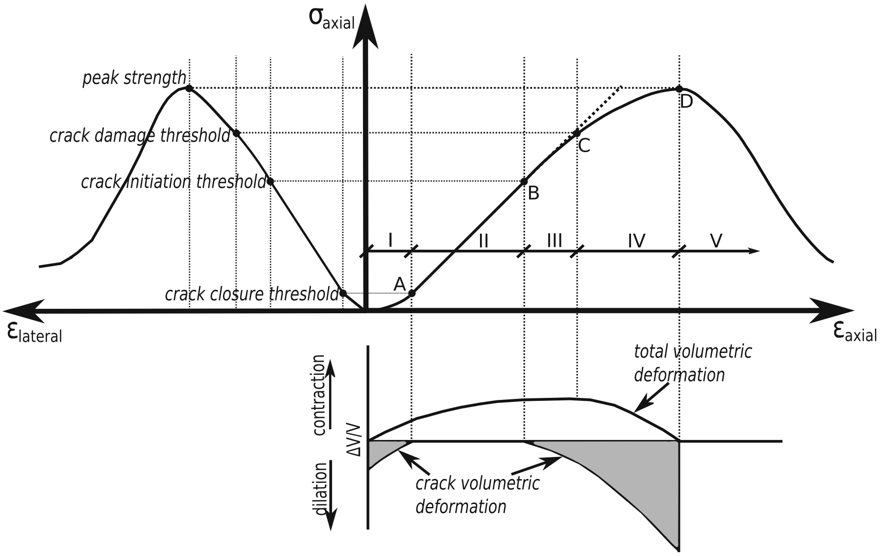


Fig. 6 Stress-strain curve showing the elements of crack development

Stage I is associated to microcrack closure and the initial flaws in the material which continues with stage II, a linear elastic stage. The inelastic behaviour starts at the beginning of stage III and until the end of stage, the hardening response accompanied by fracture process zone with microcrack initiation, can be observed. With an increase of a loading program, stage IV is activated. The stress value at the beginning of this stage (point C) can vary between 50–90% of ultimate strength, while the rest of the stage is characterized by the nonlinear behaviour and more rapid increase of lateral deformation. At the point D, the ultimate strength of specimen is reached and the larger macro-cracks start to propagate through the sample leading to softening of the specimen. At this point, the volumetric strain starts to reverse from a compressive to dilatation behaviour.

The constitutive relations need to be defined outside and at the discontinuity. The constitutive models are constructed within the framework of thermodynamics for a stress resultant beam formulation.

The beam longitudinal and transversal directions are enhanced with additional kinematics, representing modes I and II with softening behaviour, while the rotations keep their standard elastic form. The first two stages of rock failure (up to point B) are kept elastic, with respect to stage I being finished soon after the loading is applied. The linear elastic behaviour is finished when the point B is reached, continuing with hardening. When stage III is activated, significant damage caused by micro-crack propagation starts to occur in the specimen and increases until the highest peak point (point D). The constitutive model for latter stages, which represents a fracture process zone, is chosen as classical plasticity model with isotropic hardening. When the critical point is reached, the complete failure of the specimen is enabled through the exponential softening law. This invokes the enhanced kinematics activation and occurrence of the displacement jumps. The carrying capacity of element reduces with increase in the displacement jump.

In the following equations, the development for the failure of the beam in modes I and II is presented. When the loading starts and softening has not formed yet, the classical elasto-plastic model is considered. The total strains can be additively decomposed into elastic and plastic components

$$\begin{aligned} \varepsilon &= \bar{\varepsilon}^e + \bar{\varepsilon}^p \\ \gamma &= \bar{\gamma}^e + \bar{\gamma}^p. \end{aligned} \tag{27}$$

Strain energy functions depend upon elastic strains and hardening variables, $\bar{\xi}_u$, $\bar{\xi}_v$:

$$\begin{aligned} \psi_u\left(\varepsilon, \bar{\varepsilon}^p, \bar{\xi}_u\right) &= \frac{1}{2} EA \left(\varepsilon - \bar{\varepsilon}^p\right)^2 + \frac{1}{2} \bar{\xi}_u^2 \bar{K}_u \\ \psi_v\left(\gamma, \bar{\gamma}^p, \bar{\xi}_v\right) &= \frac{1}{2} GA \left(\gamma - \bar{\gamma}^p\right)^2 + \frac{1}{2} \bar{\xi}_v^2 \bar{K}_v, \end{aligned} \tag{28}$$

where $\overline{K}_u$ and $\overline{K}_v$ denote isotropic hardening modulus for longitudinal and transversal direction. The yield criterion is defined as

$$\begin{aligned} \overline{\Phi}_u\left(N, \overline{q}_u\right) &= N - \left(N_y - \overline{q}_u\right) \leq 0 \\ \overline{\Phi}_v\left(T, \overline{q}_v\right) &= |T| - \left(T_y - \overline{q}_v\right) \leq 0, \end{aligned} \tag{29}$$

where N_y and T_y represent the forces at yielding point. The state equations are

$$\begin{aligned} N &= EA\left(\varepsilon - \overline{\varepsilon}^p\right) \\ T &= GA(\gamma - \overline{\gamma}^p). \end{aligned} \tag{30}$$

and

$$\begin{aligned} \overline{q}_u &= -\overline{K}_u \overline{\xi}_u \\ \overline{q}_v &= -\overline{K}_v \overline{\xi}_v. \end{aligned} \tag{31}$$

For the inelastic case, the principle of maximum dissipation is considered, the evolution laws are obtained as

$$\begin{aligned} \dot{\overline{\varepsilon}}^p &= \dot{\overline{\lambda}}_u \frac{\partial \overline{\Phi}_u}{\partial N} = \dot{\overline{\lambda}}_u sign(N); \quad \dot{\overline{\xi}}_u = \dot{\overline{\lambda}}_u \frac{\partial \overline{\Phi}_u}{\partial \overline{q}_u} = \dot{\overline{\lambda}}_u \\ \dot{\overline{\gamma}}^p &= \dot{\overline{\lambda}}_v \frac{\partial \overline{\Phi}_v}{\partial T} = \dot{\overline{\lambda}}_v sign(T); \quad \dot{\overline{\xi}}_v = \dot{\overline{\lambda}}_v \frac{\partial \overline{\Phi}_v}{\partial \overline{q}_v} = \dot{\overline{\lambda}}_v, \end{aligned} \tag{32}$$

where the plastic multiplier parameters $\overline{\lambda}_u$ and $\overline{\lambda}_v$ have been introduced to participate in evolution equations obtained from Kuhn-Tucker optimality conditions [6]. The constitutive equations for the elastoplastic case are

$$\dot{N} = \begin{cases} EA\dot{\varepsilon}; & \dot{\overline{\lambda}}_u = 0 \\ \frac{EA\overline{K}_u}{EA+\overline{K}_u}\dot{\varepsilon}; & \dot{\overline{\lambda}}_u > 0 \end{cases}, \quad \dot{T} = \begin{cases} GA\dot{\gamma}; & \dot{\overline{\lambda}}_v = 0 \\ \frac{GA\overline{K}_v}{GA+\overline{K}_v}\dot{\gamma}; & \dot{\overline{\lambda}}_v > 0. \end{cases} \tag{33}$$

Accompanying loading/unloading conditions and consistency condition obey $\dot{\lambda}\Phi = 0, \dot{\lambda} \geq 0, \Phi \leq 0, \dot{\lambda}\dot{\Phi} = 0$.

Once the ultimate failure point is reached, enhanced kinematics needs to be activated. All further plastic deformation will be accumulated at the discontinuity section, that once passed the peak resistance. The corresponding strain fields containing regular and singular components are obtained:

$$\begin{aligned} \varepsilon &= \overline{\varepsilon} + \overline{\overline{\varepsilon}} = \overline{\varepsilon}^e + \overline{\varepsilon}^p + \overline{\overline{\varepsilon}} \\ \gamma &= \overline{\gamma} + \overline{\overline{\gamma}} = \overline{\gamma}^e + \overline{\gamma}^p + \overline{\overline{\gamma}}. \end{aligned} \tag{34}$$

The failure criteria for mode I and mode II failure are defined as

$$\begin{aligned} \overline{\overline{\Phi}}_u\left(t_u, \overline{\overline{q}}_u\right) &= t_u - \left(N_u - \overline{\overline{q}}_u\right) \leq 0 \\ \overline{\overline{\Phi}}_v\left(t_v, \overline{\overline{q}}_v\right) &= |t_v| - \left(T_u - \overline{\overline{q}}_v\right) \leq 0, \end{aligned} \tag{35}$$

where N_u, T_u are the ultimate capacity forces and $\overline{\overline{q}}_u$, $\overline{\overline{q}}_v$ are stress-like softening variables which increase exponentially as

$$\begin{aligned}\overline{\overline{q}}_u &= N_u \left(1 - exp\left(-\overline{\overline{\xi}}_u \frac{N_u}{G_{f,u}}\right)\right)\\ \overline{\overline{q}}_v &= T_u \left(1 - exp\left(-\overline{\overline{\xi}}_v \frac{T_u}{G_{f,v}}\right)\right),\end{aligned} \tag{36}$$

and t_u, t_v are traction forces at the discontinuity obtained from equilibrium equations (26). The evolution of internal variables in softening states

$$\begin{aligned}\dot{\alpha}_u &= \dot{\overline{\overline{\lambda}}}_u \frac{\partial \overline{\overline{\Phi}}_u}{\partial N} = \dot{\overline{\overline{\lambda}}}_u sign(N); \;\; \dot{\overline{\overline{\xi}}}_u = \dot{\overline{\overline{\lambda}}}_u \frac{\partial \overline{\overline{\Phi}}_u}{\partial \overline{\overline{q}}_u} = \dot{\overline{\overline{\lambda}}}_u\\ \dot{\alpha}_v &= \dot{\overline{\overline{\lambda}}}_v \frac{\partial \overline{\overline{\Phi}}_v}{\partial T} = \dot{\overline{\overline{\lambda}}}_v sign(T); \;\; \dot{\overline{\overline{\xi}}}_v = \dot{\overline{\overline{\lambda}}}_v \frac{\partial \overline{\overline{\Phi}}_v}{\partial \overline{\overline{q}}_v} = \dot{\overline{\overline{\lambda}}}_v,\end{aligned} \tag{37}$$

where $\overline{\overline{\lambda}}$ is the plastic multiplier associated with the softening behaviour and α is an equivalent to the accumulated plastic strain at the discontinuity.

2.4 The Finite Element Equations of a Coupled Poroplastic Problem

In this section, the final finite element implementation aspects accounting for each single element contribution, further denoted with subscript e, are presented.

The regular part of weak form (24/1) leads to the element residual equation

$$\mathbf{r}_d = \mathbf{F}^{ext} - \mathbf{A}_{e=1}^{n_{el}} \int_0^{l_e} \mathbf{B}^{d,T} \boldsymbol{\sigma} dx, \tag{38}$$

where the total stress resultants $\boldsymbol{\sigma}$ are obtained in terms of effective stress resultants $\boldsymbol{\sigma}'$ and pore pressures $\mathbf{p}$ in (8). The symbol $\mathbf{A}_{e=1}^{n_{el}}$ denotes the finite element assembly operator for all element contributions. The effective stress resultants $\boldsymbol{\sigma}'$ are calculated in terms of regular parts of enhanced strain field (23). The enhanced strain parameters $\boldsymbol{\alpha}$, in each element where localization occurs, are obtained by solving the local equilibrium of the effective stresses

$$\mathbf{h}^{(e)} = \int_0^{l_e} \overline{\mathbf{G}} \boldsymbol{\sigma}' dx + \mathbf{t}', \tag{39}$$

where $\mathbf{t}'$ represent the corresponding effective tractions acting at the discontinuity. The local equilibrium equation in (39) offers the benefit of local computation of the enhanced parameters. Subsequent static condensation of these parameters allows to keep standard matrix at the global level. The local computation algorithm and numerical procedure are described in the next subsection.

Upon introducing the finite element interpolations, the coupled fluid equation (10) results with the finite element residual form

$$\mathbf{r}_p = \mathbf{Q}^{ext} - \mathbf{A}_{e=1}^{n_{el}}\Bigg[\int_0^{l_e} \mathbf{N}^{p,T} M^{-1}\mathbf{N}^p dx\dot{\mathbf{p}}_e - \\ - \int_0^{l_e} \mathbf{N}^{p,T}\alpha\mathbf{B}^d dx\dot{\mathbf{d}}_e - \int_0^{l_e} \mathbf{B}^{p,T} k_f \mathbf{B}^p dx\mathbf{p}_e\Bigg], \tag{40}$$

where $\mathbf{Q}^{ext}$ represent the external applied fluxes and imposed pressures. The consistent linearization of the Eqs. (38) and (40) leads to a set of linear algebraic equations

$$\mathbf{r}_d^{(i)} - \mathbf{A}_{e=1}^{n_{el}}\Bigg[\mathbf{K}_e\Delta\mathbf{d}_e - \mathbf{L}_e\Delta\boldsymbol{\alpha}_e - \mathbf{Q}_e\Delta\mathbf{p}_e\Bigg] = 0 \tag{41}$$

and

$$\mathbf{r}_p^{(i)} - \mathbf{A}_{e=1}^{n_{el}}\left[\frac{1}{\Delta t}\mathbf{Q}_e^T\Delta\mathbf{d}_e + \left(\mathbf{H}_e + \frac{1}{\Delta t}\mathbf{S}_e\right)\Delta\mathbf{p}_e\right] = 0 \tag{42}$$

in the increments $\Delta t = t_{n+1}^{(i+1)} - t_{n+1}^{(i)}$, where (i) denotes iteration counter within the time interval $[t_n, t_{n+1}]$. The matrices are evaluated in the previous iteration (i) where all values are known. The element stiffness matrix $\mathbf{K}_e$ is defined as

$$\mathbf{K}_e = \int_0^{l_e} \mathbf{B}^{d,T}\mathbf{D}_{sk}\mathbf{B}^d dx \tag{43}$$

and the localized contribution matrix

$$\mathbf{L}_e = \int_0^{l_e} \mathbf{B}^{d,T}\mathbf{D}_{sk}\overline{\mathbf{G}}dx. \tag{44}$$

The compressibility matrix $\mathbf{S}_e$, the permeability matrix $\mathbf{H}_e$ and the coupling matrix $\mathbf{Q}_e$ are given by

$$\mathbf{S}_e = \int_0^{l_e} \mathbf{N}^{p,T} M^{-1}\mathbf{N}^p dx, \tag{45}$$

$$\mathbf{H}_e = \int_0^{l_e} \mathbf{B}^{p,T} k_f \mathbf{B}^p dx, \tag{46}$$

$$\mathbf{Q}_e = \int_0^{l_e} \mathbf{B}^{d,T} b\mathbf{N}^p dx. \tag{47}$$

The linearization of local equilibrium equation in (39) results with

$$\mathbf{h}_e^{(i)} - \mathbf{L}_e^T \varDelta \mathbf{d}_e - \mathbf{F}_e \varDelta \boldsymbol{\alpha}_e = 0, \tag{48}$$

where

$$\mathbf{F}_e = \int_0^{l_e} \overline{\mathbf{G}}^T \mathbf{D}_{sk} \overline{\mathbf{G}} + \mathbf{K}_{dis}. \tag{49}$$

Matrix $\mathbf{K}_{dis}$ contains consistent tangent stiffness components for the discontinuity obtained as a derivatives of the exponential softening laws from (36) with respect to the corresponding displacement jumps.

The enhanced strain parameters $\varDelta \boldsymbol{\alpha}$ can be obtained by the local operator split solution procedure and return mapping algorithm presented in the next section. Finally, the static condensation strategy serves for local elimination of the enhanced strain parameters which leads to the final statically condensed equation

$$\mathbf{r}_d^{(i)} - \mathbf{A}_{e=1}^{n_{el}} \left[\left(\mathbf{K}_e - \mathbf{L}_e^T \mathbf{F}_e^{-1} \mathbf{L}_e \right) \varDelta \mathbf{d}_e - \mathbf{Q}_e \varDelta \mathbf{p}_e \right] = 0. \tag{50}$$

2.5 The Operator Split Algorithm

The operator split is an element-wise algorithm performed for each directional component with its ultimate goal of computing the internal variables related to discontinuity. After computing the internal variables locally, the global solution procedure with Newton incremental/iterative procedure can be performed.

It is assumed that the best iterative value of displacements $u_{n+1}^{(i)}$ and $v_{n+1}^{(i)}$ for which the trial values of the traction forces are obtained

$$t_{*,n+1}^{trial} = - \int_0^{l_e} \overline{G} \left[EA \left(\sum_{a=1}^{2} B_a^d u_{a,n+1}^{(i)} + \overline{G} \alpha_{*,n} \right) \right] \tag{51}$$

where $\alpha_{*,n}$ represents the discontinuity parameters at previous time for softening plastic deformation. The * denotes each directional component of the Timoshenko beam. Later on, the trial value of failure functions ought to be calculated

$$\overline{\overline{\varPhi}}_{*,n+1}^{trial} = t_{*,n+1}^{trial} - \left(N_u - \overline{\overline{q}}_{*,n} \right). \tag{52}$$

If the trial values of the failure functions are negative or zero, the elastic trial step is accepted for final, with no modification of the plastic strain from the previous time step

$$\alpha_{*,n+1} = \alpha_{*,n}; \quad \overline{\overline{\xi}}_{*,n+1} = \overline{\overline{\xi}}_{*,n}, \tag{53}$$

The plastic softening parameter will remain intact, while the traction force will be changed due to displacement increment.

On the other hand, if the trial values of failure functions are positive, the current step is in the softening plasticity and there is a need to modify the elastic strain and internal variables $\alpha_{*,n}$, in order to re-establish the plastic admissibility at discontinuity. The internal softening plasticity variables ought to be updated by using evolution equations

$$\alpha_{*,n+1} = \alpha_{*,n} + \bar{\bar{\lambda}}_{*,n+1} sign\left(t^{trial}_{*,n+1}\right) \tag{54}$$

and

$$\bar{\bar{\xi}}_{*,n+1} = \bar{\bar{\xi}}_{*,n} + \bar{\bar{\lambda}}_{*,n+1} \tag{55}$$

where $\bar{\bar{\lambda}}_{*,n+1}$, $\bar{\bar{\lambda}}_{*,n+1}$ are softening plastic multipliers. The value of the plastic multiplier is determined from the conditions $\bar{\bar{\varPhi}}_{*,n+1} \leq tol$ and the solutions of a nonlinear equations are obtained iteratively using the Newton-Raphson method

$$\bar{\bar{\varPhi}}_{*,n+1} = \bar{\bar{\varPhi}}^{trial}_{*,n+1} + \left(\bar{\bar{q}}_{*,n+1} - \bar{\bar{q}}_{*,n}\right) + EA\overline{G}\bar{\bar{\lambda}}_{*,n+1} \leq tol \tag{56}$$

In the plastic softening step, the traction forces are produced by a change of discontinuity parameters α_*.

3 Numerical Simulations

In this section, the numerical simulations for several numerical tests are presented. The uniaxial tension and compression tests are performed on heterogeneous 2D rock specimens. The influence of heterogeneity with different distributions of phase I and II (strong and weak phase) are studied. Fluid-saturated rock sample with localized shear band formation development is presented as well. Presented numerical model formulations are implemented into the research version of the computer code FEAP [32].

3.1 Preparation of 2D Plain Strain Rock Specimens

2D plane strain rock specimens are constructed. The specimens are of dimensions 10 × 10 cm (with unit thickness) and are meshed with triangles by means of Delaunay algorithm. The specimen has 253 nodes and 704 elements (Fig. 7). Timoshenko beam elements are positioned on each edge of every triangle in the specimen. Their geometric properties represent the corresponding part in specimen volume. The main hypothesis in constructing the lattice model is that the cells connected by cohesive

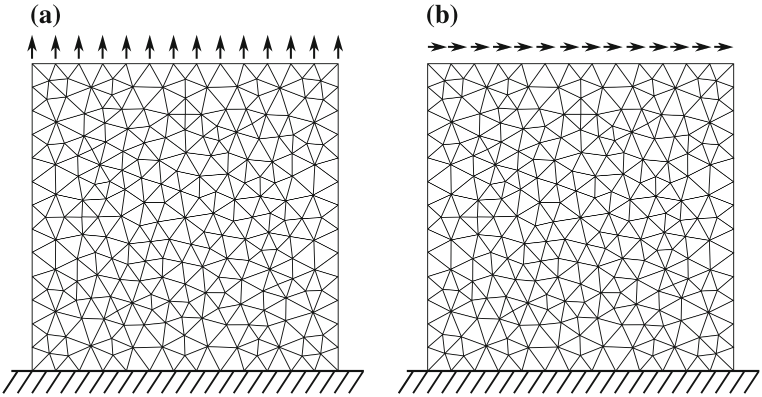


Fig. 7 A homogeneous 2D plain strain specimen is constructed. Uniaxial tension (**a**) and shear test (**b**) are performed in linear elastic regime to validate the model

links (beams) correspond to the representative part of the specimen which have homogeneous properties, while the heterogeneities are introduced through the cohesive links. Thus, the Voronoi cells are derived from Delaunay triangulation and the beam cross sections are computed from the length of the common size of the neighbouring cells (Fig. 8). The material parameters are taken the same as in the equivalent standard continuum.

In order to validate the lattice model parameters, the tension and shear tests are conducted in the linear elastic regime on the proposed homogeneous specimen (shown in Fig. 7) in two versions: lattice model and equivalent standard continuum model with triangular solid elements. The material parameters are the same for each test version: $E = 1000\,\mathrm{kN/cm}^2$, $\nu = 0.2$. The results are presented in Fig. 9a, b.

The equivalent standard continuum model (with triangles as finite elements) operate only in linear elastic regime and its response matches with linear elastic regime of lattice models before the failure phase, showing that the proposed model is capable of reproducing classical linear elastic continuum with such computed lattice parameters.

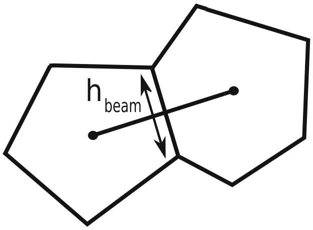


Fig. 8 Beam cross sections are computed from the length of the common size of the neighbouring cells

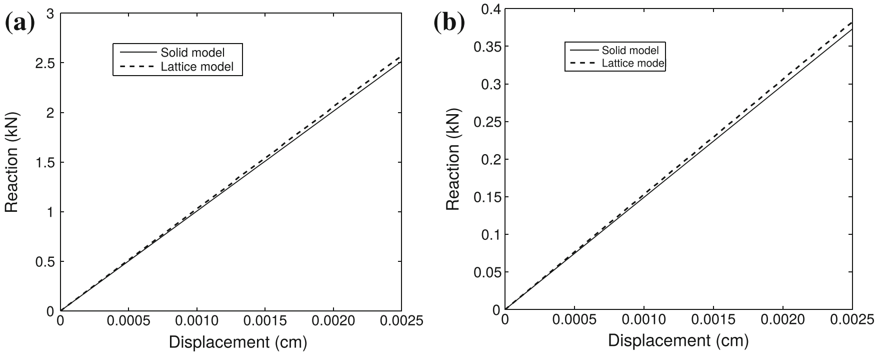


Fig. 9 Response of homogeneous specimen in linear elastic regime for **a** tension test and **b** shear test in two versions: solid model with triangles and lattice model

3.2 Influence of Heterogeneity in Tension and Compression Tests

In this example, the influence of heterogeneity on a global response is studied. Three different specimens with the same geometric properties (same specimen size), but different levels of heterogeneity are subjected to uniaxial tension and compression tests. Table 1 summarizes the mechanical and geometric characteristics of the specimen used for these experiments. The corresponding macroscopic results are shown in Fig. 10a, b.

The specimens are given different initial properties, specifically with 40, 50 and 60 % of phase II material. With an increase of phase II material, the global modulus of elasticity decreases. This is the result of more elements of phase II representing initial weaker material, which makes the global response of specimen more ductile and also with a somewhat lower value of modulus of elasticity. However, it can also be seen from global exponential curve that, when a ratio of phase II material increases, the failure of the specimen becomes more ductile in fracture process zone creation, but also more brittle in the softening response phase, for when the fracture

Table 1 Mechanical and geometric characteristics of the specimen

Phase I	Phase II
$E = 7000\,\text{kN/cm}^2$	$E = 1000\,\text{kN/cm}^2$
	$\nu = 0.2$
$\nu = 0.2$	$\sigma_u = 2.2\,\text{MPa}$, $\tau_u = 1.15\,\text{MPa}$
Tension fr. energ.:	$G_f^{(u)} = 10\,\text{N/m}$, $G_f^{(v)} = 1.5\,\text{N/m}$
Comp. fr. energ.:	$G_f^{(u)} = 350\,\text{N/m}$, $G_f^{(v)} = 10\,\text{N/m}$
Dimensions: $0.1 \times 0.1 \times 0.01$ m; 40, 50, 60 % phase II	

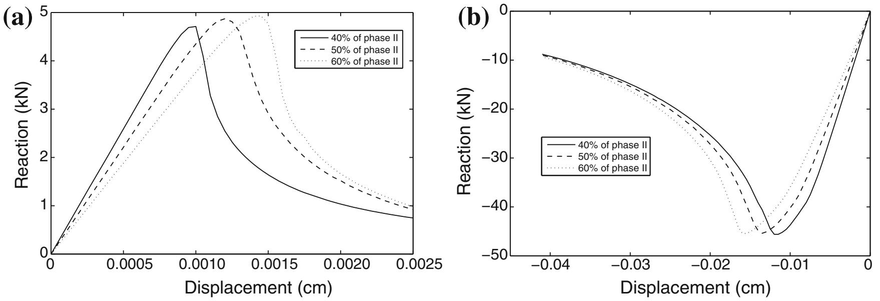


Fig. 10 The computed macroscopic response with different levels of heterogeneity for: **a** uniaxial tension test and **b** uniaxial compression test

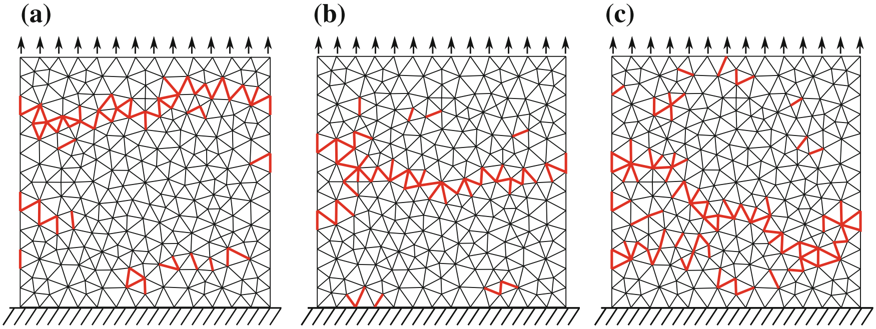


Fig. 11 Final failure patterns created in tension test for specimens with: **a** 40 % of phase II, **b** 50 % of phase II and **c** 60 % of phase II (broken links are *red coloured*) (Color Online)

starts the complete failure happens faster. This is due to appearance of many more potential macro-cracks, which drives more quickly the stress to zero.

The failure patterns of three different heterogeneous specimens are shown in Figs. 11 and 12. Figure 11 presents the final macro-cracks at the end of tension test computations for the specimens with 40, 50 and 60 % of phase II material. It is observed that one dominant macro-crack is present in all of the specimens inducing the final failure mechanism. However, in each specimen the macro-crack formed differently depending on the initial heterogeneity which decides the crack path. Failure due to mode I is more pronounced in tension test.

The ultimate shear strength is defined by the Mohr-Coulomb failure criterion

$$\tau_f = \tau_u + \sigma_c \cdot \tan(\phi), \tag{57}$$

where τ_u represents cohesion-like value of ultimate shear force when compression force is equal to zero, σ_c represents the compression force and ϕ is internal angle of friction. Figure 12 reveals the final cracks formed at the end of compression tests where not only one macro-crack is enough to break the specimens. Contrary to

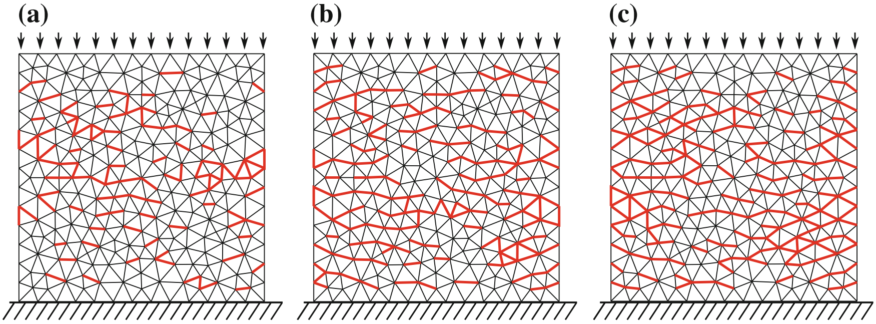


Fig. 12 Final failure patterns created in compression test for specimens with: **a** 40 % of phase II, **b** 50 % of phase II and **c** 60 % of phase II (broken links are *red coloured*) (Color Online)

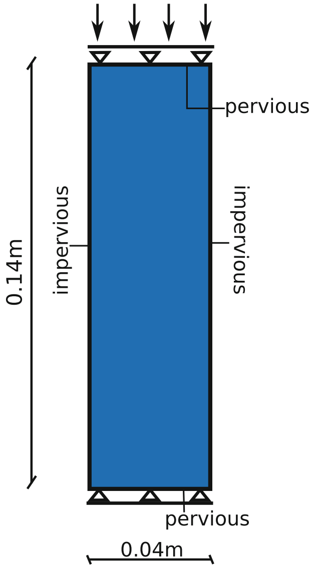
tension test crack patterns, in compression test much more macro-cracks are needed to drive the specimens to failure and these are influenced more by mode II mechanism, compared to tension test, which forms the final crack patterns together with mode I. It is important to note that red coloured links in Figs. 11 and 12 represent the failed cohesive links. However, the actual cracks are localized inside elements and enable the crack propagation between the Voronoi cells, which are dual to Dealunay triangulation.

In either tension or compression, the difference in reduction of the peak stresses in different specimens remains fairly mild. Having approximately the same peak resistance is quite realistic to expect for the similar failure pattern is created once the threshold is reached. However, the similar peak stresses in compression test leads to conclusion that despite the variations in heterogeneity, crack propagation patterns in each of the samples remain similar with similar failure mechanism present in all of them, which can be observed in Fig. 12. Specifically, this means that more defects were present in the specimens with more phase II material which made the material softer, but at the same time these were not crucial for complete failure which was caused by similar macro-cracks in all specimens. This leads to conclusion that difference in heterogeneity, that was used here: 40, 50, 60 % of phase II, is not as significant to lead to drastically different values of ultimate stresses.

3.3 Drained Compression Test of the Poro-plastic Sample with the Localized Failure

The fluid saturated rock sample under compression test is considered in this section. The geometry of the sample and boundary conditions imposed on the displacement and pore pressure fields are shown in Fig. 13. The external load is applied via constant velocity $v_0 = 5 \times 10^{-4}$ m/s imposed on the top base. With the aim of observing the coupling effects as well, the tests are then repeated with the imposed constant velocity

Fig. 13 Geometry of the poroplastic sample and imposed boundary conditions



$v_0 = 1.5 \times 10^{-3}$ m/s. The chosen material parameters listed in Table 2 correspond to the limestone fully saturated with the water. The value of hydraulic permeability of the sample obtained from the parameters in the Table 2 is equal to $K_h = \rho_w g K_f = 1 \times 10^{-8}$ m/s, where the procedure of computing lattice permeabilities is used. Such procedure is performed to find equivalent permeabilities when the fluid flows across

Table 2 Material parameters considered in the numerical simulations of poro-plastic sample

Drained Young modulus	$E_{sk} = 50\,\text{GPa}$
Drained Poisson ratio	$\nu_{sk} = 0.25$
Tensile yield stress	$\sigma_{y,t} = 12\,\text{MPa}$
Shear yield stress	$\tau_y = 23\,\text{MPa}$
Hardening modulus	$\overline{K} = 5\,\text{GPa}$
Tensile strength	$\sigma_{u,t} = 13\,\text{MPa}$
Shear strength	$\tau_u = 25\,\text{MPa}$
Angle of friction	$\phi = 35°$
Fracture energies	$G_{f,u} = 300\,\text{N/m}$; $G_{f,v} = 600\,\text{N/m}$
Biot coefficient	$b = 0.8$
Biot modulus	$M = 16.9\,\text{GPa}$
Porosity	$n_f = 0.1$
Permeability	$K_f = 1 \times 10^{-9}\,\text{m}^2/(\text{kPa/s})$
Fluid density	$\rho_w = 1000\,\text{kg/m}^3$

the discrete lattice network. Associating K_f with given permeability and k_f with lattice permeability, the following expression is obtained

$$k_f = \frac{K_f}{c}, \tag{58}$$

with $c = h_f / l_e$ being the coefficient of modification of permeability for given lattice. Here, h_f denotes the shortest distance between the two centroids of neighbouring triangles and l_e is the length of given element. See [31] for more elaborate explanation of this procedure.

The final goal is to investigate the influence of heterogeneity upon the localized failure of the proposed sample. The presented discrete model formulation is capable of considering the influence of heterogeneity. Here, the two-phase representation is adopted, where the second phase takes the slightly weaker properties in terms of material strengths ($\sigma_{u,t} = 12\,\text{MPa}$; $\tau_u = 24\,\text{MPa}$). The two-phases are distributed randomly throughout the sample and each phase participates with equal number of elements. The differences in two samples are brought by the different distributions of the phases when the random sampling is performed two times in a row. Figures 14 and 15 show the displacements and pore pressures of the heterogeneous samples 1 and 2 plotted in the deformed mesh at the final time step of the simulation. These results are obtained with the imposed constant velocity of $v_0 = 5 \times 10^{-4}$ m/s. It can be observed from the deformed meshes of both samples that the localized macro cracks propagate differently in two cases only because of the slight difference in initial heterogeneity distributions. Macro-cracks also formed the irregular geometries that propagated through the weaker parts of the material. The main strength of the presented discrete model is in simulating the heterogeneous materials where macro-cracks propagate through the material's weaker phases, avoid the stiffer ones and

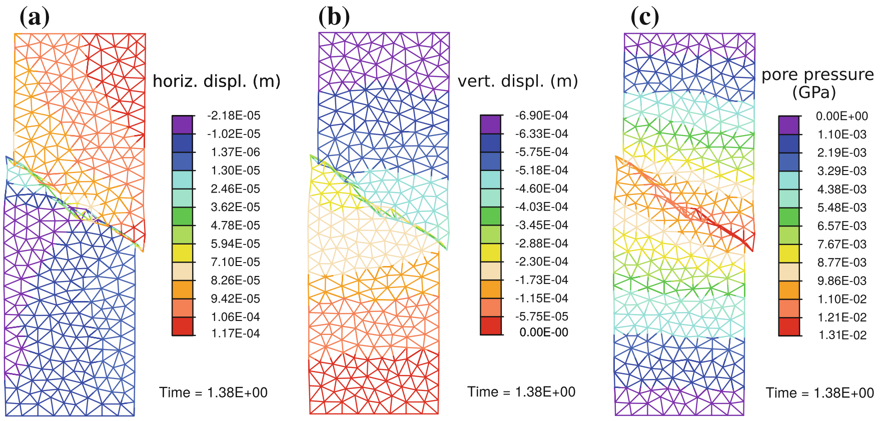


Fig. 14 The state of the 1st heterogeneous sample after the compression test (imposed velocity $v_0 = 5 \times 10^{-4}$ m/s): **a** horizontal displacement **b** vertical displacement **c** pore pressure

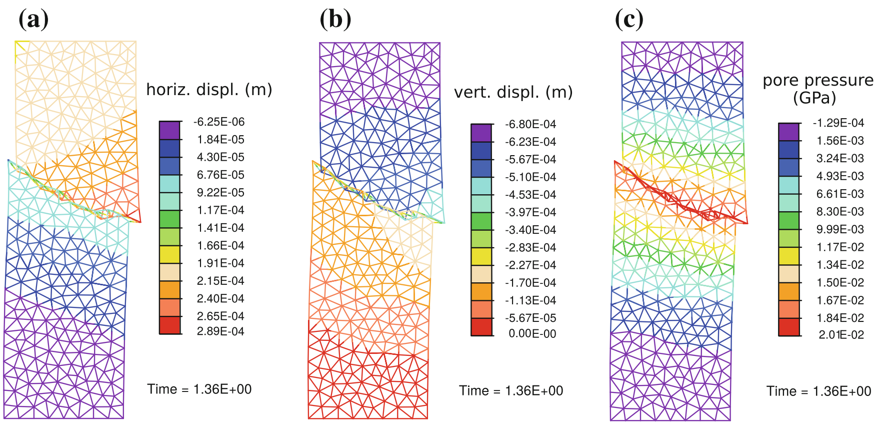


Fig. 15 The state of the 2nd heterogeneous sample after the compression test (imposed velocity $v_0 = 5 \times 10^{-4}$ m/s): **a** horizontal displacement **b** vertical displacement **c** pore pressure

exhibit the irregular geometries. When it comes to the pore pressures, previoussides and reach their highest values near the localized zone.

To investigate the coupling effects, the two heterogeneous samples are put under compression test with a different rate of imposed vertical displacement on the top base $v_0 = 1.5 \times 10^{-3}$ m/s. The macroscopic curves including the cumulative vertical reaction and pore pressure in the centre of the sample in the close neighboured of the localized zone are presented in Fig. 16, for two heterogeneous samples and different imposed velocities obtained within the compression tests.

The macroscopic vertical reactions indicate that higher rates of imposed displacement cause the samples to be more resistant (larger ultimate stress) and more ductile

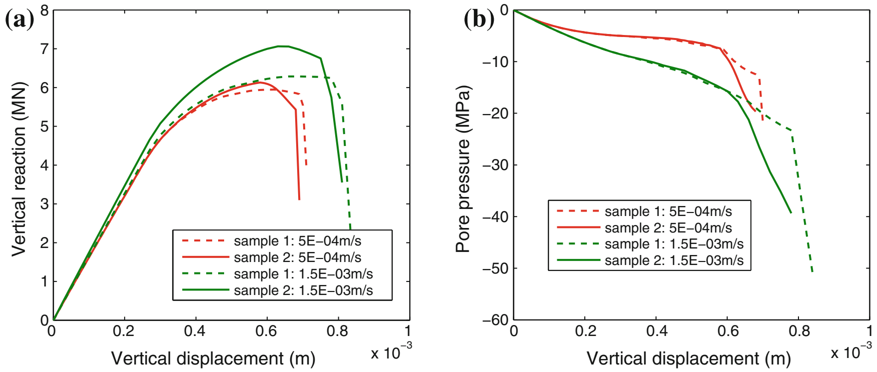


Fig. 16 Macroscopic curves of the poro-plastic sample obtained within the compression test **a** cumulative vertical reaction versus impose displacement **b** pore pressure at the sample centre versus imposed displacement

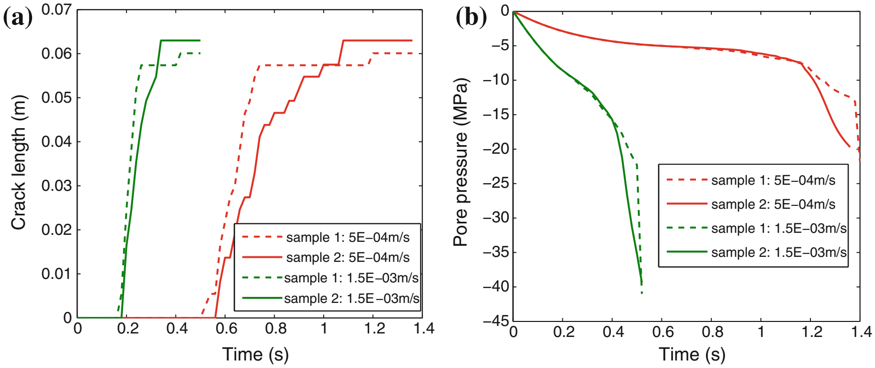


Fig. 17 **a** Crack length versus time **b** pore pressure at the sample centre versus time

(larger displacement is needed to drive the samples to the failure). This is due to an increase of pore pressure which is brought by shorter time left for drainage at the sample centre (Fig. 16b).

The pronounced coupling effects are more obvious when it comes to the nonlinear behaviour and formation of localization zone. In the beginning of the test, the vertical reaction is less influenced by higher pore pressure.

No coupling effect is observed in the geometry of the macro-crack for each sample when it comes to the localized zone formation. More precisely, the discontinuity still propagated through the same elements for different imposed velocities.

The differences with respect to heterogeneities seem to increase in the nonlinear zone with the higher imposed velocity. Namely, the increase of flow through cracks in localization zone, together with the ‘faster’ loading, induces the higher rates of pore pressures making the heterogeneities’ influence even more profound.

As can be seen from Fig. 17a, where time history of the crack length is presented, cracks start to propagate at some point in time when the external load produces significant stress triggering the crack. The cracks then propagate quickly through the samples. The plots for samples with applied faster external load ($v_0 = 1.5 \times 10^{-3}$ m/s) show, that in these cases, cracks propagate more quickly and the tests are completed in less time. Figure 17b presents the time evolution of pore pressure in the centre of the sample and in the close neighboured of the crack, showing the shorter time needed for completion of test and faster rate of the pore pressure increase.

4 Conclusions

In this chapter the discrete element modelling suitable for describing the fracture process with localized failure zones in heterogeneous non-saturated and fully fluid-saturated poro-plastic medium is presented, where coupling between the fluid and solid obey the Biot theory of poroplasticity. The localized failure mechanisms are incorporated through the enhanced kinematics of Timoshenko beams that act as cohesive links between the grains of heterogeneous rock material. The embedded discontinuities can represent the failure modes I and II, as well as their combination. The fluid flow is governed by the Darcy law with assumed continuous pore pressure field.

The model ingredients are incorporated into the framework of embedded discontinuity finite element method, where the computation of the enhanced discontinuity parameters requires only local element equilibrium. Further use of the static condensation of the enhanced parameters at the element level, leads to the computationally very efficient approach and numerical implementation that fits within the standard finite element code architecture.

The main strength of the proposed discrete model lies in its ability to account for material heterogeneities with localized macro-cracks propagating throughout the weaker parts of the material and forming the irregular geometries. Such a phenomenon is presented by the numerical simulations of two samples with equal geometries and material properties, but slightly different distribution of material heterogeneities throughout samples, which present different behaviour in terms of localized macro-crack propagation. The solid-fluid coupling plays important role here as well, bringing the variations in macroscopic responses and compliance of the samples. It is important to emphasise that heterogeneous effects become more pronounced with the coupling effects and higher rates of the imposed velocities.

References

1. Terzaghi, K.: Theoretical Soil Mechanics. Wiley (1943)
2. Biot, M.A.: Mechanics of incremental deformations. John Wiley & Sons (1965)
3. Lewis, R.W., Schrefler, B.A.: The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media. Wiley, Chichester, 2nd edition (1998)
4. Zienkiewicz, O.C., Taylor, R.L: Finite Element Method. Butterworth-Heinemann, 5th edition (2000)
5. Bathe, K.: Finite Element Procedures. Prentice Hall, New Jersey (2006)
6. Ibrahimbegovic, A.: Nonlinear Solid Mechanics: Theoretical Formulations and Finite Element Solution Methods. Springer (2009)
7. Simo, J.C., Oliver, J., Armero, F.: An analysis of strong discontinuities induced by strain-softening in rate-independent inelastic solids. Comput. Mech. **12**, 277–296 (1993)
8. Simo, J.C., Rifai, M.S.: A class of mixed assumed strain methods and the method of incompatible modes. Int. J. Numer. Meth. Engng. **29(8)**, 1595–1638 (1990)
9. Ortiz, M., Leroy, Y., Needleman, A.: A finite element method for localized failure analysis. Comput. Methods Appl. Mech. Engrg. **61**, 189–214 (1987)
10. Ibrahimbegovic, A., Melnyk, S.: Embedded discontinuity finite element method for modeling of localized failure in heterogeneous materials with structured mesh: an alternative to extended finite element method. Comput. Mech. **40**, 149–155 (2007)
11. Moes, N., Dolbow, J., Belytschko, T.: A finite element method for crack growth without remeshing. Int. J. Numer. Meth. Engng. **46**, 131–150 (1999)
12. Fries, T.P., Belytschko, T: The intrinsic XFEM: a method for arbitrary discontinuities without additional unknowns. Int. J. Numer. Meth. Engng. **68**, 1358–13850 (2006)

13. Fries, T.P., Belytschko, T: The extended/generalized finite element method: An overview of the method and its applications. Int. J. Numer. Meth. Engng. **84**, 253–304 (2010)
14. Rethore, J., de Borst, R., Abellan, M.A: A two-scale approach for fluid flow in fractured porous media. Int. J. Numer. Meth. Engng. **71**, 780–800 (2007)
15. Rethore, J., de Borst, R., Abellan, M.A: A two-scale model for fluid flow in an unsaturated porous medium with cohesive cracks. Comput. Mech. **42**, 227–238 (2008)
16. de Borst, R., Rethore, J., Abellan, M.A: A numerical approach for arbitrary cracks in a fluid-saturated medium. Arch. Appl. Mech. **75**, 595–606 (2006)
17. Benkemoun, N., Gelet, R., Roubin, E., Colliat, J.B.: Poroelastic two-phase material modeling: theoretical formulation and embedded finite element method implementation. Int. J. Numer. Anal. Meth. Geomech. **39**, 1255–1275 (2015)
18. Callari, C., Armero, F.A.: Finite element methods for the analysis of strong discontinuities in coupled poro-plastic media. Comput. Methods Appl. Mech. Engrg. **191**, 4371–4400 (2002)
19. Callari, C., Armero, F., Abati, A.: Strong discontinuities in partially saturated poroplastic solids. Comput. Methods Appl. Mech. Engrg. **199**, 1513–1535 (2010)
20. Schrefler, B.A., Secchi, S., Simoni, L.: On adaptive refinement techniques in multi-field problems including cohesive fracture. Comput. Methods Appl. Mech. Engrg. **195**, 444–461 (2006)
21. Secchi, S., Schrefler, B.A.: A method for 3D hydraulic fracturing simulation. Int. J. Fract. **178**, 245–258 (2012)
22. Nikolic, M., Ibrahimbegovic, A., Miscevic, P.: Brittle and ductile failure of rocks: Embedded discontinuity approach for representing mode i and mode ii failure mechanisms. Int. J. Numer. Meth. Engng. **102**, 1507–1526 (2015)
23. Nikolic, M., Ibrahimbegovic, A.: Rock mechanics model capable of representing initial heterogeneities and full set of 3d failure mechanisms. Comput. Methods Appl. Mech. Engrg. **290**, 209–227 (2015)
24. Ostoja-Starzewski, M.: Lattice models in micromechanics. Appl. Mech. Rev. **55**, 35–60 (2002)
25. Cusatis, G., Bazant, Z., Cedolin, L.: Confinement-shear lattice CSL model for fracture propagation in concrete. Comput. Methods Appl. Mech. Engrg. **195**, 7154–7171 (2006)
26. Vassaux, M., Richard, B., Ragueneau, F., Millard, A., Delaplace, A.: Lattice models applied to cyclic behavior description of quasi brittle materials: advantages of implicit integration. Int. J. Numer. Anal. Meth. Geomech. **39**, 775–798 (2015)
27. Karihaloo, B., Shao, P., Xiao, Q.: Lattice modelling of the failure of particle composites. Eng. Fract. Mech. **70**, 2385–2406 (2003)
28. Bui, N., Ngo, M., Nikolic, M., Brancherie, D., Ibrahimbegovic, A.: Enriched timoshenko beam finite element for modeling bending and shear failure of reinforced concrete frames. Comput. Struct. **143**, 9–18 (2014)
29. Jukic, M., Brank, B., Ibrahimbegovic, A.: Embedded discontinuity finite element formulation for failure analysis of planar reinforced concrete beams and frames. Eng. Struct. **50**, 115–125 (2013)
30. Pham, B., Brancherie, D., Davenne, L., Ibrahimbegovic, A.: Stress-resultant models for ultimate load design of reinforced concrete frames and multi-scale parameter estimates. Comput. Mech. **51**, 347–360 (2013)
31. Nikolic, M., Ibrahimbegovic, A., Miscevic, P.: Discrete element model for the analysis of fluid-saturated fractured poro-plastic medium based on sharp crack representation with embedded strong discontinuities. Comput. Methods Appl. Mech. Engrg. **298**, 407–427 (2016)
32. Taylor R.L.: FEAP - A Finite Element Analysis Program, University of California at Berkeley

Reliability Calculus on Crack Propagation Problem with a Markov Renewal Process

Chrysanthi A. Papamichail, Salim Bouzebda and Nikolaos Limnios

Abstract This chapter concerns a stochastic differential system that describes the evolution of a degradation mechanism, the fatigue crack propagation. A Markov process will be considered as the perturbing process of the system that models the crack evolution. With the help of Markov renewal theory, we study the reliability of a structure and propose for it a new analytical solution. The method we propose reduces the complexity of the reliability calculus compared with the previous resolution method. As numerical applications, we tested our method on a numerical example and on an experimental data set, which gave results in good agreement with a Monte Carlo estimation.

1 Introduction and Motivation

To begin with, we explain our interest in the mechanical problem of fatigue crack growth and in reliability of stochastic models and give a basic bibliographic frame of the approach we rely on. In everyday industrial life, mechanical structures with variable life cycles are used and thus undergo the degradation process that leads to fatigue aging. The crack-growth problem concerns many industrial fields, such as aeronautics, nuclear and electrical power stations, etc. On the one hand, there is the need to avoid failure of the structure, caused by crack propagation, and, on the other hand, the wish of industry for maximum exploitation of the structure. Consequently, models that quantify the degradation level are necessary and important to be developed.

C.A. Papamichail · S. Bouzebda · N. Limnios (✉)
Sorbonne Universités, Université de Technologie de Compiègne,
rue du Dr Schweitzer, CS 60319, 60203 Compiegne Cedex, France
e-mail: nikolaos.limnios@utc.fr

S. Bouzebda
e-mail: salim.bouzebda@utc.fr

C.A. Papamichail
e-mail: chrysanthi.papamichail@utc.fr

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_13

Fatigue crack growth problematic is an issue of general engineering interest [1, 2]. Deterministic models (i.e., through finite-elements analysis) have been provided from structural mechanics, but do not adequately describe the degradation mechanisms. Admittedly, probabilistic modeling is required to describe such degradation process. The theory of stochastic differential equations (e.g., [3]) is essential here. Stochastic dynamical systems are found well adapted to the modelling of fatigue crack growth [4, 5].

In present work, the dynamical evolution of the increasing stochastic process Z_t, on $\mathbb{R}_+^*$, with initial condition $Z_0 = z_0$, is modeled by a first order stochastic dynamical system of the following form

$$\begin{cases} \frac{\mathrm{d}Z_t}{dt} = \mathbf{C}(Z_t, X_t), \\ Z_0 = z_0, \end{cases} \tag{1}$$

where $(Z_t, t \in \mathbb{R}_+)$ is the stochastic process that represents the crack length, $(X_t, t \in \mathbb{R}_+)$ is a stochastic process of values in the finite state space E and serves as the driving jump process that handles the randomness of the system, and, C is a function from $\mathbb{R}_+^d \times E$ to $\mathbb{R}_+^d$ with the appropriate existence and uniqueness properties. Concerning perturbation problems, Krylov-Bogoliubov averaging method can be used to approximate oscillatory processes in non-linear mechanics. The theorem known as "Bogolyubov's principle", presented by [6], gives an asymptotic convergence of (1), when a change of scale $t \to t/\epsilon$, with $\epsilon \to 0$ takes place. In this case, the process $X_{t/\epsilon}$ has the impact on the system that it would have after a long time interval, $t/\epsilon \to \infty$, and the process Z_t^ϵ converges weakly to the solution of the deterministic system.

In our study, X_t is assumed to be a jump Markov process and the coupled process (Z_t, X_t) has the Markov property (see e.g., [7] for Markov and semi-Markov processes and [8] for Markov models).

Results on the stochastic approximation of dynamical systems, by weak convergence techniques are encountered in [9]. General and particular schemes of proofs for average, diffusion, and Poisson approximations of stochastic systems are presented, allowing one to simplify complex systems and obtain numerically tractable models. All these systems are switched by Markov and semi-Markov processes whose phase space is considered in asymptotic split and merging schemes.

The process (Z_t, X_t) belongs to the class of stochastic models sometimes called piecewise deterministic Markov processes (PDMP). The theory for PDMPs, which are alternative to diffusion processes, underlies in present analysis and is introduced in [10, 11] and further developed in [12], where it is based exclusively on the theory of marked point processes. The class of PDMPs is considered and recognized as a powerful modelling tool for complex systems. The main objective of [13] consists in presenting mathematical tools recently developed by the authors on theoretical and numerical aspects of simulation and optimization for PDMPs. This book is focused on the computation of expectation of functionals of PDMPs with applications to the evaluation of service times and on solving optimal control problems with applications to maintenance optimization. Optimal control of PDMPs [14–16], and optimal stopping have been studied, [17]. Results in the control theory for the long-run average continuous control problem of PDMPs are presented in [18] and stability and ergodicity of PDMPs are studied in [19]. Azais [20], proposed and analyzed nonparametric estimation methods for both the features governing the randomness of PDMPs. Riedler [21], presented an almost sure convergence analysis for numerical simulation algorithms for PDMPs. PDMPs form a general class of stochastic hybrid models covering a large number of problems such as engineering systems, operation research, management science, biology, internet traffic, networks, reliability, computer science, neuroscience (i.e., [21]), mobile robotics, finance and insurance etc.

This PDMP in our case is associated to a Markov renewal process (MRP) (see e.g., [7, 22] for Markov renewal processes, [23, 24] for Markov renewal theory). In [25] the perturbing process is associated to a Markov process whereas in [26] to a semi-Markov process, and the associated dynamical system is respectively described.

Here, we are particularly interested in reliability of the stochastic model (1). Admittedly, any advance in stochastic processes is sooner or later applied in reliability theory. Reliability, as a measure, quantifies the capability of a system or a service to perform its expected job under the specified conditions of use over an intended period of time. This capability can be used to compare the performance of different types of systems and take decision of their suitability. Another advantage is that there is no need to redefine or modify the quantification of reliability for different kinds of engineering products or systems. Due to the immense interest and research progress in reliability theory, it is impossible to provide any exhaustive literature on it. We shall restrict ourselves to some authors we consider inspiring for our study, with the risk, however, to have omitted others with remarkable, as well, contribution to the field.

Barlow's and Proschan's books, [27, 28], have influenced the field of reliability engineering and statistical reliability for many years. Among Barlow's achievements is his work on the theory and methodology of modeling the failure rate of systems and components. Some of his work was closely related to engineering applications, such as his work on accelerated degradation models and on fault tree analysis in order to demonstrate the safety of nuclear plants. Another lasting contribution of his is the reliability importance measure of the components in a system, referred to as Barlow and Proschan's Importance Measure, which is widely used in practice and always used for comparison with new or alternative measures. The books of Gnedenko et al. [29], and, Gnedenko and Ushakov [30], are of considerable influence, as well. In [30], the authors developed advanced statistical methods for reliability analysis. The book of [31] includes mathematical models associated with probabilistic methods of reliability analysis. Also, it provides a detailed treatment of reliability indices, the structure function, load-strength reliability models, distributions with monotone intensity functions, repairable systems, the Markov models, analysis of performance effectiveness, optimal technical diagnosis and heuristic methods in reliability. The set-theoretic approach to reliability theory and the central concepts of set theory to the phenomena are considered by [32]. The authors present methods of finding estimates for reliability parameters based on observations and methods of testing reliability hypotheses, as well as a method that increases the reliability of manufactured articles-redundancy.

Bedford offers an insight into mathematical modeling in reliability theory and its applications [33]. Interesting topics in his book are the relations among aging and stochastic dependence, as well as how competing risks arise in reliability and maintenance analysis through the ways in which data is censored. Nakagawa [34], gives a basic treatment of stochastic processes, such as Markov renewal processes, with related reliability studies and applications. Aven and Jensen [35], provide an up-to date presentation of some of the classical areas of reliability based on a more advanced probabilistic framework using the modern theory of stochastic processes. Also, they adopt a (semi-) martingale approach for analyzing failure-prone systems. With the aid of this general approach, they formulate a unifying theory of both nonrepairable and repairable systems, with Markov processes as a special case, among others. Birolini investigated renewal and alternating renewal processes, Markov processes with a finite state space, semi-Markov processes, regenerative stochastic processes and some kinds of non-regenerative stochastic processes used in the modeling of reliability problems, along with the respective reliability models, [36]. He gave emphasis to the theoretical and computational limitations involved in the various analytical procedures for reliability resolution, as well as to the unification and extension of the reliability models known in the literature. In [37], the state-of-the art of reliability engineering is presented, based also on the author's experience in industry. In [38], the authors presented homogeneous and non-homogeneous semi-Markov models with finitely many states for real life problems of risk management such as reliability and proposed basic algorithms for the respective theoretical models.

More specifically, reliability of models that are described with PDMPs also attracts research attention. The semi-Markov processes generalize the renewal processes and the Markov jump processes with applications in reliability, [39]. The importance of hitting times to reliability is indicated in [40]. This paper systematically models reliability for semi-Markov processes with a general state space, which generalizes results from finite state spaces. Dynamic reliability in the frame of PDMP is studied by [41]. Estimates for some functionals of PDMP, in view of a sensitivity analysis in dynamic reliability, are computed by [42]. The theory of dynamic reliability and how it models random loads, which is a common situation in structural reliability problems, is considered in [43].

Cocozza-Thivent and Mercier studied dynamic reliability models. The authors proposed a finite-volume scheme that approximates the probability measures that are the solution to the probability of the state of the system, in [44] and they characterized the marginal distribution of the PDMP, in [45].

The rest of this chapter is organized as follows. In Sect. 2, the stochastic model is described in detail. As necessary, basic elements of Markov Renewal Theory are included, along with the semi-Markov kernel and the transition function of the Markov process. In Sect. 3, reliability calculation is presented, first, as it follows from [46]. At this point we propose our method on reliability calculation which is the main contribution of the present work. This method offers a deviation in calculation procedure compared to the previous one. Also, the procedure of its numerical implementation and the accompanying algorithms are provided. Next, Sect. 4 describes a method for estimation of the reliability function, as presented and applied in [25]. Its utility is to offer an estimator for reliability, as comparison measure of the method we proposed for reliability calculus. Sect. 5 includes our numerical results on reliability as application of our calculation method presented in Sect. 3, in comparison with the estimation method of Sect. 4. A few concluding remarks are given in Sect. 6.

2 The Model Settings and Elements of Markov Renewal Theory

In this Section, the model (1) is thoroughly described and the essential theoretical background is provided, as well. Particularly, the following points are to be exposed here:

1. The Markov process X_t is defined, as well as its probability transition function and infinitesimal generator. Also, it is necessary here to give the basic frame of **Markov renewal theory**.
2. The coupled process (Z_t, X_t) is described along with the associated probability transition function, infinitesimal generator and Markov Renewal Equation
3. The conditions of the function C are given.

In this section, the principal references are [7, 26, 46].

2.1 Basic Definitions

Assuming that the process X_t of the model is a jump Markov process it is necessary to begin with the respective definitions. First, let us define a Markov process and its transition probability function, in the following way:

Definition 1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $(X_t, t \in \mathbb{R}_+)$ a random process with values in measurable state space E of σ-algebra $\mathcal{E}$. We note as $\mathcal{F}_t$ the σ-algebra of events generated by $(X_s, 0 \leq s \leq t)$ and $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ the associated filtration. The process X_t is a Markov process if for all $B \in \mathcal{E}$ and all $s, t \in \mathbb{R}_+$, such that $0 \leq s < t$, it satisfies

$$\mathbb{P}(X_t \in B|\mathcal{F}_s) = \mathbb{P}(X_t \in B|X_s), \qquad p.s.$$

Also, X_t is **homogeneous** with respect to time if for all $s, t \in \mathbb{R}_+$ and all $x \in E$

$$\mathbb{P}(X_t \in B|X_0 = x) = \mathbb{P}(X_{t+s} \in B|X_s = x). \tag{2}$$

For a Markov homogeneous process we denote as $P(x, B, t)$ the probability (2). The function defined as $P : (x, B, t) \rightarrow P(x, B, t)$ for $x \in E$, $B \in \mathcal{E}$, $t \in \mathbb{R}_+$ is called **transition function** of the process.

We are concerned only about homogeneous Markov processes. The essential property of Markov processes is on Chapman-Kolmogorov equation:

Definition 2 The transition function of a homogeneous Markov process verifies the equation:

$$P(x, B, t) = \int_E P(x, dy, s)P(y, B, t-s), \qquad 0 \leq s \leq t,$$

called Chapman-Kolmogorov equation.

It is indispensable for a Markov process that its infinitesimal generator is defined, after the definition of the expectancy:

Definition 3 The expectancy with respect to the probability mesure $P(x, \cdot, t)$ of a real function f, measurable and bounded, is defined by

$$\mathbb{E}_x f(X_t) = \mathbb{E}[f(X_t)|X_0 = x] = \int_E f(y)P(x, dy, t).$$

The expectation $\mathbb{E}_x[\cdot]$ serves to interpret the transition function as an operator $\mathcal{P}_t$

$$\mathcal{P}_t f(x) = \mathbb{E}_x f(X_t).$$

The set $\{\mathcal{P}_t, t \geq 0\}$ is called semi-group of transition functions. Now, the **infinitesimal generator** of the process can be defined.

Definition 4 The infinitesimal generator of a Markov process is an operator $\mathcal{A}$ defined by the following limit

$$\mathcal{A}f(x) = \lim_{t \to 0} \frac{\mathcal{P}_t f(x) - f(x)}{t},$$

as long as it exists, for all $x \in E$. The set of functions f, measurable and bounded, for which this limit exists, is called domain of $\mathcal{A}$. The generator corresponds to the derivative with respect to the time of expectancy calculated on 0^+.

This, along with the initial law $\alpha(B) = \mathbb{P}(X_0 \in B)$, $B \in \mathcal{E}$, characterizes completely the X_t.

Next, we pass to the definition of the **jump Markov processes**, X_t being one, which are a particular set among Markov processes:

Definition 5 The jump Markov process is a Markov process with values in a countable set E, where the evolution is realized with jump from state to state and almost all trajectories (i.e. with probability one) are constant except for isolated points corresponding to jumps.

The letters i, j are used for countable or finite state spaces, which is the case of the present study. We denote as $P : (i, j, t) \to P_{ij}(t)$ for $i, j \in E, t \in \mathbb{R}_+$ the transition function associated to a Markov process with countable state space, i.e. with respect to the general case, $P_{ij}(t) = P(i, \{j\}, t)$. We set the function $f : i \to \mathbb{1}_j(i)$ with $i, j \in E$ where $\mathbb{1}_j(i)$ defines the indicator function equal to 1 if $i = j$, 0 if not. So, $\mathcal{P}_t f(i) = \mathcal{P}_t \mathbb{1}_j(i) = P_{ij}(t)$ and we can represent the generator with a matrix $\mathbf{A} = (a_{ij})_{i,j \in E}$ verifying the relation:

Definition 6 The matrix $\mathbf{A} = (a_{ij})_{i,j \in E}$ of generator of a Markov process with countable state space is given by

$$a_{ij} = \lim_{t \to 0} \frac{P_{ij}(t) - \delta_{ij}}{t}, \quad \text{where } \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if not.} \end{cases}$$

For the jump Markov processes, we consider only generators stable and conservative, i.e. verifying

$$\begin{cases} a_{ij} \geq 0, & \text{if } \forall i \neq j \\ a_{ii} = -a_i = -\sum_{l \in E, l \neq i} a_{il}. \end{cases} \tag{3}$$

Since the state space E is a countable part of $\mathbb{N}$, the generator is represented by a square matrix of dimensions $s \times s$, where s is the cardinal of E.

2.2 Markov Renewal Process and Semi-Markov Kernel

At this point, the notion of a **Markov Renewal Process** should be introduced. **Semi-Markov processes** are a natural generalization of Markov processes: the constraint that all states sojourn times follow exponential distribution is relaxed. A Markov Renewal Process of two components is associated to a semi-Markov process:

1. one that describes the states successively visited and
2. another for the instants of change of state for the process.

Let $0 = S_0 \leq S_1 \leq \cdots \leq S_n \leq S_{n+1} \leq \ldots$ be the sequence of random variables that describe the successive instants of jump of X_t and $W_n = S_{n+1} - S_n$, for $n \in \mathbb{N}$, the sojourn times of the states.

The sequence $(J_n, n \in \mathbb{N})$, composed by the successively visited states taken up by X_t is a Markov chain called **embedded Markov chain** (EMC). Note that S_0 may also take positive values. Let $\mathbb{N}$ be the set of non-negative integers. Then, X_t is connected to (J_n, S_n) through

$$X_t = J_n, \quad \text{if } S_n \leq t < S_{n+1}, \ t \geq 0 \quad \text{and} \quad J_n = X_{S_n}, \ n \in \mathbb{N}.$$

The embedded chain J_n corresponds to successive values of X_t on the intervals $[S_n, S_{n+1})$. The process X_t can be written

$$X_t = \sum_{n \in \mathbb{N}} J_n \mathbb{1}_{[S_n, S_{n+1})}(t), \tag{4}$$

where the indicator function $\mathbb{1}_A$ is defined as

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{if} \, x \in A, \\ 0, & \text{if} \, x \notin A. \end{cases}$$

The embedded chain J_n has the same initial law as X_t and is completely characterized by the properties:

1. If X_t is on state i on the interval $[S_n, S_{n+1})$, the law of sojourn time conditional at state i is exponential of parameter a_i
2. The law of change of state for the embedded Markov chain J_n is given by transition matrix $\mathbf{P} = (p_{ij})_{i,j \in E}$ verifying

$$p_{ij} = \begin{cases} \frac{a_{ij}}{a_i}, & \text{if } i \neq j \text{ and } a_i \neq 0, \\ 0, & \text{if not.} \end{cases}$$

We suppose that the process X_t is an homogeneous jump Markov process with countable state space E, generator $\mathbf{A}$ and initial law $(\alpha(i))_{i \in E}$.

Definition 7 The stochastic process $(J_n, S_n)_{n \in \mathbb{N}}$ is said to be a Markov renewal process (MRP), with state space E, if it satisfies, a.s., the following equality

$$\begin{aligned} &\mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t | J_0, \ldots, J_n; S_1, \ldots S_n) \\ &\quad = \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t | J_n), \end{aligned}$$

for all $j \in E$, all $t \geq 0$ and all $n \in \mathbb{N}$. In this case, X_t is called a semi-Markov process (SMP).

Remark 1 We assume that the above probability is independent of n and S_n, and in this case the MRP is called time homogeneous. Only time-homogeneous MRP are considered in the sequel.

The MRP $(J_n, S_n)_{n \in \mathbb{N}}$ is determined by

1. the initial distribution α, with $\alpha_i = \mathbb{P}(J_0 = i), i \in E$, and
2. by the transition kernel

$$Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t | J_n = i),$$

called the **semi-Markov kernel**.

The EMC J_n has E as its state space and transition probabilities $p_{ij} := Q_{ij}(\infty) := \lim_{t\to\infty} Q_{ij}(t)$. To denote here that $Q_{ii}(t) \equiv 0$, for all $i \in E$, but in general we can consider semi-Markov kernels by dropping this hypothesis.

An important point is the following decomposition of the semi-Markov kernel

$$Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \le t | J_n = i) = p_{ij} F_{ij}(t), \quad t \ge 0, i, j \in E,$$

where $F_{ij}(t) := \mathbb{P}(S_{n+1} - S_n \le t | J_n = i, J_{n+1} = j)$ is the **conditional distribution function** of the sojourn time in the state i given that the next visited state is $j, i \neq j$.

Let us also define the **distribution function** $H_i(t) := \sum_{j\in E} Q_{ij}(t)$ and its mean value m_i, which is the **mean sojourn time** of X_t in state i. In general, Q_{ij} is a subdistribution, i.e., $Q_{ij}(\infty) \le 1$, hence H_i is a distribution function, $H_i(\infty) = 1$, and $Q_{ij}(0-) = H_i(0-) = 0$.

Markov process is a particular case of semi-Markov process and its semi-Markov kernel is as follows.

Example 1 A Markov process with state space $E = \mathbb{N}$ and generating matrix $\mathbf{A} = (a_{ij})_{i,j\in E}$ is a special semi-Markov process with semi-Markov kernel

$$Q_{ij}(t) = \frac{a_{ij}}{a_i}(1 - \mathrm{e}^{-a_i t}), \quad i \neq j, \quad a_i \neq 0,$$

where $a_i := -a_{ii}$, $i \in E$, and $Q_{ij}(t) = 0$, if $i = j$ or $a_i = 0$. In this case the transition function of the EMC is $p_{ij} = a_{ij}/a_i$ and we recover an exponential distribution for the conditional distribution function of the sojourn time such as $F_i(t) = 1 - \mathrm{e}^{-a_i t}$, with $t \ge 0$.

We introduce the counting process $(N(t), t \ge 0)$ which counts the number of jumps of X_t in the time interval $(0, t]$, by $N(t) := \sup\{n \ge 0 : S_n \le t\}$. Also, define $N_i(t)$ to be the number of visits of X_t to state $i \in E$ in the time interval $(0, t]$. That is to say,

$$N_i(t) := \sum_{n=0}^{N(t)} \mathbb{1}_{\{J_n=i\}} = \sum_{n=0}^{\infty} \mathbb{1}_{\{J_n=i, S_n\le t\}}.$$

If we consider the renewal process $(S_n^i)_{n\ge 0}$ of successive times of visits to state i, then $N_i(t)$ is the counting process of renewals. Now a semi-Markov process X_t is said to be regular if

$$\mathbb{P}_i(N(t) < \infty) = 1,$$

for any $t \ge 0$ and any $i \in E$. As usual, $\mathbb{P}_i(\cdot)$ means $\mathbb{P}(\cdot | J_0 = i)$, and $\mathbb{E}_i$ is the corresponding expectation. For **regular semi-Markov processes** we have $S_n < S_{n+1}$, for any $n \in \mathbb{N}$, and $S_n \to \infty$. In the sequel, we are concerned with regular semi-Markov processes, and in particular with regular Markov processes.

2.3 Markov Renewal Equation

Next, we should refer to **Markov renewal equation** (MRE), an essential tool in semi-Markov theory which can be solved using the so-called **Markov renewal function**. Necessary for the definition of this function is the **convolution in the Stieljes-sense**. For $\phi(i,t), i \in E, t \geq 0$ a real-valued measurable function, the convolution of ϕ by Q is defined by

$$Q * \phi(i,t) := \sum_{k \in E} \int_0^t Q_{ik}(ds)\phi(k, t-s).$$

Now, consider the n-fold convolution of Q by itself. For any $i, j \in E$,

$$Q_{ij}^{(n)}(t) = \begin{cases} \sum_{k \in E} \int_0^t Q_{ik}(ds) Q_{kj}^{(n-1)}(t-s), & n \geq 2, \\ Q_{ij}(t), & n = 1, \\ \delta_{ij} \mathbb{1}_{\{t \geq 0\}}, & n = 0, \end{cases} \tag{5}$$

where δ_{ij} is the Kronecker delta, that is to say, $\delta_{ij} = 1$ if $i = j$, 0 otherwise. It is easy to prove the following fundamental equality

$$Q_{ij}^{(n)}(t) = \mathbb{P}_i(J_n = j, S_n \leq t).$$

The **Markov renewal function** $\psi_{ij}, i, j \in E, t \geq 0$ is defined by

$$\begin{aligned} \psi_{ij}(t) := \mathbb{E}_i[N_j(t)] &= \mathbb{E}_i \sum_{n=0}^{\infty} \mathbb{1}_{\{J_n = j, S_n \leq t\}} \\ &= \sum_{n=0}^{\infty} \mathbb{P}_i(J_n = j, S_n \leq t) \\ &= \sum_{n=0}^{\infty} Q_{ij}^{(n)}(t). \end{aligned}$$

In matrix form, this is written as

$$\psi(t) = (I(t) - Q(t))^{(-1)} = \sum_{n=0}^{\infty} Q^{(n)}(t),$$

or alternatively

$$\psi(t) = I(t) + Q * \psi(t), \tag{6}$$

where $I(t) = I$ (the identity matrix), if $t \geq 0$ and $I(t) = 0$, if $t < 0$.

Equation (6) is a special case of what is called a **Markov Renewal Equation** (MRE). A general MRE is one of the following form:

$$\Theta(t) = L(t) + Q * \Theta(t), \tag{7}$$

where $\Theta(t) = (\Theta_{ij}(t))_{i,j \in E}$, $L(t) = (L_{ij}(t))_{i,j \in E}$ are matrix-valued measurable functions, with $\Theta_{ij}(t) = L_{ij}(t) = 0$ for $t < 0$. The function $L(t)$ is known while $\Theta(t)$ is the unknown. The (i, j) entry of the Eq. (7) is

$$\Theta_{ij}(t) = L_{ij}(t) + \sum_{k \in E} \int_0^t Q_{ik}(du)\Theta_{kj}(t-u), \tag{8}$$

If $\Theta(t)$ and $L(t)$ are vector valued functions, the jth element is written

$$\Theta_j(t) = L_j(t) + \sum_{k \in E} \int_0^t Q_{jk}(du)\Theta_k(t-u), \tag{9}$$

Without any loss of generality, Eq. (9) can be considered in the place of Eq. (8).

Let $\mathcal{M}$ be the space of all bounded vectors $\Theta(t)$ such that $||\Theta(t)|| = \sup_{i \in E} |\Theta_i(t)|$ is bounded with respect to t on the bounded intervals of $\mathbb{R}_+$. The Markov Renewal Theorem that follows gives results on the existence and unicity of a solution to a MRE as Eq. (7).

Theorem 1 ([23]). *Equation (7) has a solution Θ belonging to $\mathcal{M}$, if and only if $\psi * L$ belongs to $\mathcal{M}$. Any solution Θ can be represented in the form $\Theta(t) = \psi * L(t) + C(t)$, where C satisfies the equation $C * L(t) = C(t)$, $C(t) \in \mathcal{M}$. A unique solution of (7) of the form $\Theta(t) = \psi * L(t)$ exists if one of the following five conditions are fulfilled:*

1. *The state space E is finite.*
2. *The EMC is irreducible and positive recurrent.*
3. $\sup_{i \in E} H_i(t) < 1$ *for some* $t > 0$.
4. *$L_{ij}(t)$ is uniformly bounded in i for every j and $t \in \mathbb{R}_+$, and for every i there exists a $c > 0$ such that $H_i(c) < 1 - \epsilon$. In this case, the unique solution is uniformly bounded in i, for every $j \in E$ and $t > 0$.*

It can be shown that the **semi-Markov transition function** defined as

$$P_{ij}(t) := \mathbb{P}(X_t = j | X_0 = i), \quad i, j \in E, \quad t \geq 0,$$

which is the conditional marginal law of the process, satisfies a particular MRE, which will be an essential result in the sequel. We can find, e.g., in [7, 22, 23] the following result.

Proposition 1 ([26]). *The transition function $P(t) = (P_{ij}(t))$ satisfies the MRE:*

$$P(t) = I(t) - H(t) + Q * P(t),$$

which has the unique solution

$$P(t) = \psi * (I(t) - H(t)),$$

and, for any $i, j \in E$,

$$\lim_{t\to\infty} P_{ij}(t) = \nu_i m_i / m =: \pi_i.$$

Here $H(t) = diag(H_i(t))$ *is a diagonal matrix.*

To note here that, ν is the stationary distribution of the embedded Markov chain J_n and π is the **stationary distribution** of the Markov process, as defined:

Definition 8 Let X_t be an irreducible Markov process, with finite state space E and generator $\mathbf{A} = (a_{ij})_{i,j\in E}$. The probability π, unique solution of the equation

$$\sum_{j\in E} \pi(j) a_{ij} = 0, \quad i \in E,$$

is called stationary law of process X_t.

The stationary law π has the properties, see, e.g., [9, 58]:

Property 1 ([47]). *Let* X_t *be an ergodic Markov process, of transition function* $P_{ij}(t)$ *and of stationary law* π. *We have the following properties:*

(a) For all $i, j \in E$, $\lim_{t\to\infty} P_{ij}(t) = \pi(j)$.
(b) For all function $f : E \to \mathbb{R}$,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(X_t) dt = \sum_{j\in E} \pi(j) f(j) \ \ a.s.$$

2.4 Further Model Settings

Now, we turn back to the model (1), which was presented in [25], and consider the process Z_t and the properties of the function C.

Let $(Z_t, t \in \mathbb{R}_+)$ be an increasing stochastic process on $\mathbb{R}_+^*$. Its initial condition is $Z_0 = z_0$ with $z \in \mathbb{R}_+$ is a realisation of the random variable Z_0 that admits a probability law β defined as $\beta(B) = \mathbb{P}(Z_0 \in B)$, with $B \in \mathcal{B}$, where $\mathcal{B}$ is the σ-algebra of Borel sets of $\mathbb{R}$. Looking towards reliability analysis, we define for the process Z_t a set of working states $U = [z_0, \Delta)$, with $0 < z_0 < \Delta$, and a set of down states $D = [\Delta, \infty)$. Its time evolution is continuous, with positive values, thus Z_t necessarily passes through the point Δ while reaching the set of down states.

Properties of degradation process Z_t:

1. The degradation process Z_t takes its values in $\mathbb{R}^d_+$.
2. The level of degradation evolves in monotone increasing way in time.
3. The degradation domain limits to a critical threshold not to be overpassed, noted as Δ. The failure time τ is defined as the random variable that describes the time that Z_t enters failure domain defined by

$$\tau = \inf\{t \geq 0 : Z_t \geq \Delta\}.$$

4. The trajectories of degradation level are the only observable data. They are measured from an initial instant up to the instant that the degradation level reaches the degradation threshold Δ.

For a function $\mathbf{C}$ we set the assumptions

Assumption 1 (a) The function $\mathbf{C} : \mathbb{R} \times E \to \mathbb{R}$ is measurable, of class C^1 on $\mathbb{R} \times E$. (b) There is a function $h : E \to \mathbb{R}$ such that, for $x, y \in \mathbb{R}$ and $i \in E$,

$$|\mathbf{C}(x, i) - \mathbf{C}(y, i)| \leq h(i)|x - y|,$$

which is that $\mathbf{C}$ is Lipschitz with respect to its first component.

Assumption 2 The function $\mathbf{C} : (y, i) \to \mathbf{C}(y, i)$ is continuous on $\mathbb{R}$ for fixed $i \in E$, bounded and Lipschitz with respect to its first component.

For $y \in \mathbb{R}$ fixed, there is a mean function $\mathbf{C}_0$ defined as

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{C}(y, X_s) ds = \mathbf{C}_0(y), \quad \text{a.s.}$$

Assumption 3 The function $\mathbf{C} : (z, i) \to \mathbf{C}(z, i)$ is positive for all $z \in \mathbb{R}_+$, $i \in E$, so, the trajectories of Z_t are increasing.

Assumption 4 The support of the initial law β of Z_t is a set $\mathcal{D}_0 = [0, A]$, such that $A < \Delta$, where Δ is the threshold of system failure.

Under these assumptions, Eq. (1) admits a unique solution, using the standard results of existence and uniqueness of a solution for the classic differential systems [48, 49]. So, Markov character of process X_t is not necessary in order to construct a solution for Z_t but uniquely the fact that its trajectories are constant between isolated jumps, which is true for any jump process. If we fix X_t to i, we obtain

$$\begin{cases} \frac{\mathrm{d}z_t}{dt} = \mathbf{C}(z_t, i), \\ z_0 = z, \end{cases} \tag{10}$$

with $z \in \mathbb{R}_+$ and $i \in E$. The Eq. (10) constitutes a classic Cauchy problem [50] which under Assumption 1 has a solution

$$z_t = \phi_{z,i}(t), \quad t \geq 0.$$

The jump process X_t is constant on the successive intervals $[S_n, S_{n+1}), n \in \mathbb{N}$. A unique process solution Z_t is constructed by pieces, associated to X_t and to function C, by solving successively the differential equation of type (10). In order to determine Z_t up to instant t we calculate the pieces of trajectory of $(Z_s, 0 \leq s \leq t)$

$$Z_s = \begin{cases} \phi_{z,J_0}(s), s \in [S_0, S_1), \\ \phi_{Z_{S_1},J_1}(s - S_1), s \in [S_1, S_2), \\ \vdots \\ \phi_{Z_{S_n},J_n}(s - S_n), s \in [S_n, t), t \in [S_n, S_{n+1}). \end{cases} \tag{11}$$

Proposition 2 *Let the stochastic differential Eq.(1) with $(X_t, t \in \mathbb{R}_+)$ a jump process and let the Assumption 1 be verified for the function* **C**. *So there is a unique process $(Z_t, t \in \mathbb{R}_+)$ solution of this system with continuous trajectories.*

We suppose that the variables Z_0 and X_0 are independent so the initial law of the couple (Z_t, X_t) is given by

$$\mu(B, i) = \mathbb{P}(Z_0 \in B, X_0 = i) = \mathbb{P}(Z_0 \in B)\mathbb{P}(J_0 = i) = \alpha(i)\beta(B). \tag{12}$$

Proposition 3 *Since Z_0 and X_0 are independent, the couple (Z_t, X_t) is a Markov process with values in the state space $\mathbb{R}_+ \times E$.*

Proof Let $\widetilde{\mathcal{F}}_t$ the σ-algebra of events produced by the couple process $(Z_s, X_s, 0 \leq s \leq t)$. By conditioning, we have for all $s, t \in \mathbb{R}_+$ such that,

$$\mathbb{P}(Z_t \in B, X_t = j|\widetilde{\mathcal{F}}_s) = \mathbb{P}(X_t = j|\widetilde{\mathcal{F}}_s) \times \mathbb{P}(Z_t \in B|\widetilde{\mathcal{F}}_s, X_t = j),$$

where $B \in \mathcal{B}^d$. The process X_t being Markov and for X_0 independent of Z_0, the first term of the product becomes:

$$\mathbb{P}(X_t = j|\widetilde{\mathcal{F}}_s) = \mathbb{P}(X_t = j|X_s).$$

The second term of the equation, since Z_t is determined by (1) and by the process $(Z_s, X_s, 0 \leq s < t)$, becomes

$$\mathbb{P}(Z_t \in B|\widetilde{\mathcal{F}}_s, X_t = j) = \mathbb{P}(Z_t \in B|X_s, Z_s),$$

provided that X_0, Z_0 are independent. Finally

$$\mathbb{P}(Z_t \in B, X_t = j|\widetilde{\mathcal{F}}_s) = \mathbb{P}(Z_t \in B, X_t = j|X_s, Z_s).$$

The Markov quality of the couple (Z_t, X_t) in addition to the piecewise construction of Z_t trajectories as presented in (11) justifies the name of **piecewise deterministic Markov processes** attributed to (Z_t, X_t).

Now, the functions that characterize (Z_t, X_t) are

- the transition function, for which the entire following paragraph, Sect. 2.5, is devoted and
- the infinitesimal generator of the couple.

One of the most important results for the PDMPs such as the couple (Z_t, X_t) concerns the infinitesimal generator $\mathcal{B}$, which is given in an explicit form in the following proposition:

Proposition 4 ([51]). *The infinitesimal generator $\mathcal{B}$ of the couple (Z_t, X_t) refers to the real functions $f : (z, i) \to f(z, i)$, for $z \in \mathbb{R}^d$, $i \in E$ bounded and differentiable with respect to first argument:*

$$\mathcal{B} f(z, i) = \sum_{k=1}^{d} C_k(z, i) \frac{\partial}{\partial z_k} f(z, i) + \sum_{j \in E} a_{ij} [f(z, j) - f(z, i)],$$

where C_k corresponds to the kth component of $C(z, i)$.

2.5 The Transition Probability Function of the PDMP and Its Markov Renewal Equation

The **transition probability function** P defined by

$$P_{ij}(z, B, t) := \mathbb{P}_{z,i}(Z_t \in B, X_t = j), \quad i, j \in E, B \in \mathcal{B}, \tag{13}$$

where B is a subset of $\mathcal{B}$, the Borel σ field of $\mathbb{R}_+$ and

$$\mathbb{P}_{z,i}(\cdot) := \mathbb{P}(\cdot | Z_0 = z, X_0 = i).$$

A Markov process is a special MRP, thus we may associate to (Z_t, X_t) the extended MRP $(\zeta_n, J_n, S_n, n \in \mathbb{N})$ such as

$$\zeta_n = Z_{S_n}, \qquad J_n = X_{S_n}, \quad n \in \mathbb{N}.$$

The process (J_n, S_n) is a standard MRP, while (ζ_n, J_n, S_n) is an extended one. The associated semi-Markov kernel Q is defined, for $t > 0$, by

$$Q_{ij}(z, B, t) := \mathbb{P}(Z_{n+1} \in B, J_{n+1} = j, S_{n+1} - S_n \leq t | Z_n = z, J_n = i). \tag{14}$$

The Stieltjes-convolution is denoted by "$*$", hence, the successive n-fold convolutions of the semi-Markov kernel Q are defined recursively. For $n = 0, 1$

$$Q_{ij}^{(0)}(z, B, t) = \mathbb{1}_{\{i=j\}}\mathbb{1}_B(z)\mathbb{1}_{\mathbb{R}_+}(t), \qquad Q_{ij}^{(1)}(z, B, t) = Q_{ij}(z, B, t),$$

where $\mathbb{1}_B(x)$ is the indicator function, i.e., $\mathbb{1}_B(x) = 1$ if $x \in B$, 0 otherwise. For $n \geq 2$, the n-fold convolution turns to

$$\begin{aligned} Q_{ij}^{(n)}(z, B, t) &:= (Q * Q^{(n-1)})_{ij}(z, B, t) \\ &= \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, dy, ds) Q_{kj}^{(n-1)}(y, B, t - s). \end{aligned} \tag{15}$$

The Markov renewal function Ψ, which plays a central role, is defined by [7]

$$\Psi_{ij}(z, B, t) = \sum_{n \geq 0} Q_{ij}^{(n)}(z, B, t). \tag{16}$$

In the case at hand, we have (ζ_n, J_n, S_n) a normal MRP, that is, $\Psi_{ij}(z, B, t) < \infty$ for any fixed $t > 0$, $z > 0$, $B \in \mathcal{B}$ and $i, j \in E$.

A MRE has the following form

$$\Theta_{ij}(z, B, t) = g_{ij}(z, B, t) + (Q * \Theta)_{ij}(z, B, t), \tag{17}$$

where g is a known function defined on $\mathbb{R}_+ \times E \times \mathbb{R}_+$ and Θ is the unknown function. The solution is given by

$$\Theta_{ij}(z, B, t) = (\Psi * g)_{ij}(z, B, t). \tag{18}$$

Applying the Markov renewal theory, a MRE for P is introduced. First, Q is to calculate:

Lemma 1 *The semi-Markov kernel Q of the MRP (ζ_n, J_n, S_n) verifies for $i \neq j$,*

$$Q_{ij}(z, B, dt) = a_{ij}\mathrm{e}^{-a_i t}\delta_{\phi_{z,i}(t)}(B)dt,$$

where $\delta_x(B)$ is the Dirac distribution, equal to 1 if $x \in B$, 0 otherwise.

Proof Assuming that $S_0 = 0$, and conditioning on Eq. (14), we get

$$Q_{ij}(z, B, dt) = \mathbb{P}_{z,i}(J_1 = j, S_1 \in dt)\mathbb{P}_{z,i}(\zeta_1 \in B | J_1 = j, S_1 = t).$$

First, since S_1 and J_1 are independent, we have $\mathbb{P}_{z,i}(J_1 = j, S_1 \in dt) = a_{ij}\,\mathrm{e}^{-a_i t}\,dt$ from the usual results of Markov theory. Second, Z_t is fully characterized by $\phi_{z,i}(t)$ before the first jump time S_1, thus $\mathbb{P}_{z,i}(\zeta_1 \in B | J_1 = j, S_1 = t) = \mathbb{P}_{z,i}(Z_t \in B) =$

$\delta_{\phi_{z,i}(t)}(B)$. Indeed the probability $\mathbb{P}_{z,i}(Z_t \in B)$ is zero everywhere, except for the time points where B is reached. We hence get the expected result.

Here follows the solution of a Markov Renewal Equation (MRE) for the transition function of the PDMP.

Proposition 5 ([25]). *The transition function P of (Z_t, X_t) is governed by the MRE*

$$P_{ij}(z, B, t) = g_{ij}(z, B, t) + (Q * P)_{ij}(z, B, t),$$

whose solution is $P_{ij}(z, B, t) = (\Psi * g)_{ij}(z, B, t)$ and where

$$g_{ij}(z, B, t) = \mathrm{e}^{-a_i t}\, \mathbb{1}_{\{i=j\}} \mathbb{1}_B(\phi_{z,i}(t)).$$

Proof It is convenient to make appear S_1 in (13). Hence,

$$P_{ij}(z, B, t) = \mathbb{P}_{z,i}(Z_t \in B, X_t = j, S_1 > t) + \mathbb{P}_{z,i}(Z_t \in B, X_t = j, S_1 \leq t),$$

where P_1 and P_2 denote the first and the second term, respectively, on the right part of the equation.

Before the first jump of X_t, $i = j$ and Z_t evolves according to $\phi_{z,i}(t)$. Thus,

$$P_1 = \mathrm{e}^{-\alpha_i t}\, \mathbb{1}_B(\phi_{z,i}(t)) \mathbb{1}_{\{i=j\}}.$$

From the Total Probability Theorem, it holds for P_2 that

$$P_2 = \sum_{k \in E, k \neq i} \int_0^t \mathbb{P}_{z,i}(Z_t \in B, X_t = j | J_1 = k, S_1 = s) \mathbb{P}_{z,i}(J_1 = k, S_1 \in \mathrm{d}s).$$

As long as $\mathbb{P}_{z,i}(Y_1 = k, S_1 \in \mathrm{d}s) = a_{ik} \mathrm{E}^{-a_i s} \mathrm{d}s$, and noticing that

$$\mathbb{P}_{z,i}(Z_t \in B, X_t = j | Y_1 = k, S_1 = s) = P_{kj}(\phi_{z,i}(s), B, t - s),$$

then P_2 is fully characterized. Finally,

$$P_{ij}(z, B, t) = \mathrm{e}^{-a_i t}\, \mathbb{1}_{\{i=j\}} \mathbb{1}_B(\phi_{z,i}(t)) + \sum_{k \in E, k \neq i} a_{ik} \int_0^t \mathrm{e}^{-a_i s}\, P_{kj}(\phi_{z,i}(s), B, t - s) \mathrm{d}s,$$

which may be written, with Q given by Lemma 1 and g given by Proposition 5, as

$$P_{ij}(z, B, t) = g_{ij}(z, B, t) + \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, \mathrm{d}y, \mathrm{d}s) P_{kj}(y, B, t - s).$$

This last equation is of the general form (17) whose solution is given by (18).

2.6 Practical Calculation of Semi-Markov Kernel

As already given in [47], the semi-Markov kernel $Q_{ij}(z, B, t)$ of the system can be calculated with integration of $Q_{ij}(z, B, \mathrm{d}t)$, Lemma 1.

First, in order to resolve this integral, let us give the definition of $t_{z,i}(y)$, the time that Z_t needed to reach $y > z$, without jump of X_t, with initial conditions $Z_0 = z$, $X_0 = i$:

$$t_{z,i}(y) = \inf\{t \geq 0 : \phi_{z,i}(t) \geq y\},\ y \in \mathbb{R}_+. \tag{19}$$

The function $t_{z,i}(y)$ is the generalized inverse of the function $\phi_{z,i}(t)$.

Applying Lemma 1, for all B of $\mathcal{B}_+$,

$$Q_{ij}(z, B, t) = \int_0^t Q_{ij}(z, B, \mathrm{d}t) = a_{ij} \int_0^t \mathrm{e}^{-a_i s}\, \mathbb{1}_B(\phi_{z,i}(t))\mathrm{d}s. \tag{20}$$

What is useful in this problem frame, is to calculate the kernel, at time t, for the subsets of the set B, which is the subset of the 'up' states U and the subset of the 'down' states D:

- For $B := U$ the indicator function $\mathbb{1}_B(\phi_{z,i}(t)) = 0$, when the function ϕ is out of the 'up' states subset, i.e. when $\phi_{z,i}(t)$ reaches the point Δ. By (19), this corresponds to the instant $t_{z,i}(\Delta)$ and the kernel is written as

$$\begin{aligned} Q_{ij}(z, U, t) &= a_{ij} \int_0^t \mathrm{e}^{-a_i s}\, \mathbb{1}_U(\phi_{z,i}(t))\mathrm{d}s = a_{ij} \int_0^{\min(t_{z,i}(\Delta),t)} \mathrm{e}^{-a_i s}\, \mathrm{d}s \\ &= p_{ij}\left(1 - \mathrm{e}^{-a_i \min(t_{z,i}(\Delta),t)}\right). \end{aligned} \tag{21}$$

- For $B := D$ the indicator function $\mathbb{1}_D(\phi_{z,i}(t)) = 0$, when the function ϕ has not reached the 'down' states, i.e. when $\phi_{z,i}(t) < \Delta$, so

$$Q_{ij}(z, D, t) = a_{ij} \int_0^t \mathrm{e}^{-a_i s}\, \mathbb{1}_D(\phi_{z,i}(t))\mathrm{d}s = a_{ij} \int_{t_{z,i}(\Delta)}^t \mathrm{e}^{-a_i s}\, \mathrm{d}s \tag{22}$$

$$= p_{ij}(\mathrm{e}^{-a_i t_{z,i}(\Delta)} - \mathrm{e}^{-a_i t})\mathbb{1}_{\{t > t_{z,i}(\Delta)\}}. \tag{23}$$

3 Reliability Calculus

As an application, let us now study reliability of the system. The reliability $R(t)$ of a system at time $t \in \mathbb{R}_+$, starting to operate at time $t = 0$, is defined as the probability that the system has operated without failure in the interval $[0, t]$, i.e.

$$R(t) = \mathbb{P}((Z_t, X_t) \in U \times E) = \sum_{i,j \in E} \mu(B, i) P_{ij}(z, U, t). \tag{24}$$

In the sequel, we briefly describe the previously proposed method for reliability calculus and then we present our method along with details on its practical implementation.

3.1 Previous Method on Reliability Calculus

In order to obtain a calculable form of $R(t)$, a first approach, as already described in [46] is to plug-in the solution of $P_{ij}(z, B, t)$, Proposition 5, in Eq. (24):

$$R(t) = \sum_{i,j \in E} \mu(B, i) \times (\Psi * g)_{ij}(z, B, t). \tag{25}$$

In turn, the Eq. (25) requires to calculate:

1. the convolution between the functions Ψ and g: $(\Psi * g)_{ij}(z, B, t)$, which includes:
2. the Markov renewal function Ψ,

$$\Psi_{ij}(z, B, t) = \sum_{n \geq 0} Q_{ij}^{(n)}(z, B, t)$$

which in turn requires:

3. the n-power of kernel convolution, $Q_{ij}^{(n)}$, Eq. (15), which requires:
4. the semi-Markov kernel, Q_{ij}.

3.2 A New Method on Reliability Calculus

Here we present a second approach for the demanding calculation of reliability. Its advantage is its lower calculation complexity, since the recursive calculation is directly applied on reliability itself. To make the idea more comprehensible, we will first refer to a case where the process of interest is $(X_t)_{t \geq 0}$, and later pass to our case of the coupled process.

The novelty to calculate reliability recursively with recurrence on itself exists in [52], for semi-Markov processes. Following the notation of our problem setting, let $(X_t)_{t \geq 0}$ be the semi-Markov process that describes the evolution in time of the studied system. Let us assume that the system starts to work at time $t = 0$ and the event $\{X_t = i\}$, $i \in E$ means that the system is in the operating mode i at time t. The reliability of the system at time t is defined as the probability that the system has been functioning without failure in the interval $[0, t]$, that is

$$R(t) = \mathbb{P}(X_s \in E, \forall s \in [0,t])$$

If the conditional reliability $R_i(t)$, $i \in E$, at time t is defined as

$$R_i(t) = \mathbb{P}(X_s \in E, \forall s \in [0,t] | X_0 = i), \quad i \in E, \tag{26}$$

then for any initial distribution α, it is $R(t) = \sum_{i \in E} \alpha(i) R_i(t)$. The conditional reliability satisfies the following Markov Renewal Equation (MRE)

$$R_i(t) = 1 - H_i(t) + \sum_{j \in U} \int_0^t Q_{ij}(ds) R_j(t-s). \tag{27}$$

Equation (27) can be approximated by

$$R_i(t) \approx 1 - H_i(t) + \sum_{j \in U} \sum_{l=1}^{k} (Q_{ij}(t_l) - Q_{ij}(t_{l-1})) R_j(t - t_l), \tag{28}$$

where $0 = t_0 < t_1 < \ldots < t_k = t$, which can be used recursively in order to obtain $R_i(t)$, starting from an initial point $R_i(0) = 1$, [7]. Similar MRE for the conditional reliability is encountered in [53].

Comparative results are encountered in [54], on the reliability function of an object with the failure rate modeled by a semi-Markov process defined on an at most countable state space. The solution of the conditional reliability function is proved to be unique in the class of the measurable and uniformly bounded failure rate functions.

Let $\{w(t) : t \geq 0\}$ denote a random load process defined on at most countable state space W, with a nonnegative and right continuous trajectories. Suppose that the failure rate of an object, denoted by $\{\lambda(t) : t \geq 0\}$ depends on the random load process $\lambda(t) = g(w(t))$, where $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a monotonic and continuous function. It is obvious that $\{\lambda(t) : t \geq 0\}$ is a random process described on at most countable state space $\Lambda = g(W)$ with a nonnegative, right continuous trajectories. The reliability of a component with random failure rate is defined as follows

$$R(t) = \mathbb{E}[\mathrm{e}^{-\int_0^t \lambda(x) dx}] \tag{29}$$

and has all the properties of a classical reliability function.

Suppose that random load $\{w(t) : t \geq 0\}$ is a semi-Markov process defined on a discrete set $W = \{w_j : j \in E\}$, where $E = \{1, 2, \ldots\}$ or $E = \{1, 2, \ldots, m\}$ and $w_i \in \mathbb{R}$, $0 \leq w_1 < w_2 < \ldots$ Assume that this process is defined by an initial distribution

$$\alpha = \left\{ \mathbb{P}(w(0) = w_i) : i \in E \right\} \tag{30}$$

and a kernel

$$\begin{aligned} Q(t) &= [Q_{ij}(t) : i, j \in E], \\ Q_{ij}(t) &= \mathbb{P}(S_{n+1} - S_n \le t, w(S_{n+1}) = w_j | w(S_n) = w_i). \end{aligned} \tag{31}$$

For the monotonic function $g, \{\lambda(t) = g(w(t)) : t \ge 0\}$ is the semi-Markov process on $\Lambda = g(W)$ with the same kernel. Let g be an increasing function. Then the state space Λ consist of the real numbers $\lambda_0, \lambda_1, \lambda_2, \ldots$ such that $0 \le \lambda_0 < \lambda_1 < \lambda_2 < \ldots$ Grabski [54] defined a conditional reliability function as

$$R_i(t) = \mathbb{E}[\mathrm{e}^{-\int_0^t \lambda(x)dx} | \lambda(0) = \lambda_i]. \tag{32}$$

It is obvious, then, that

$$R(t) = \sum_{j \in E} P\{\lambda(0) = \lambda_i\} R_i(t). \tag{33}$$

The following theorem is an extension of the result proved by [55].

Theorem 2 ([54]). *If the failure rate function $\{\Lambda(t) : t \ge 0\}$ is a regular semi-Markov process on discrete state space with a kernel $Q(t) = [Q_{ij} : i, j \in E]$, then the conditional reliability functions $R_i(t), i \in E$, satisfy the equation*

$$R_i(t) = \mathrm{e}^{-\lambda_i t}(1 - H_i(t)) + \int_0^t \mathrm{e}^{-\lambda_i x} \sum_{j \in E} R_j(t - x) \mathrm{d}Q_{ij}(x), \quad i \in E. \tag{34}$$

The solution is unique in the class of the measurable and uniformly bounded functions space.

Theorem 1 of [55] introduced the reliability function assuming that the failure rate of an element is a special semi-Markov process with a finite state space or a piecewise Markov process with finite state space.

Turning back to the reliability of the coupled process now (Z_t, X_t) that we are interested in, we should work equivalently to the previously cited authors, that is applying the recursion on the conditional reliability. Thus, we start by defining conditional reliability as:

$$R_i(z, U, t) := \mathbb{P}((Z_t, X_t) \in U \times E | Z_0 = z, J_0 = i).$$

This can be written with the following form

$$R_i(z, B, t) = \sum_{j \in E} P_{ij}(z, B, t), \tag{35}$$

that we plug into (24), in order to obtain reliability:

$$R(t) = \sum_{i \in E} \mu(B, i) R_i(z, B, t). \tag{36}$$

So, we focus on the calculation of conditional reliability:

Following the solution of $P_{ij}(z, B, t)$ as given in Proposition 5, that is

$$P_{ij}(z, B, t) = g_{ij}(z, B, t) + \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, \mathrm{d}y, \mathrm{d}s) P_{kj}(y, B, t-s),$$

we apply the (35) on it, thus

$$R_i(z, B, t) = \sum_{j \in E} g_{ij}(z, B, t) + \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, \mathrm{d}y, \mathrm{d}s) R_k(y, B, t-s), \tag{37}$$

or

$$R_i(z, B, t) = (1 - H_i(t)) \mathbb{1}_B(\phi_{z,i}(t)) + \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, \mathrm{d}y, \mathrm{d}s) R_k(y, B, t-s), \tag{38}$$

and if (Z_t, X_t) is in particular Markov process

$$R_i(z, B, t) = e^{-a_i t} \mathbf{1}_B(\phi_{z,i}(t)) + \sum_{k \in E \setminus \{i\}} a_{ik} \int_0^t e^{-a_i s} R_k(\phi_{z,i}(s), B, t-s) ds. \tag{39}$$

From Theorem 1, it becomes obvious that $R_i(z, B, t)$ is the unique solution of the above Markov Renewal equations, (38) and (39).

In Eqs. (38) and (39) it becomes evident that the conditional reliability $R_i(z, B, t)$ is calculated recursively by itself. In terms of economizing on calculation time compared to previous method, we deviate steps 1–3, which demand convolution two times (step 1 and step 3) and pass through Markov Renewal function. In essence, from the previous method, we apply only step 4, that is we calculate the semi-Markov kernel, Q_{ij} and also we calculate recursively the conditional reliability function by (38)/(39), so convolution appears only once. Last, reliability is calculated simply by (36). A comparison between the two methods for reliability calculus can be seen in Fig. 1. It is clear that the second diagram describes a shorter and less complex algorithm. The complexity increases in both cases of algorithm as the size of state space E increases, which makes the second calculus method even more preferable. To note, that a possible drawback of the method we propose, as one may conclude also from Fig. 1, is that it cannot serve for the calculation of the transition probability function, P_{ij}, which was an intermediate outcome of the calculus method of [46]. However, when reliability is the outcome one wishes to obtain, which is admittedly most often the case, the second method is recommended.

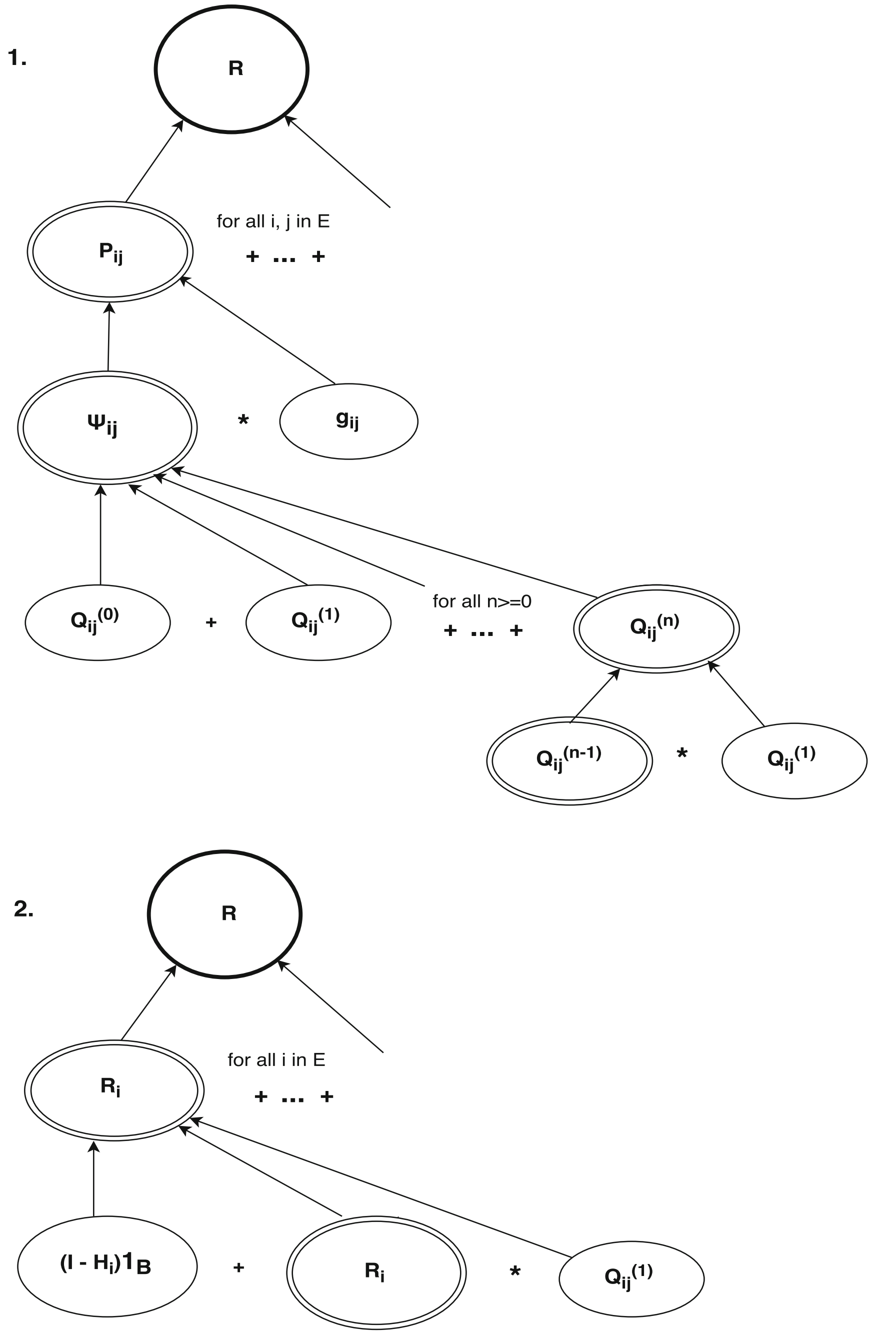


Fig. 1 Comparative diagrams for the two methods of Reliability Calculus. *Diagram 1* represents the method of [46]. *Diagram 2* represents the method proposed in present study, as described in Sect. 3.2. Calculations are given in matrix form. *Double ellipse* is implemented for intermediate calculations and *plain ellipse* for the terminal calculations.

3.3 Practical Implementation

Numerically, we will discretize conditional reliability function in order to calculate it. A discretized function shall have # as an exponent. The discretization will be applied at the same time on the intervals $B = [z_0, \Delta)$ and $[0, t]$. These two numerical partitions will be:

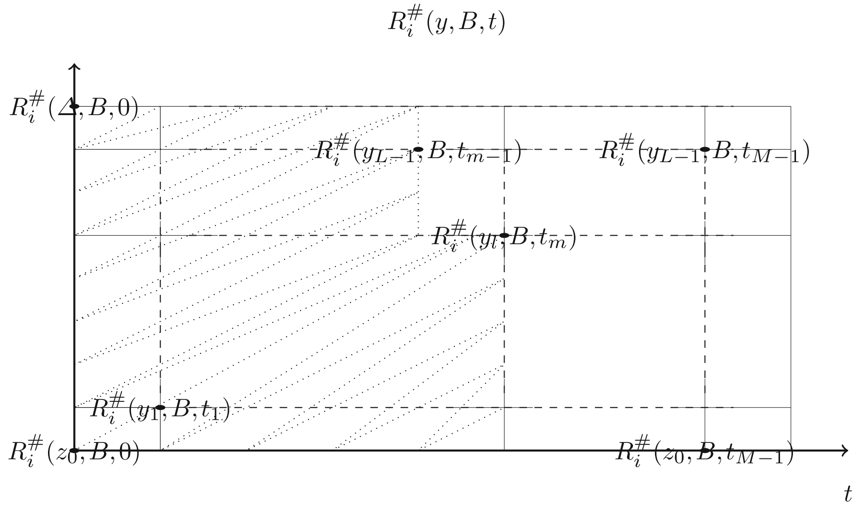
$$I_y = \{z_0 = y_0 < y_1 < \cdots < y_l < \cdots < y_L = \Delta\},$$
$$I_t = \{0 = t_0 < t_1 < \cdots < t_m < \cdots < t_M = t\}.$$

The steps of discretization of I_t and I_y are considered regular: for $l = 0, \ldots, L-2$, $m = 0, \ldots, M-2$,

$$y_{l+1} - y_l = y_{l+2} - y_{l+1} = \Delta y, \quad t_{m+1} - t_m = t_{m+2} - t_{m+1} = \Delta t.$$

Effectively, the variables L, M represent the number of discretization steps for $[z_0, \Delta), [0, t]$.

In order to make the discretization method used here more perceivable, we give the following graphical representation. It depicts the discretized conditional reliability $R_i^{\#}(y_l, B, t_m)$ with the numerical partitions of I_y and I_t on the vertical and horizontal axis respectively. When $R_i^{\#}(y_l, B, t_m)$ is about to be calculated, the algorithm has at his disposal all values for conditional reliability, $\forall k \in E, k \neq i$, such that $t < t_m, \forall y$ and for $t = t_m$, for $y < y_l$, as represented by the shadowed area, though not depicted $\forall k \in E$:



We provide the discretized form of conditional reliability function and its implementation algorithm for both the case when the coupled process is a Markov process (Eq. (40), Algorithm I) and the case when it is a semi-Markov process (Eq. (41), Algorithm II).

So, when (Z_t, X_t) is in particular Markov process, the function (39) gets the form

$$R_i^{\#}(y_l, B, t_m) = \mathrm{e}^{-a_i t_m}\, \mathbb{1}_B(\phi_{y_l,i}(t_m)) + \sum_{s\in[t_1,t_m]} \sum_{k\in E, k\neq i} p_{ik} R_k^{\#}(\phi_{z,i}(s), B, t_m - s)(\mathrm{e}^{-a_i(s-\Delta t)} - \mathrm{e}^{-a_i s}) \tag{40}$$

and the algorithm for its calculation is:

Algorithm I: Reliability Calculation for Markov Process

```
» Input: $z, A, \Delta, \alpha, L, M$
$I_y := \{z = y_0 < y_1 < \cdots < y_l < \cdots < y_L = \Delta\}$
$I_t := \{0 = t_0 < t_1 < \cdots < t_m < \cdots < t_M = t\}$

» Calculation of conditional reliability
for $t_m \in I_t, y_l \in I_y, i \in E$ do:
  if $\phi_{y_l,i}(t_m) < \Delta$ do
    $R_i^{\#}(y_l, B, t_m) := \mathrm{e}^{-a_i t_m}$
  end if
  for $s \in [t_1, t_m]$
    if $\phi_{y_l,i}(s) < \Delta$ do
      for $k \in E$ do:
        if $k \neq i$ do:
          $R_i^{\#}(y_l, B, t_m) := R_i^{\#}(y_l, B, t_m) + p_{ik}(\mathrm{e}^{-a_i(s-\Delta t)} - \mathrm{e}^{-a_i s})R_k^{\#}(\phi_{y_l,i}(s),$
          $B, t_m - s)$
        end if
      end for
    end if
  end for
end for

» Calculation of reliability
for $t_m \in I_t$ do:
  $R(t_m) := \sum_{i\in E} \mu(B, i) R_i^{\#}(y_l, B, t_m)$
end for
```

Respectively, when (Z_t, X_t) is a semi-Markov process, the discretized form of function (38) will be:

$$R_i^{\#}(y_l, B, t_m) = (1 - H_i(t_m))\mathbb{1}_B(\phi_{y_l,i}(t_m)) + \sum_{s\in[t_1,t_m]} \sum_{y_n\in I_y} \sum_{k\in E} \Delta Q^{\#}(y_l, y_n, s) R_k^{\#}(y_n, B, t_m - s) \tag{41}$$

where the kernel difference is given by

$$\Delta Q^{\#}_{ij}(y_l, y_n, s) = [Q^{\#}_{ij}(y_l, y_n, s) - Q^{\#}_{ij}(y_l, y_{n-1}, s)] \\ - [Q^{\#}_{ij}(y_l, y_n, s - \Delta t) - Q^{\#}_{ij}(y_l, y_{n-1}, s - \Delta t)]. \tag{42}$$

Here in particular, the semi-Markov kernel can be written as

$$Q^{\#}_{ij}(y_l, y_n, s) = p_{ij}(1 - \mathrm{e}^{-a_i min(t_{y_l,i}(y_n),s)}) \tag{43}$$

And the algorithm for this calculation:

Algorithm II: Reliability Calculation for semi-Markov Process

```
» Input: $z, A, \Delta, \alpha, L, M$
$I_y := \{z = y_0 < y_1 < \cdots < y_l < \cdots < y_L = \Delta\}$
$I_t := \{0 = t_0 < t_1 < \cdots < t_m < \cdots < t_M = t\}$
function for the kernel difference $\Delta Q^{\#}(y_l, y_n, s)$, Equation 42

» Calculation of conditional reliability
for $t_m \in I_t, y_l \in I_y, i \in E$ do:
    if $\phi_{y_l,i}(t_m) < \Delta$ do
        $R^{\#}_i(y_l, B, t_m) := 1 - H_i(t_m)$
    end if
    for $s \in [t_1, t_m]$
        for $y_n \in I_y$ do
            for $k \in E$ do:
                $R^{\#}_i(y_l, B, t_m) := R^{\#}_i(y_l, B, t_m) + \Delta Q^{\#}(y_l, y_n, s) R^{\#}_k(y_n, B, t_m - s)$
            end for
        end for
    end for
end for

» Calculation of reliability
for $t_m \in I_t$ do:
    $R(t_m) := \sum_{i \in E} \mu(B, i) R^{\#}_i(y_l, B, t_m)$
end for
```

4 Estimation of Reliability

The aim is to obtain an estimation of reliability result in order to compare it with this of reliability calculus. The empirical estimator of reliability presented in this paragraph comes from the resolution of the Paris model function, which is

$$
\begin{cases} \frac{\mathrm{d}Z_t}{dt} = a_0 Z_t^b \times v(X_t), \\ Z_0 = z_0. \end{cases} \tag{44}
$$

The method we describe here and apply in the sequence is thoroughly presented in [47].

If we denote as $(\varsigma_t^k)_{k=1,\dots,N}$ some N simulated trajectories of process Z_t, the reliability of the system, given as

$$
R(t) = \mathbb{P}(Z_t < \Delta),
$$

can be estimated by the empirical estimator

$$
\widehat{R}(t) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{\{\varsigma_t^k < \Delta\}}. \tag{45}
$$

So as to simulate Z_t trajectories, we distinguish two cases of solution for $Z_t = \phi(t)$, one solution when $b = 1$ and another when $b \neq 1$.

In the case when $b = 1$:

$$
Z_t = \phi(t) := \begin{cases} \phi_{z,J_0}(t) = z\,\mathrm{e}^{a_0 t v(J_0)}, \quad t \in [S_0, S_1), \\ \phi_{Z_{S_1},J_1}(t) = Z_{S_1}\,\mathrm{e}^{a_0(t-S_1)v(J_1)}, \quad t \in [S_1, S_2), \\ \dots \\ \phi_{Z_{S_n},J_n}(t) = Z_{S_n}\,\mathrm{e}^{a_0(t-S_n)v(J_n)}, \quad t \in [S_n, S_{n+1}). \end{cases} \tag{46}
$$

In the case when $b \neq 1$:

$$
Z_t = \phi(t) := \begin{cases} \phi_{z,J_0}(t) = (z^{1-b} + a_0(1-b)tv(J_0))^{\frac{1}{(1-b)}}, \quad t \in [S_0, S_1), \\ \phi_{Z_{S_1},J_1}(t) = (Z_{S_1}^{1-b} + a_0(1-b)(t-S_1)v(J_1))^{\frac{1}{(1-b)}}, \quad t \in [S_1, S_2), \\ \dots \\ \phi_{Z_{S_n},J_n}(t) = (Z_{S_n}^{1-b} + a_0(1-b)(t-S_n)v(J_n))^{\frac{1}{(1-b)}}, \quad t \in [S_n, S_{n+1}). \end{cases} \tag{47}
$$

In order, however, to simulate the trajectories of Z_t, as necessary for the empirical estimator of reliability (45), the numerical cost of calculating the function ϕ in the successive jump intervals is high. As an alternative, the following will be applied:

- With integration of (44), when the parameter $b = 1$, the result is

$$
\int_0^t \frac{\mathrm{d}Z_s}{Z_s} = a_0 \int_0^t v(X_s)ds \Leftrightarrow \ln\left(\frac{Z_t}{z}\right) = a_0 \int_0^t v(X_s)ds, \tag{48}
$$

so $Z_t = z_0\,\mathrm{e}^{a_0 V_t}$, where V_t is the process of cost accumulation associated to the jump Markov process

$$V_t = \int_0^t v(X_s)ds. \tag{49}$$

The process V_t is piecewise linear and easily calculable from the values of $v(X_t)$ that has constant values between the jumps. From the above, this equivalence follows

$$\{Z_t < \Delta\} \Leftrightarrow \left\{V_t < \frac{1}{a_0}\ln\left(\frac{\Delta}{z_0}\right)\right\}. \tag{50}$$

- Respectively, when $b \neq 1$, integration on (44) gives

$$\int_0^t \frac{\mathrm{d}Z_s}{Z_s^b} = a_0 \int_0^t v(X_s)ds = a_0 V_t, \tag{51}$$

thus $Z_t = (z^{(1-b)} + a_0(1-b)V_t)^{1/(1-b)}$. From the above, equivalent events for the system function on the working "up" states will be

$$\{Z_t < \Delta\} \Leftrightarrow \left\{V_t < \frac{\Delta^{1-b} - z_0^{1-b}}{a_0(1-b)}\right\}. \tag{52}$$

This equivalence makes in practice possible to work with the process V_t instead of the process Z_t. The failure time can be determined accurately if we consider that system failure takes place when the process Z_t crosses the threshold Δ between the instants S_n and S_{n+1} or when V_t crosses its threshold in the same time interval. By linear interpolation between the values of V_t at instants S_n and S_{n+1}, we obtain the exact value of time failure τ.

So, for $t \in [S_n, S_{n+1})$, the process V_t is given by the linear function

$$V_t = V_{S_n} + \frac{V_{S_{n+1}} - V_{S_n}}{S_{n+1} - S_n}(t - S_n).$$

At time τ, it is $V_\tau = \frac{1}{a_0}\ln(\Delta/z_0)$, if $b = 1$ and $V_\tau = \frac{\Delta^{1-b} - z_0^{1-b}}{a_0(1-b)}$, if $b \neq 1$, so the exact value of failure time should be:

- if $b = 1$,

$$\tau = S_n + \frac{W_n\left(\frac{1}{a_0}\ln(\Delta/z_0) - V_{S_n}\right)}{V_{S_{n+1}} - V_{S_n}} \tag{53}$$

- if $b \neq 1$,

$$\tau = S_n + \frac{W_n\left(\frac{\Delta^{1-b} - z_0^{1-b}}{a_0(1-b)} - V_{S_n}\right)}{V_{S_{n+1}} - V_{S_n}}. \tag{54}$$

It is concluded that for the estimation of reliability (45), instead of simulating N trajectories of Z_t, we gain important calculation time and precision if we calculate the failure time associated to each trajectory. This means that we need N values of time failure, $(\tau_k)_{k=1,\dots,N}$. Thus, the reliability estimator transforms into

$$\widehat{R}(t) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{\{t<\tau_k\}}. \tag{55}$$

5 Simulation Results

This part includes two numerical applications that implement the method for reliability calculus we proposed and verify its performance by comparison with reliability empirical estimator. The first application concerns a given example. This means that we dispose all the parameters of the stochastic model, the reliability of which we are interested in. The second application is conducted on an experimental data set, which means that the model is not known and has to be estimated in advance.

5.1 Simulation Results on a Given Example

We provide here an illustration from numerical implementation of the methodology proposed. We applied reliability calculus as in Eq. (40) and Algorithm I in comparison with the estimated reliability from Sect. 4 and Eq. (55).

The numerical example studies the dynamical system (44), with parameters $a_0 = 0.01, b = 1, z_0 = 1$ and $\Delta = 10$. The randomizing process X_t is a five-state Markov process with $E = \{1, 2, 3, 4, 5\}$, a matrix generator given by

$$A = \begin{pmatrix} -0.2 & 0.16 & 0 & 0.04 & 0 \\ 0.12 & -0.2 & 0.08 & 0 & 0 \\ 0.14 & 0 & -0.2 & 0 & 0.06 \\ 0 & 0.07 & 0 & -0.1 & 0.03 \\ 0 & 0 & 0.05 & 0.05 & -0.1 \end{pmatrix},$$

and initial law

$$\alpha = \begin{pmatrix} 0.25 \; 0.5 \; 0.25 \; 0 \; 0 \end{pmatrix}.$$

The function v is a one-to-one mapping of state space E' that associates to each element of E an element of E'. It serves as a “normalized” version of the process X_t and has the same generator as X_t. For our example, it is $v : \{1, 2, 3, 4, 5\} \rightarrow \{0.5, 1, 1.5, 2, 4\}$, so $E' = \{0.5, 1, 1.5, 2, 4\}$.

The reliability of the system of our numerical example gets evaluated by implementing Algorithm I, Sect. 3.3. The Markov Renewal Equation is resolved with the foreseen discretization method, where $M = L = 500$ points of discretization. With

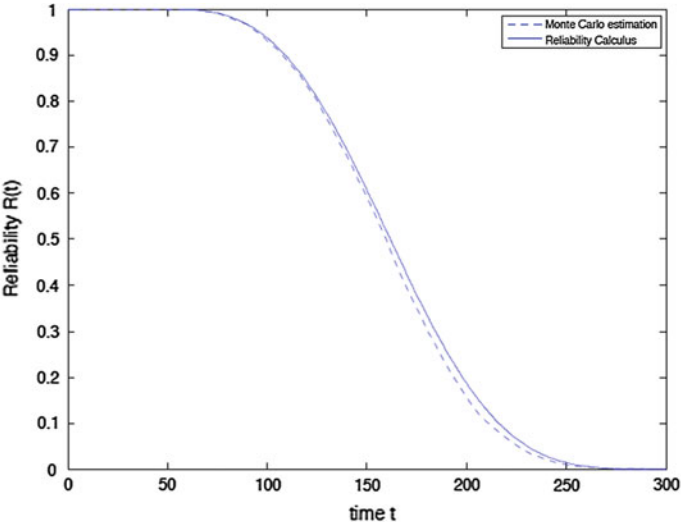


Fig. 2 Reliability function: numerical resolution of MRE compared with Monte Carlo estimate

the same input data for the model, a Monte Carlo estimation of $K = 5000$ iterations is conducted, according to Sect. 4. More precisely, K trajectories of $(Z_t^k)_{k=1,\ldots,K}$ or alternatively K values of time failure $(\tau_k)_{k=1,\ldots,K}$ are simulated and used for the estimation of $\widehat{R}(t) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{\{t<\tau_k\}}$. The outcome of Monte Carlo estimation serves as a comparison measure for the result of the MRE resolution. The result of the simultaneous implementation of the two reliability methods is depicted on Fig. 2. Evidently, the two reliability results appear very similar. We observe a slight discrepancy, though, that can be attributed to the number of discretization points, that is for larger M, L we expect a more precise curve, or to the number of Monte Carlo iterations, that is for larger K, probably the estimation curve would be even closer to the calculus curve. However, the proximity of the two curves is sufficiently close. To note, here, that similar results are obtained in [47], following, however, a different calculus procedure, Sect. 3.1. What is particular, though, in this study, is that we achieved to reduce the complexity of this previous resolution procedure and propose for the new calculus method an algorithm that is more time efficient.

5.2 *Simulation Results on Experimental Data*

At this point, we are interested in applying the methodology we proposed in Sect. 3.2 on a real data set. Before proceeding to the implementation, we shall initially refer to this data set, which we used for this application.

To begin with, the data comes from crack propagation experiments and, in particular, from fatigue tests conducted at the Aerospatiale-Matra laboratory in Suresnes

(France). This data set was utilized in [56] and gave remarkable results. In [57], the interested reader can study a fatigue crack propagation analysis, by means of dual boundary element method. Applying the Paris equation, a deterministic and a probabilistic approach are provided, giving results in good agreement with those from the experimental tests. Also, sufficient information on the experimental procedure for the tests that produced this data is provided. Here, we refer to this procedure briefly. The data originates from multiple site damage (MSD) phenomenon, that is the simultaneous occurrence of fatigue cracks in the same structural element. In fact, 6 aluminum 2024T3 specimens were exploited in the following way. Fatigue tests on a plane with 14 free holes were achieved by applying the load on the transversal direction of the plane. The tests are done under imposed load with a constant amplitude of a ration $R = 0.1$ and with a maximum stress level of 100 MPa. The evolution of the crack lengths with the number of loading cycles was measured using an optical microscope.

We should underline that the experimental procedure of Aerospatiale-Matra laboratory is not exactly our initial consideration of the cracking problem. We are more concerned about mono-cracks, without any interaction among them on the contrary to the multi-crack of MSD. In addition, our model seems rather conservative to consider $P(Z_0 = z_0)$, which is not the case of Aerospatiale-Matra data. In Fig. 3 we see the trajectories of Z_t, consisting the data set. This time the initial value of Z_t, i.e., Z_0, is not a unique value for all trajectories, so we assume that it follows a distribution, whose law is estimated with a machine learning method from experimental data. Moreover, the failure point, Δ, is not a fixed value but varies, on the contrary. For the

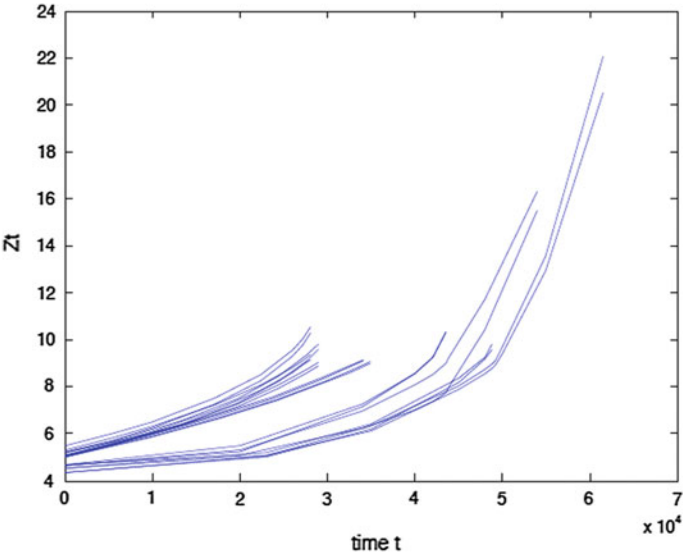


Fig. 3 Aerospatiale-Matra data set

numerical implementation, we considered the mean value of the various experimental Δ as the Δ that the two reliability methods implemented in the previous example indispensably require.

As it becomes obvious from Fig. 3 that depicts the data set, only the degradation process Z_t is now observable. This means that in order to apply the model considered here, the jump Markov process X_t, not being available, has to be estimated. The procedure for the estimation of the jump process is described in detail in Chap. 4 of [47]. Here, we shall only mention the outline of how we obtain in the end an estimation of the initial law α and the generator A of the process X_t, with Z_t as unique input data:

1. We estimate $\dot{Z}_t = \frac{\mathrm{d}Z_t}{\mathrm{d}t}$, that is $\widehat{\dot{Z}}_t$
2. Applying the following hypothesis

Hypothesis 3 For the system (1), there is a function G such that we can write X_t explicitly as a function of $\dot{Z}_t$ and Z_t, that is $X_t = G(\dot{Z}_t, Z_t)$.

We obtain an estimation of the X_t trajectories, that is $(\widetilde{X}_t^k)_{k=1,\dots,K}$.

3. We reduce the state space $\widetilde{E}$ of $\widetilde{X}_t^k$ to $\widehat{E}$ of $\widehat{X}_t^k$, by placing in the same state-group the states that are sufficiently close with the appropriate classification algorithm.
4. With the maximum likelihood method, we estimate the initial law, $\widehat{\alpha}$, and the infinitesimal generator, $\widehat{A}$. (In [47], strong convergence and asymptotic normality results are given for $\widehat{A}$.)

Disposing now the estimates of the initial law, $\widehat{\alpha}$, and the infinitesimal generator, $\widehat{A}$, for the jump process X_t, the procedure described for the previous numerical application is again followed. This means that, with the estimated model as input data instead of a given one as previously, reliability is calculated as proposed in Sect. 3.3 and estimated with Monte Carlo method according to Sect. 4. What is new, though, this time, is an empirical estimator for reliability that is derived directly from the experimental data by Eq. (55). This estimator actually has no relation with the model considered in the present study, since it just originates from the real data set. It serves, thus, as a comparison measure for the result of the MRE resolution and the MC estimator, both of which come from a model assumed to represent crack propagation and for which crack propagation data is used in order to estimate. Figure 4 gives the three reliability curves in a common axis system, after simultaneous application of the three, as mentioned here, reliability methods. And we notice good similarity results among the three curves. First, this implementation confirms the good performance of the reliability calculus method we proposed (Sect. 3.3), that, as expected after the previous implementation, is in agreement with the MC estimation result. We see that these two methods gave curves that are very close between them and this time, as a second remark, are quite close to the empirical reliability estimator that comes directly from the data. It becomes obvious, so, that this estimator validates the stochastic model considered in this study on the problem of crack propagation. In other words, the outcome of a model supposed to describe fatigue crack growth

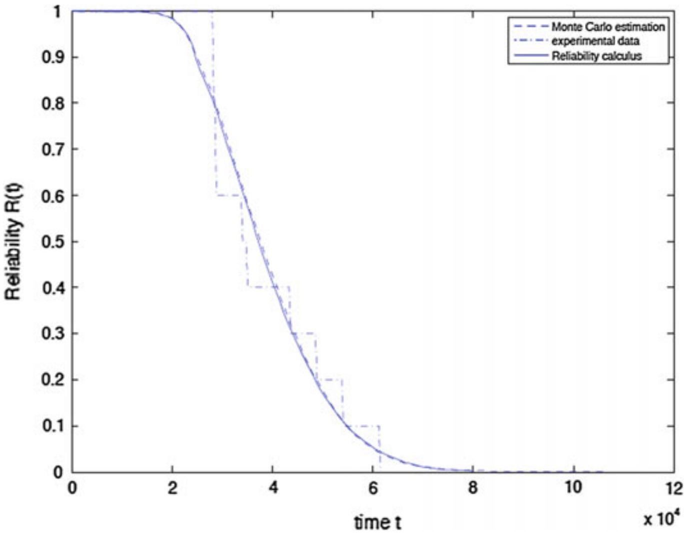


Fig. 4 Reliability function: numerical resolution of MRE compared with Monte Carlo estimate (using the estimates of the initial law and the generator of the jump process), followingly compared with reliability empirical estimator based on the experimental data

phenomenon 'agrees', indeed, with experimental data of this phenomenon. Interestingly enough, it handle s a version of crack propagation phenomenon, MSD, that is differentiated from our initial research pretext.

In an attempt to justify the discrepancy noticed between the data-originated curve and the estimated-model-originated ones, we should refer to the small size of the data set, that is only 20 trajectories were exploited in this application. Given that this data set served as the input for the estimation of the model, we understand that the model has not been precisely estimated. Besides, the fact that the distribution law of z_0 had to be estimated and that Δ had to be replaced by its mean value estimator makes clear some further loss of precision. As stated in step 3 of the procedure of estimation of the jump process, the state space $\widetilde{E}$ gets reduced to $\widehat{E}$ and this definitely entails some approximation error. Taking all the above into consideration, the proximity result of Fig. 4 is very satisfying.

6 Conclusions

In this chapter, reliability of a stochastic dynamical model that describes fatigue crack propagation was studied, by means of the Markov Renewal theory. A previously proposed method for reliability calculation was considered. We relied on the theoretical results of other authors, on the reliability of a single Markov or semi-Markov process, in order to extend their approach to the case of a coupled Markov process. To remind, here, that the stochastic dynamical model we consider is characterized by a coupled Markov. What consists also a principal accomplishment of the present study, is that we proposed a new, more time-efficient, method for reliability calculation, with a detailed description of its implementation algorithm. It would be interesting a study of computational complexity in order to explicitly compare the two calculus methods. This demands, however, some basic knowledge of theoretical computer science, that is an interdisciplinary approach.

In the sequel, we validated our theoretical results via numerical simulations. A first numerical application was conducted on a given example of stochastic model when all the model parameters were known. The implementation gave very satisfying results, with the resulting curve of our reliability calculus method almost to coincide with the curve a Monte Carlo reliability estimation gave. After this encouraging result, this calculus method and the MC estimation method were applied on a experimental data set. This time, however, the model was not given but had to be estimated from the available data. Again the two methods gave very good similarity results. An empirical estimator for reliability, calculated only with the real data, was compared with the results that the estimated model produced and was sufficiently close to them. This application could motivate future research into multiple site damage, provoked by fatigue crack growth, by appropriately adapting the stochastic model of this study.

References

1. P.C. Paris and F. Erdogan. A critical analysis of crack propagation laws. *J. Fluids Eng.*, 85:528–533, 1963.
2. D. A. Virkler, B. M. Hillberry, and P. K. Goel. The statistical nature of fatigue crack propagation. Technical report, School of Mechanical Engineering Purdue University West Lafayette, Indiana, 1978.
3. B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Hochschultext/Universitext. U.S. Government Printing Office, 2003. ISBN 9783540047582.
4. Y.K. Lin and J.N. Yang. A stochastic theory of fatigue crack propagation. *AIAA J.*, 23:117–124, 1985. ISSN 0001-1452; 1533-385X/e.
5. Kazimierz Sobczyk, editor. *Stochastic approach to fatigue: experiments, modelling and reliability estimation.* Wien: Springer-Verlag, 1993. ISBN 3-211-82452-9/pbk.
6. N.N. Bogolyubov and Mitropolskii. *Asymptotic Methods in the Theory of Non-linear Oscillations* Gordon and Breach, 1961.
7. N. Limnios and G. Oprişan. *Semi-Markov processes and reliability*. Statistics for Industry and Technology. Birkhäuser Boston, Inc., Boston, MA, 2001, reprint Springer Science and Business Media, 2012. ISBN 0-8176-4196-3.
8. R.A. Howard. *Dynamic probabilistic systems: volume i: markov models*. Series in decision and control. John Wiley And Sons, Incorporated, 1971.
9. V.S. Koroliuk and N. Limnios. *Stochastic Systems in Merging Phase Space*, World Scientific, N.Y., 2005. ISBN 9789812703125.
10. M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B*, 46(3):353–388, 1984. ISSN 0035-9246. With discussion.
11. M. H. A. Davis. *Markov models and optimization*, volume 49 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1993. ISBN 0-412-31410-X.

12. Martin Jacobsen. *Point process theory and applications*. Probability and its Applications. Birkhäuser Boston, Inc., Boston, MA, 2006. ISBN 978-0-8176-4215-0; 0-8176-4215-3. Marked point and piecewise deterministic processes.
13. B.Ã. de Saporta, F.Ã. Dufour, and H. Zhang. *Numerical Methods for Simulation and Optimization of Piecewise Deterministic Markov Processes*. Wiley, 2015. ISBN 9781848218390.
14. M. A. H. Dempster. Optimal control of piecewise deterministic Markov processes. In *Applied stochastic analysis (London, 1989)*, volume 5 of *Stochastics Monogr.*, pages 303–325. Gordon and Breach, New York, 1991.
15. M. A. H. Dempster and J. J. Ye. Necessary and sufficient optimality conditions for control of piecewise deterministic Markov processes. *Stochastics Stochastics Rep.*, 40(3–4):125–145, 1992. ISSN 1045-1129.
16. D. Vermes. Optimal control of piecewise deterministic Markov process. *Stochastics*, 14(3):165–207, 1985. ISSN 0090-9491.
17. U. S. Gugerli. Optimal stopping of a piecewise-deterministic Markov process. *Stochastics*, 19(4):221–236, 1986. ISSN 0090-9491.
18. Oswaldo Luiz do Valle Costa and François Dufour. *Continuous average control of piecewise deterministic Markov processes*. Springer Briefs in Mathematics. Springer, New York, 2013. ISBN 978-1-4614-6982-7; 978-1-4614-6983-4.
19. O. L. V. Costa and F. Dufour. Stability and ergodicity of piecewise deterministic Markov processes. *SIAM J. Control Optim.*, 47(2):1053–1077, 2008. ISSN 0363-0129.
20. R. Azais. *Estimation Non Paramétrique pour les Processus Markoviens Déterministes par Morceaux*. PhD thesis, Université Bordeaux 1, 2013.
21. Martin G. Riedler. Almost sure convergence of numerical approximations for piecewise deterministic Markov processes. *J. Comput. Appl. Math.*, 239:50–71, 2013. ISSN 0377-0427.
22. Ronald Pyke. Markov renewal processes: definitions and preliminary properties. *The Annals of Mathematical Statistics*, pages 1231–1242, 1961.
23. Erhan Cinlar. Markov renewal theory. *Advances in Applied Probability*, 1(2):123–187, 1969.
24. V. M. Shurenkov. On the theory of markov renewal. *Theory Probab.Appl.*, 29(2):247–265, 1984.
25. J. Chiquet, N. Limnios and M. Eid. Piecewise deterministic markov processes applied to fatigue crack growth modelling. *Journal of Statistical Planning and Inference*, 139(5):1657–1667, 2009. ISSN 0378-3758. Special Issue on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing.
26. Julien Chiquet and Nikolaos Limnios. Dynamical systems with semi-Markovian perturbations and their use in structural reliability. In *Stochastic reliability and maintenance modeling*, volume 9 of *Springer Ser. Reliab. Eng.*, pages 191–218. Springer, London, 2013.
27. R.E. Barlow and F. Proschan. *Statistical theory of reliability and life testing: probability models*. International series in decision processes. Holt, Rinehart and Winston, 1975. a. ISBN 9780030858536.
28. Richard E. Barlow and Frank Proschan. Mathematical theory of reliability, b. Wiley, N.Y., 1965.
29. B.V. Gnedenko, Y.K. Belyayev, and A.D. Solovyev. *Mathematical Methods of Reliability Theory*. Academic Press, 1969, N.Y. (translated from Russian,1965).
30. B. Gnedenko, I.V. Pavlov, I.A. Ushakov, I.A. Ushakov, and S. Chakravarty. *Statistical Reliability Engineering*. A Wiley Interscience publication, 1999. a. ISBN 9780471123569.
31. B. Gnedenko, I.A. Ushakov, and J. Falk. *Probabilistic Reliability Engineering*. Wiley-Interscience publication, 1995. b. ISBN 9780471305026.
32. B.V. Gnedenko, Y.K. Belyayev, A.D. Solovyev, Z.W. Birnbaum, and E. Lukacs. *Mathematical Methods of Reliability Theory*. Probability and mathematical statistics. c. ISBN 9781483263519. Elsevier Science, 2014.
33. T. Bedford. *Advances in Mathematical Modeling for Reliability*. ISBN 9781586038656. IOS Press, 2008.
34. T. Nakagawa. *Stochastic Processes: with Applications to Reliability Theory*. Springer Series in Reliability Engineering, 2011. ISBN 9780857292742.

35. T. Aven and U. Jensen. *Stochastic Models in Reliability*. Applications of mathematics. Springer, 1999. ISBN 9780387986333.
36. A. Birolini. *On the Use of Stochastic Processes in Modeling Reliability Problems*. Lecture Notes in Economics and Mathematical Systems. Springer Berlin Heidelberg, 2012. ISBN 9783642465536.
37. A. Birolini. *Reliability engineering : Theory and practice (7th ed.)*. Springer Berlin Heidelberg, 2014. ISBN 9783642395345.
38. J. Janssen and R. Manca. *Semi-Markov Risk Models for Finance, Insurance and Reliability*. Springer US, 2007. ISBN 9780387707303.
39. Nikolaos Limnios and Gheorghe Oprisan. *Semi-Markov processes and reliability*. Springer Science & Business Media, 2012.
40. Nikolaos Limnios. Reliability measures of semi-Markov systems with general state space. *Methodol. Comput. Appl. Probab.*, 14(4):895–917, 2012. ISSN 1387-5841.
41. Huilong Zhang, K. Gonzalez, François Dufour, and Yves Dutuit. Piecewise deterministic Markov processes and dynamic reliability. Proceedings of the Institution of Mechanical Engineers Part O Journal of Risk and Reliability, 222(04): 545–551, 2008.
42. Sophie Mercier and Michel Roussignol. Sensitivity estimates in dynamic reliability. In *Advances in mathematical modeling for reliability*, pages 208–216. IOS, Amsterdam, 2008.
43. G. Becker, L. Camarinopoulos, and D. Kabranis. Dynamic reliability under random shocks. *Reliability Engineering and System Safety*, 77(3):239–251, 2002.
44. C. Cocozza-Thivent, R. Eymard, and S. Mercier. A finite-volume scheme for dynamic reliability models. *IMA Journal of Numerical Analysis*, 26(3):446–471, 2006a.
45. C. Cocozza-Thivent, R. Eymard, S. Mercier, and M. Roussignol. Characterization of the marginal distributions of markov processes used in dynamic reliability. *Journal of Applied Mathematics and Stochastic Analysis*, 2006, 2006b.
46. Julien Chiquet and Nikolaos Limnios. A method to compute the transition function of a piecewise deterministic Markov process with application to reliability. *Statist. Probab. Lett.*, 78(12):1397–1403, 2008. ISSN 0167-7152.
47. J. Chiquet. *Modélisation et estimation des processus de dégradation avec application en fiabilité des structures*. PhD thesis, Université de Technologie de Compiègne, 2007.
48. R. Bellman. *Stability Theory of Differential Equations*. Dover Books on Mathematics, 2013. ISBN 9780486150130.
49. M.W. Hirsch, S. Smale, and R.L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Elsevier Science, 2012. ISBN 9780123820112.
50. J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Dover Publications, 2014. ISBN 9780486781488.
51. Bernard Lapeyre, Étienne Pardoux, and Rémi Sentis. *Méthodes de Monte-Carlo pour les équations de transport et de diffusion*, volume 29 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1998. ISBN 3-540-63393-6.
52. Sonia Malefaki, Nikolaos Limnios, and Pierre Dersin. Reliability of maintained systems under a semi-markov setting. *Reliability Engineering & System Safety*, 131:282–290, 2014.
53. Sophie Mercier and Michel Roussignol. Asymptotic failure rate of a markov deteriorating system with preventive maintenance. *J. Appl. Probab.*, 40(1):1–19, 03 2003.
54. Franciszek Grabski. The reliability of an object with semi-Markov failure rate. *Appl. Math. Comput.*, 135(1):1–16, 2003. ISSN 0096-3003.
55. I. Kopocińska and B. Kopociński. On system reliability under random load of elements. *Zastos. Mat.*, 17(1):5–14, 1980/81. ISSN 0044-1899.
56. H. Kebir. *Approches déterministe et probabiliste de la prévision de la durée de vie de structures aéronautiques à l' aide de la méthode des équations intégrales duales*. PhD thesis, Université de Technologie de Compiègne, 2011.
57. H. Kebir, J. M. Roelandt and J. Gaudin. Monte-carlo simulations of life expectancy using the dual boundary element method. *Engineering Fracture Mechanics*, 68(12):1371–1384, 2001. ISSN 0013-7944.
58. V. V. Anisimov. Switching Processes in Queueing Models. Wiley, 2010.

Multi-scale Simulation of Newtonian and Non-Newtonian Multi-phase Flows

Juan Luis Prieto

Abstract This work is devoted to the multi-scale simulation of Newtonian and non-Newtonian multi-phase flows using a level-set method to capture the fluid interface along with Brownian Dynamics simulations to account for the viscoelastic effects of the fluid. The Navier-Stokes equations are solved by a second order accurate semi-Lagrangian scheme, evolving the level-set function along the characteristic curves of the flow. Marker particles are added to correct the shape of the free surface, using a Semi-Lagrangian Particle Level-Set method taking into account viscous and surface tension effects. Non-Newtonian flows are modeled by means of a micro-macro, multi-scale approach in which stochastic, partial differential equations are solved using a variance-reduced technique on a number of ensembles of dumbbells scattered over the domain, with the Finitely Extensible Non-linear Elastic (FENE) kinetic model. Several benchmark problems for Newtonian and non-Newtonian fluids are presented, ranging from imposed flows to complex, high density and viscosity ratio flows featuring highly viscoelastic effects. The results highlight the versatility, accuracy and robustness of the proposed technique.

1 Introduction

The field of interface capturing techniques is a-changing. Ever increasing demands for the understanding of multi-phase flows in biological sensors, crystal propagation phenomena, combustion processes, or virtual surgery, along with a rapid development of computational resources have led to a rich framework of numerical methods, each of them best suited to specific situations [8, 18, 37].

If we focus on purely Newtonian applications, several computational techniques are available [19–21, 38]. Among them, Level-Set methods, a sort of front-capturing technique, hold a privilege position: since the seminal work by Osher and Sethian [24], much effort [23, 32, 35] has been devoted to the application and improvement

J.L. Prieto (✉)
Department of Energy Engineering, E.T.S. Ingenieros Industriales,
Universidad Politécnica de Madrid, Madrid, Spain
e-mail: juanluis.prieto@upm.es

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_14

of this elegant method, capable of dealing with topological changes in a natural way, providing also, if required, geometrical magnitudes such as the normal and curvature of the interface. One conspicuous example is the Hybrid Particle Level-Set (HPLS) method [9] of Enright and collaborators, in which marker (massless) particles are added close to the interface to better preserve its shape and mass conservation properties. Further analyses in this direction were carried out in [10, 27, 36] and recently, by Bermejo and Prieto in [2].

Research on non-Newtonian , multi-phase flows has received much less attention, especially, if a Finite Element discretization is considered jointly with a level-set method: here, the work of Pillapakkam and Singh [28], Pillapakkam et al. [29] in droplet deformation and rising bubble configurations stands out. Other front-tracking and front-capturing techniques have successfully been applied to viscoelastic, multi-phase flows: thus, Bonito et al. [4], Oishi et al. [22], Pasquali and Scriven [26], Foteinopoulou and Laso [11], Adami et al. [1], or Zainali and co-workers [39], to name a few. However, the number of studies making use of micro-macro techniques such as the CONNFFESSIT (Calculation of Non-Newtonian Flow: Finite Elements and Stochastic Simulation Technique) approach is smaller still: in this regard, we can cite the works by Cormenzana et al. [7] or Grande and others [12] for free-surface, viscoelastic flows.

The purpose of this paper is to offer an overview of the capabilities of a multi-scale, micro-macro approach in which Finite Element, semi-Lagrangian, Particle-Level Set methods are used in combination with Brownian Dynamics simulations for the accurate and robust computation of Newtonian and non-Newtonian, multi-phase flows. Thus, after this brief Introduction, we present in Sect. 2 the mathematical background required to follow the subsequent explanations; then, Sect. 3 is concerned with the numerical procedures involved in the multi-scale approach, with an emphasis on the efficient implementation of stochastic computations. Later, Sect. 4 tests the method under several benchmark configurations of Newtonian and non-Newtonian flows. Finally, some conclusions and future efforts are collected in Sect. 5.

2 Mathematical Background

The simulation of multi-phase Newtonian and non-Newtonian flows with the level-set method entails the solution of the Navier-Stokes equations along with the advection of the auxiliary function that provides the location of the interface at each time step, namely, the *level-set* function. If a non-Newtonian fluid is considered, then the polymer stress should be estimated by a closed-form constitutive equation, directly solving the Fokker-Planck equation in the configuration space, or integrating the internal degrees of freedom of particles that convey the molecular information of the polymer and are the stochastic equivalent to the Fokker-Planck equations. We use here the latter approach to compute the stress tensor as a right-hand-side term in the momentum equation, thus using a multi-scale approach in which the “micro” and the “macro” scales interact.

2.1 Macro-Scale Equations

Let us then consider the Navier-Stokes (NS) equations for a Newtonian, incompressible fluid of constant density ρ and viscosity μ in a bounded domain $D \subset \mathbb{R}^2$ and in a time interval $[0, T]$. Let Γ be the boundary of D, and let Γ^s, Γ^i and Γ^o denote existing solid, inflow and outflow boundaries. Under these conditions, the NS equations read:

$$\begin{cases} \rho \dfrac{D\mathbf{v}}{Dt} + \nabla p = \mu \Delta \mathbf{v} \quad \text{in } D \times (0, T], \\ \nabla \cdot \mathbf{v} = 0 \quad \text{in } D \times (0, T]. \end{cases} \tag{1}$$

Additionally, one should impose initial

$$\mathbf{v}(\mathbf{x}, 0) = \mathbf{v_0}(\mathbf{x}) \ \forall \mathbf{x} \in D, \tag{2}$$

and boundary conditions to the flow

$$\begin{aligned} \mathbf{v}(\mathbf{x}, t) &= \mathbf{0} \quad \text{on } \Gamma^s \ \forall t, \\ -p\mathbf{n} + \mu \frac{\partial \mathbf{v}}{\partial \mathbf{n}} &= \mathbf{b}(\mathbf{x}, t) \ \text{ on } \quad \Gamma^o \ \forall t, \\ -\mathbf{n} \cdot \mathbf{v}(\mathbf{x}, t) &= a(\mathbf{x}, t) \quad \text{on } \Gamma^i \ \forall t. \end{aligned} \tag{3}$$

where $\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla$ represents the total derivative operator.

2.1.1 The Level-Set Approach

The idea is to represent the interface between two fluids as the zero level-set of the implicit function

$$\phi(\mathbf{x}(t), t) - C = 0, \tag{4}$$

where $\mathbf{x}(t)$ denotes the coordinates of the points of the flow domain, and C is a constant. Thus, $\phi(\mathbf{x}, t)$ satisfies:

$$\frac{D\phi}{Dt} = \frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi = 0, \tag{5}$$

when $\mathbf{v} = \frac{d\mathbf{x}}{dt}$. The level-set function ϕ is initially computed as the signed distance function to the interface $\Gamma_{ls} \equiv (\phi = 0)$, so that we choose $\phi < 0$ as the interior fluid (or phase), and $\phi > 0$ as the outer fluid (or phase). The variation of both density and viscosity across the interface Γ_{ls} can now be represented in terms of ϕ as:

$$\begin{aligned} \rho(\phi) &= \rho_2 + (\rho_1 - \rho_2) H(\phi), \\ \mu(\phi) &= \mu_2 + (\mu_1 - \mu_2) H(\phi), \end{aligned} \tag{6}$$

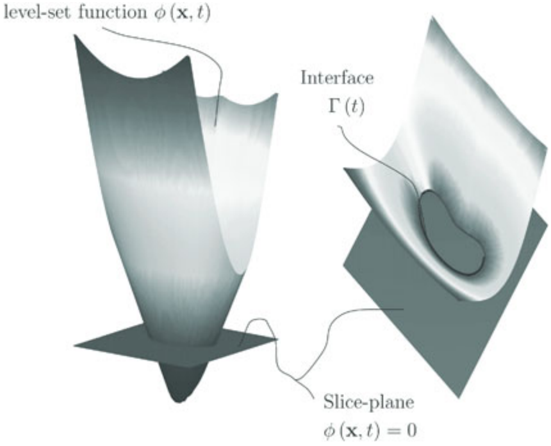


Fig. 1 Level-set function ϕ and interface Γ captured as the zero-isocontour of ϕ

where sub-indexes 1, 2 represent the outer and inner fluids to the interface Γ_{ls}; and H is the Heaviside function. Geometric magnitudes such as normal vector and curvature of the surface Γ_{ls} may be computed in terms of ϕ as:

$$\begin{aligned}\mathbf{n} &= \frac{\nabla\phi}{\|\nabla\phi\|},\\ \kappa &= -\nabla\cdot\mathbf{n},\end{aligned} \tag{7}$$

which eases the addition of surface tension effects. Here, we model the surface tension as a force defined along the interface Γ_{ls} as in [2], resulting in the term $\sigma\kappa(\phi)\,\delta(\phi)\,\nabla(\phi)$ which is to be added to the momentum equation, δ being the Dirac function and σ the surface tension coefficient (Fig. 1).

Although initially a signed distance function, ϕ usually loses this property as the numerical simulation progresses, so that $\|\nabla\phi\| \neq 1$ where $\|\cdot\|$ denotes the Euclidean norm of a vector. Under such circumstances, the isocontours of the level-set function cease to be equally spaced, geometric magnitudes are no more properly computed, and shape irregularities may arise at the interface. Hence, some *reinitialization* procedure is required to ensure that ϕ be a signed distance function involving the solution of a hyperbolic problem [2, 35]; or e.g. by means of a "direct" (geometric) reinitialization of ϕ, both options being used in this work.

Hence, the resulting NS equations for multi-phase incompressible Newtonian flows can be rewritten as:

$$\begin{cases} \dfrac{D\left(\rho\left(\phi\right)\mathbf{v}\right)}{Dt} + \nabla p - \nabla\cdot\left(\mu\left(\phi\right)\nabla\mathbf{v}\right) = \rho\left(\phi\right)\mathbf{g} + \sigma\kappa\left(\phi\right)\delta\left(\phi\right)\nabla\phi, \\ \nabla\cdot\mathbf{v} = 0, \end{cases} \tag{8}$$

in $D \times (0, T]$. System (8) along with the initial and boundary conditions Eqs. (2) and (3) the evolution of the level-set function Eq. (5), and the dependence of density and viscosity according to Eq. (6), represent the macroscopic equations to be solved at each time step.

2.1.2 Particle Level-Set Method

Proposed by Enright et al. [9] as a mixed Eulerian level-set approach making use of Lagrangian marker particles, the Hybrid Particle Level-Set (HPLS) method, has proved successful at improving mass and shape preservation in a number of benchmark problems [10, 15]. The idea is to use massless particles (*marker particles*) that following the flow, assist and correct the level-set function in under-resolved regions.

A common implementation of the HPLS method consists of three stages: *identification* of error, *quantification* of error, and *correction* of the level-set function; optionally, a *reseeding* strategy may be devised and performed after those steps. Here, we are using the Quasi-Monotone Semi-Lagrangian Particle Level-Set (QMSL-PLS) method of Bermejo and Prieto [2] for all interface-capturing computations (Fig. 2).

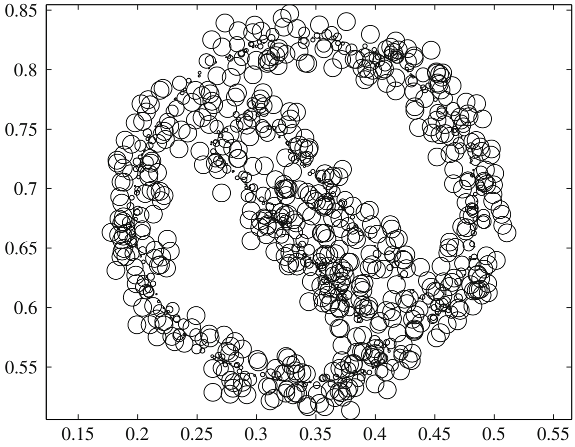


Fig. 2 Adaptive radii of marker particles for the *error quantification* stage of the QMSL-PLS method [2] in the slotted-cylinder problem

2.1.3 Dimensionless Form of the Macroscopic Equations

The dimensionless form of the macroscopic equations, choosing appropriate characteristic magnitudes, can be written as follows:

$$\left\{ \begin{array}{l} Re\rho^* \dfrac{D\mathbf{v}^*}{Dt} + \nabla p^* - \nabla \cdot \left(\mu^* (\phi) \nabla \mathbf{v}^* \right) = \dfrac{Re}{Fr^2} \rho^* (\phi) \left(-\mathbf{e}_y \right) + \dfrac{Re}{We} \kappa^* (\phi) \, \delta (\phi) \, \nabla \phi , \\ \nabla \cdot \mathbf{v}^* = 0, \end{array} \right. \tag{9}$$

in $D^* \times (0, T^*]$, along with the following initial and boundary conditions

$$\mathbf{v}^* \left(\mathbf{x}^*, 0 \right) = \mathbf{v_0}^* (\mathbf{x}) \;\; \forall \mathbf{x}^* \in D^*. \tag{10}$$

$$\begin{aligned} \mathbf{v}^* \left(\mathbf{x}^*, t^* \right) &= \mathbf{0} \text{ on } \Gamma^{s*} \; \forall t^*, \\ -p^* \mathbf{n}^* + \frac{\partial \mathbf{v}^*}{\partial \mathbf{n}^*} &= \frac{\mathbf{b} T \left(\mathbf{x}, t \right)}{\mu_1} \equiv \mathbf{b}^* \quad \text{on } \Gamma^{o*} \; \forall t^*, \\ -\mathbf{n}^* \cdot \mathbf{v}^* \left(\mathbf{x}^*, t \right) &= \frac{a T \left(\mathbf{x}, t \right)}{L} \equiv a^* \quad \text{on } \Gamma^{i*} \; \forall t^*. \end{aligned} \tag{11}$$

In Eq. (9), $\mathbf{e}_y$ represents the unitary vector in the vertical direction. The Reynolds Re, Webber We and Froude Fr dimensionless numbers are defined as:

$$Re = \frac{\rho_1 U L}{\mu_1}; \quad We = \frac{\rho_1 L U^2}{\sigma}; \quad Fr = \frac{U^2}{gL}; \tag{12}$$

with g the gravitational constant. Another relevant parameter for this work is the Capillary number Ca, which may be defined as $Ca = We/Re$.

2.2 *Micro-Scale Equations*

The idea behind the CONNFFESSIT approach and its *Brownian Dynamics* simulation technique is to envision the polymer as a Newtonian solvent in which certain entropic effects are taking place, such effects being modeled by non-interacting, massless particles (e.g. Hookean or FENE bead-spring dumbbells) that orienting themselves according to stochastic laws (equivalent to the Fokker-Planck equations in the configuration space), are influenced and do influence the macroscopic flow. Whatever the model, the dumbbells are to be evolved freely by the flow, and the position of their centers of mass $\mathbf{X}(t)$ located at any time instant; it is thus their internal degrees of freedom that further define each kinetic model.

2.2.1 Kinetic Dumbbell Model

The dumbbell version of the Rouse model with arbitrary spring elongation, i.e. the Hooke dumbbell model, is equivalent to the Oldroyd-B model of continuum mechanics, and can be represented as a number of buoyant massless dumbbells characterized by the position of their centers of mass $\mathbf{X}(t)$ and by their internal degrees of freedom $\mathbf{Q}$, the latter being defined as the vector pointing from one bead of the dumbbell to the other. As an improvement over such a model, the finite elongation introduced by the 'Finitely Extensible Non-linear Elastic' (FENE) model allows to perform more realistic computations of polymeric fluids which are known to deviate from a Gaussian behavior when stretched to their maximum. In this case, there is an elastic force for the spring connecting the dumbbells, given by

$$\mathbf{F}(\mathbf{Q}) = \frac{H\mathbf{Q}}{1 - \|\mathbf{Q}\|^2/\|\mathbf{Q}_0\|^2},$$

where $\|\mathbf{Q}_0\|^2$ denotes the maximal extensibility of the spring, and H is the spring constant. The stochastic, partial differential equation governing the internal degrees of freedom is (see Öttinger [25]):

$$d\mathbf{Q} = \left(\kappa \cdot \mathbf{Q} - \frac{1}{2\lambda}\frac{\mathbf{Q}}{1 - \|\mathbf{Q}\|^2/b}\right)dt + \sqrt{\frac{1}{\lambda}}d\mathbf{W}, \tag{13}$$

with λ the relaxation time of the polymer, $\kappa = (\nabla\mathbf{v})^T$ the transpose of the velocity gradient, $\mathbf{W}$ is a three-dimensional Wiener process, and b the parameter of maximum extensibility for the FENE model defined so that $bk_B\Theta = H\|\mathbf{Q}_0\|^2$; here, k_B is the Boltzmann constant and Θ denotes the absolute temperature. Choosing characteristic scales for length and time, the dimensionless form of Eq. (13) can be written as:

$$d\mathbf{Q} = \left(\kappa \cdot \mathbf{Q} - \frac{1}{2De}\frac{\mathbf{Q}}{1 - \|\mathbf{Q}\|^2/b}\right)dt + \frac{1}{\sqrt{De}}d\mathbf{W}, \tag{14}$$

with $De = \lambda/t_c$ the Deborah number, and t_c a characteristic time of the process. The initial conditions for $\mathbf{Q}$ in (14) are obtained from an appropriate probability density function for the FENE model [25, 30]; as for the spatial distribution, the FENE dumbbells are initially scattered over the domain in a uniform, random fashion as in [30].

2.2.2 Micro-Macro Coupling

The effect of the polymer, represented by its stress tensor τ_p, is included in the momentum equation in terms of the divergence of the polymer stress tensor which acts as an external force over the fluid, and is computed according to Kramers' expression [3]:

$$\tau_p = -nk_B\Theta\mathbf{I} + n\left\langle \mathbf{F}(\mathbf{Q}) \otimes \mathbf{Q}\right\rangle, \tag{15}$$

with n the number density of dumbbells, $\mathbf{I}$ the identity tensor, and $\langle\cdot\rangle$ the average over the configurations. Choosing the same characteristic macroscopic magnitudes as in Sects. 2.1.3 and 2.2.1, the following dimensionless momentum equation arises:

$$\begin{cases} Re\rho^* \dfrac{D\mathbf{v}^*}{Dt} + \nabla p^* - \nabla \cdot \left(\mu^* (\phi) \nabla \mathbf{v}^*\right) = \dfrac{c}{De} \nabla \cdot \tau_p + \dfrac{Re}{Fr^2} \rho^* (\phi) \left(-\mathbf{e}_y\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \dfrac{Re}{We} \kappa^* (\phi) \, \delta (\phi) \, \nabla \phi. \\ \qquad\qquad\qquad\qquad\qquad \nabla \cdot \mathbf{v}^* = 0. \end{cases} \tag{16}$$

3 Numerical Methods

The mathematical layout offered in Sect. 2 is followed now by a succinct description of the numerical methods used to accomplish the spatial and temporal discretization of the macro-scale and micro-scale models, focusing on novel techniques and efficient implementations. The interested reader is referred to [2, 5, 6, 30, 31] for a more detailed discussion on the semi-Lagrangian approach, the QMSL-PLS method, and their application to non-Newtonian fluids.

3.1 Macro-Scale Discretization

We use the method of the characteristics [2, 5] to deal with the advection terms appearing in the momentum equation (16) and in the evolution of the level-set function Eq. (5).

3.1.1 Time Discretization

Let us consider a time interval $I = (0, T]$ and N equal subdivisions of width Δt of that interval, so that $I_n = \left(t_n, t_{n+1}\right]$, with $0 \leq n \leq N - 1$. The *characteristic curves* $\mathbf{X}\left(\mathbf{x}, t; \vartheta\right)$ represent the spatial evolution of $\mathbf{X}$ with time ϑ, reaching a certain point $\mathbf{x}$ at instant t. We can make evident the dependence between the characteristics and the velocity field by

$$\frac{D\mathbf{v}\left(\mathbf{x}, t\right)}{Dt} = \frac{\partial \mathbf{v}\left(\mathbf{X}\left(\mathbf{x}, t; \vartheta\right), \vartheta\right)|_{\vartheta = t}}{\partial \vartheta} \tag{17}$$

for all $(\mathbf{x}, t)$ in $D \times (0, T)$. Thus, we can further express the characteristic curves as solution of the following system of equations:

$$\begin{cases} \dfrac{d\mathbf{X}(\mathbf{x}, t; \vartheta)}{d\vartheta} = \mathbf{v}(\mathbf{X}(\mathbf{x}, t; \vartheta), \vartheta), \\ \mathbf{X}(\mathbf{x}, t; t) = \mathbf{x}. \end{cases} \tag{18}$$

Now, we shall use a second order Backward Difference Formula (BDF2) to re-write Eq. (17) at time t_{n+1} in terms of the characteristics $\mathbf{X}$:

$$\frac{D\mathbf{v}(\mathbf{x}, t)}{Dt}\,|_{t=t_{n+1}} = \frac{3\mathbf{v}^{n+1} - 4\mathbf{v}^{n}\left(\mathbf{X}(\mathbf{x}, t_{n+1}; t_n)\right)}{2\Delta t} + \frac{\mathbf{v}^{n-1}\left(\mathbf{X}(\mathbf{x}, t_{n+1}; t_{n-1})\right)}{2\Delta t} + O\left(\Delta t^2\right), \tag{19}$$

with $\mathbf{X}(\mathbf{x}, t_{n+1}; t_n)$ and $\mathbf{X}(\mathbf{x}, t_{n+1}; t_{n-1})$ being the so-called *feet of the characteristic curves* (or departure points), which represent the spatial location at time t_n and t_{n-1} respectively, of a particle that reaches position $\mathbf{x}$ at time t_{n+1}. Their computation is carried out by a second order scheme [30] which makes use of a mid-point rule, a fixed point iterative scheme and a time-adaptive procedure, while employing a new search-and-locate algorithm to accurately track the positions of the feet of the characteristics over the mesh. The evaluation of the velocity field at the feet of the characteristic curves may be carried out by interpolation or projection.

Generalized Stokes problem The system of Eq. (9) adopts the following temporal discretization::

$$\begin{aligned} \frac{3Re}{2\Delta t}\rho^{n+1}\mathbf{v}^{n+1} - \nabla\cdot\left(\mu^{n+1}\nabla\mathbf{v}^{n+1}\right) + \nabla p^{n+1} &= \frac{Re}{Fr^2}\rho^{n+1}\left(-\mathbf{e}_y\right) + \frac{Re}{We}\kappa^{n+1}\delta\left(\phi^{n+1}\right)\nabla\left(\phi^{n+1}\right) \\ &\quad + \frac{c}{Pe}\nabla\cdot\boldsymbol{\vartheta}_p^{n+1} + \frac{2Re}{\Delta t}\rho^{n+1}\bar{\mathbf{v}}^{n} - \frac{Re}{2\Delta t}\rho^{n+1}\bar{\mathbf{v}}^{n-1}, \\ \nabla\cdot\mathbf{v}^{n+1} &= 0 \end{aligned} \tag{20}$$

where $\{\rho, \mu, \kappa\}^{n+1}$ denote the magnitudes dependent upon the level-set function at that time instant, $\{\rho, \mu, \kappa\}\left(\phi^{n+1}\right)$. The boundary conditions given by Eq. (11), may be rewritten now as:

$$\begin{aligned} \mathbf{v}^{n+1} &= \mathbf{0} \quad \text{on } \Gamma^{s*}\ \forall t^*, \\ -p^{n+1}\mathbf{n} + \frac{\partial\mathbf{v}^{n+1}}{\partial\mathbf{n}} &= \mathbf{b}^{n+1} \quad \text{on } \Gamma^{o*}\ \forall t^*, \\ -\mathbf{n}\cdot\mathbf{v}^{n+1}\left(\mathbf{x}^*, t\right) &= a^{n+1} \quad \text{on } \Gamma^{i*}\ \forall t^*. \end{aligned} \tag{21}$$

Level-set function Integrating Eq. (5) between instants t_n and t_{n+1} we get:

$$\int_{t_n}^{t_{n+1}} \frac{D\left[\phi\left(\mathbf{X}(\mathbf{x}, t_{n+1}; t), t\right)\right]}{Dt}\,\mathrm{d}t = 0 \quad \Rightarrow \quad \phi(\mathbf{x}, t_{n+1}) = \phi\left(\mathbf{X}^n\right), \tag{22}$$

where $\mathbf{x} = \mathbf{X}(\mathbf{x}, t_{n+1}; t_{n+1})$ and $\mathbf{X}^n \equiv \mathbf{X}(\mathbf{x}, t_{n+1}; t_n)$. Thus, we advance ϕ in time by evaluating the level-set function at the feet of the characteristic curves, making use of an identical procedure to that applied to the velocity field.

3.1.2 Finite Element Discretization

We employ $P_2 - P_1$ quadratic-linear polynomial approximations for velocity and pressure, and P_1 for the polymer stress tensor as in [30, 31]; a $P_1-\mathrm{iso}P_2$ approximation is used for the level-set function as was done in [2]. We then divide the computational domain D with boundary Γ into a uniformly regular partition D_H composed of triangular elements T_j, $1 \le j \le NE_1$, so that $D_H \cup \Gamma = \cup_{1\le j\le NE_1} T_j$. If we apply a red-green refinement to that partition, we get D_h, with number of elements $NE_2 = 4NE_1$. We then associate finite element spaces V_h and V_H with both partitions D_h, D_H, respectively, according to:

$$\begin{aligned} V_h &\equiv \left\{ v_h \in C^0\left(\bar{D}\right) : v_h\left|_{T_j} \in P_1\left(T_j\right), 1 \le j \le NE_2\right.\right\}, \\ V_H &\equiv \left\{ w_H \in C^0\left(\bar{D}\right) : w_H\left|_{T_k} \in P_2\left(T_k\right), 1 \le k \le NE_1\right.\right\}, \end{aligned} \tag{23}$$

where $P_m\,(T)$ denotes the set of polynomials of degree $\le m$ defined on the triangle T. Thus, any function $v_h \in V_h$ or $w_H \in V_H$ can be expressed by:

$$v_h\,(x) = \sum_{i=1}^{NN} V_i \psi_i\,(x)\,, \quad w_H\,(x) = \sum_{i=1}^{NN} W_i \bar{\psi}_i\,(x)\,, \tag{24}$$

where NN denotes the number of nodes of the partition, and $\{\psi_i\,(x)\}$, $\left\{\bar{\psi}_i\,(x)\right\}$ are the set of global basis functions in the spaces V_h, V_H, respectively. Note that $\psi_i\,(x)$ are piecewise linear polynomials, and $\bar{\psi}_i\,(x)$ are piecewise quadratic polynomials. Thus, the velocity is to be computed as a quadratic polynomial in V_h, whereas the pressure, and the polymer stress tensor belong to V_H; the level-set function ϕ is defined in a $P_1-\mathrm{iso}P_2$ space, that is, defined in partition V_h with linear approximation at each element $T_j \in D_h$, see Fig. 3.

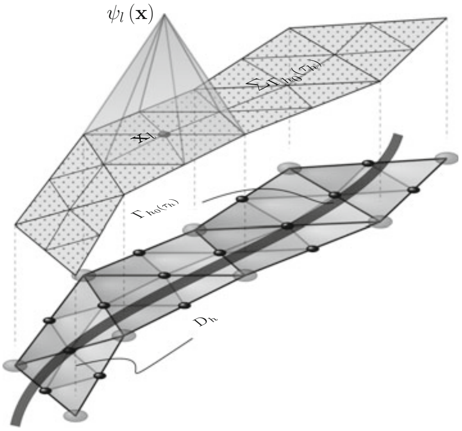


Fig. 3 Basis functions ψ_l, partition D_h and band Σ_{Γ_h} for the level-set function ϕ

3.2 Micro-Scale Discretization

We make use the Predictor-Corrector technique of [25] for the integration of the configurations of the kinetic model, as in [30, 31]; we also offer some comments about the Finite Element discretization and the variance-reduced approach.

3.2.1 Numerical Scheme for the FENE Dumbbell Model

We first apply an explicit Euler to Eq. (13) and use it as a *predictor* stage:

$$\tilde{\mathbf{Q}}^{n+1} = \mathbf{Q}^n + \left(\kappa^n \cdot \mathbf{Q}^n - \frac{1}{2\lambda}\frac{\mathbf{Q}^n}{1-\|\mathbf{Q}^n\|^2/b}\right)\Delta t + \sqrt{\frac{1}{\lambda}}\Delta\mathbf{W}^{n,n+1}. \tag{25}$$

In the second stage, we use a Crank-Nicholson scheme along with Eq. (25) to *correct* the numerical scheme and achieve (weak) second order accuracy:

$$\mathbf{Q}^{n+1}\chi = \mathbf{Q}^n + \frac{1}{2}\left(\kappa^{n+1}\cdot\tilde{\mathbf{Q}}^{n+1} + \kappa^n\cdot\mathbf{Q}^n - \frac{1}{2\lambda}\frac{\mathbf{Q}^n}{1-\|\mathbf{Q}^n\|^2/b}\right)\Delta t + \sqrt{\frac{1}{\lambda}}\Delta\mathbf{W}^{n,n+1}, \tag{26}$$

with $\chi \equiv 1 + \Delta t/[4\lambda\left(1-\|\mathbf{Q}^{n+1}\|^2/b\right)]$.

3.2.2 Finite Element Considerations

The polymer stress tensor $\tau_{p,h} \in \mathbf{R}_h$ responsible for the “micro-macro” coupling belongs to a P_1 space of linear polynomials (see [30, 31] for details), so that it can be expressed in terms of linear basis functions Ψ_k, $1 \leq k \leq MP$ as:

$$\tau_{p,h} = \sum_{k=1}^{MP} \tau_{p,k}\Psi_k, \tag{27}$$

where $\tau_{p,k}$ are the nodal values of the polymer stress tensor, and MP the number of nodes in the partition D_H.

In this work, we make use of a variance-reduced approach to the computation of the internal degrees of freedom of the dumbbells. As in the Brownian Configuration Fields (BCF) method [16] and the Lagrangian Particle Method (LPM) [13], the configurations are computed not individually for each dumbbell, but for *ensembles* of dumbbells which can be spatially correlated or not. Here, we shall use N_d dumbbells contained in each of the N_{ens} ensembles scattered over the domain, with spatially correlated Brownian processes, so that the “random-kicks” are the same for each i-th dumbbell of all the ensembles.

3.3 *Efficient Solution to the Cubic Equation in the FENE Model*

We next discuss how to efficiently compute the roots of the cubic equation arising in the FENE model. As we saw earlier, the use of the Predictor-Corrector scheme (25)–(26) proposed in [25] to integrate the configurations of a FENE fluid entails the solution of a cubic equation of the form:

$$f(x) = x^3 + Ax^2 + Bx + C = 0, \tag{28}$$

with $A = -L$, $B = -b\left(1 + \frac{\Delta t}{4De}\right)$, $C = bL$, and $L = \|\mathbf{Q}\chi\|$, see (26), to ensure that the length of the chain L remains bounded below a maximum value b. Given the large number of particles used in a typical Brownian dynamics simulation (ensembles of dumbbells in this work), along with the fact that micro-scale calculations dominate to a great extent the computational effort per iteration, this step is of utmost importance for the efficiency of the numerical scheme as a whole.

As a result, several alternatives are proposed to obtain the solution of the cubic equation: analytical solution based on trigonometric functions; iterative methods for non-linear equations; and Lookup Tables (LUTs). Presently we explore the three options, measuring the CPU time spent by each method during the computation of the configurations of a "sample run" consisting of $N_{ens} = 2.5 \times 10^5$ ensembles with $N_d = 5 \times 10^3$ dumbbells per ensemble and $b = \{20, 50, 100\}$,

3.3.1 Trigonometric Functions

The solution of the cubic equation by analytical means involve two calls to the `cos` function and another two to the `acos` function, as well as three more calls to the square-root `sqrt`. In particular, the three (real) roots of the cubic equation (28), only one of which lies in $[0, \sqrt{b}]$, are given by:

$$\left.\begin{aligned} x_1 &= \xi\sqrt{p} - \frac{A}{3} \\ x_2 &= -\tilde{\xi}\sqrt{p} - \frac{A}{3} \\ x_3 &= -x_1 - x_2 - A \end{aligned}\right\} \tag{29}$$

with $\xi \equiv 2\cos\left[\frac{\arccos(\frac{t}{2})}{3}\right]$; $\tilde{\xi} \equiv 2\cos\left[\frac{\arccos(-\frac{t}{2})}{3}\right]$; $t = \frac{q}{p^{3/2}}$; $q = -\left(\frac{2A^3-9AB+27C}{27}\right)$; $p = \frac{A^2-3B}{9}$. Though the number of CPU cycles required for each instruction is dependent on the CPU, its architecture, the memory cache, and the actual implementation of these functions in the mathematical library used in the code, the more expensive character of the trigonometric functions when compared to non-transcendental functions is in all cases observed; as it happens, for an X86 architecture, the cost of trigonometric functions can be fifty-fold that of additions or subtractions. For all the simulations performed in this paper, we have used the `math` library present in the GNU Compiler Collection (GCC) version 4.7.3.

For the "sample run" mentioned above, the time elapsed during the calculation of the configurations using the analytical approach for the cubic equation was {33.69 s, 32.46 s, 32.16 s}.

3.3.2 Iterative Methods

Next, we investigate the benefits of using an iterative scheme to retrieve the one root of the cubic equation for the FENE model lying in the interval $[0, \sqrt{b}]$ up to machine accuracy (tolerance 10^{-15}). To that end, we explore a number of numerical schemes for non-linear equations and measure the time elapsed when computing the configurations of the "sample run" aforementioned. In all cases, the understanding is that only those methods classified as "bracketing" ensure that the solution belongs to the desired interval. In view of the results collected in Table 1, the influence of the maximum extensibility parameter b is striking in the non-bracketed algorithms; accordingly, a bracketing method such as bisection, in combination with a "polishing" algorithm such as Newton-Raphson's or Halley's that uses L as initial guess, proves to be the superior option.

3.3.3 Lookup Tables

The use of Lookup Tables (LUTs) as a means to reduce the computational time spent in costly operations, such as trigonometric functions, has been widely accepted in the scientific community for decades (see e.g. [33] for a recent application of LUTs in Monte Carlo simulations). In this paper, we explore that possibility for the solution of Eq. (28), trying to build a LUT which, while offering a clear advantage time-wise, does not forgo the high-accuracy of the methods previously discussed. In Table 1 we collect the computational times of the "sample run" for different LUT sizes, in double precision. As pointed out in [34], the almost linear dependency of Eq. (28) with L makes it suitable for linear interpolation between two adjacent values, producing L^2-norm errors of $\mathscr{O}(10^{-14})$ when the size of the LUT goes beyond 7.5×10^5 (contrarily, the error was $\mathscr{O}(10^{-10})$ for an LUT size of 10^4).

Two strategies are key to the efficient implementation of an LUT in this context: memory-caching and index-mapping.

- Cache misses have a heavy impact on performance, since in that case the data must be searched in the higher-latency, main memory (RAM); hence, we propose an LUT size that fits the L3-cache of the i7-3770K CPU used in the simulation, lest the random lookups to the table (due to the underlying Brownian dynamics) further deteriorate the performance at subsequent time steps. This behavior can be observed in Table 1 for an LUT size of 10^7 doubles that do not fit the 8 MB L3-cache available in the CPU.

Table 1 Times in seconds to solve the cubic equation of a FENE fluid, $b = \{20, 50, 100\}$, with $N_{ens} = 2.5 \times 10^5$ and $N_d = 5 \times 10^3$, using analytical, iterative and LUTs techniques

Scheme	$b = 20$	$b = 50$	$b = 100$
Analytical	33.69	32.46	32.16
Newton, 2nd-order	24.73	14.20	13.76
Schröder, 3rd-order	27.49	18.09	18.08
Halley, 3rd-order	21.41	16.46	16.48
Newton modified, 4th-order	24.20	16.08	16.13
Newton modified, 5th-order	29.84	20.52	20.51
Brent, 2nd-order	50.23	49.20	48.20
Secant, 3/2-order	23.39	21.60	21.61
Newton 2nd-order+Bisection	19.58	18.41	18.09
Halley 3rd-order+Bisection	19.71	19.72	19.76
LUT size = 10^4 components	8.94	8.91	8.90
LUT size = 10^5 components	9.22	9.12	9.09
LUT size = 5×10^5 components	9.73	9.64	9.53
LUT size = 7.5×10^5 components	9.81	9.71	9.65
LUT size = 8.5×10^5 components	9.79	9.81	9.93
LUT size = 10^6 components	9.80	9.79	9.71
LUT size = 10^7components*	10.45	10.33	10.26

* This LUT size does not fit the L3-cache of the CPU. The value shown in the table corresponds to the first time step; however, after 125 time steps, the same computation takes 16.87 s ($b = 20$)

- Index-mapping avoids another possible bottleneck, namely, the search for the correct index in the table; thus, we identify, in a one-to-one fashion, each i-th entry $f(x_i)$ of the LUT with the corresponding value x_i of the independent variable, i.e. $x_i = L_i = (i+1)\frac{x_{max}}{i_{max}}$, where x_{max} is the maximum value considered for the independent variable, and i_{max} the size of the LUT table. As a consequence, the LUT behaves effectively as a "hash table", so that a searching algorithm, even an efficient one such as a binary search is no longer needed; further, the LUT size does not affect the performance of the procedure (within the L3 cache limits).

To validate this approach, we carried out additional tests with the purely extensional and purely shear flows of [14], for $b = \{20, 50, 100\}$. All results with the LUT approach overlap (within error bars of the simulation) those provided by the analytical or iterative methods. As a result, we choose an LUT size if 8.5×10^5 elements for the complex flow simulations performed in this paper, attaining an speed-up of roughly $\approx$3.4 compared with the analytical approach.

4 Results for Imposed and Complex Flows

In this section, we explore the capabilities of the method under some benchmark configurations, considering multi-phase flows of Newtonian and non-Newtonian fluids.

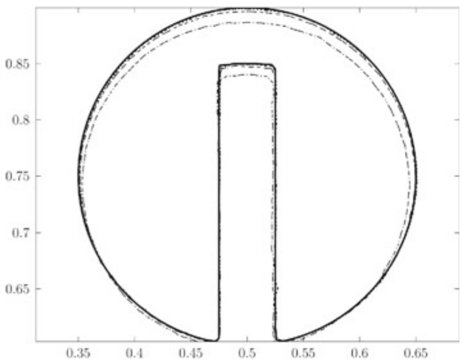


Fig. 4 Zalesak's slotted cylinder after one revolution in mesh M_1 with $N^{mk} = 1.5 \times 10^4$ marker particles and $\Delta t = 10^{-2}$ (*dashed-dotted*), $\Delta t = 2.5 \times 10^{-3}$ (*dashed*), and $\Delta t = 6.25 \times 10^{-3}$ (*dotted*). Initial solution is in *solid black*

4.1 Newtonian Fluids

The QMSL-PLS approach of [2] is used to tackle the multi-phase problems when Newtonian fluids are considered.

4.1.1 Imposed Flows

First, we ensure the correct behavior of the method under imposed, Newtonian flows. For this purpose, we consider two thoroughly revisited tests: the Zalesak's slotted cylinder and the single-vortex flow. The former is used for measuring diffusion effects in our front-capturing technique; whereas the latter is a high-vorticity flow originally proposed to test the ability of a method to deal with thin filaments of the order of the mesh resolution.

We start with Zalesak's problem. Three unstructured meshes were considered in the experiments (Table 2), though the first mesh M_1 provided enough accuracy when using our particle level-set method with a decreasing time step (Fig. 4); the influence of the number of marker particles proved rather negligible in this case, obtaining quite the same results with $N^{mk} = 1.5 \times 10^3 - 1.5 \times 10^5$.

Table 2 Number of elements T, pressure nodes MP and velocity nodes MV for mesh M_1, M_2, M_3 considered in the Zalesak and single vortex problems

Mesh	*NE*	*MP*	*MV*
M_1	5248	2705	10,657
M_2	20,992	10,657	42,305
M_3	34,030	14,216	68,461

The performance of the method in stretching filaments can be observed in Fig. 5, where we show the evolution of the single vortex problem at times $T = 1, 3, 5$ using mesh M_2 along with the velocity field given by:

$$u = -\sin^2(\pi x)\sin(2\pi y),$$
$$v = \sin^2(\pi y)^2\sin(2\pi x), \tag{30}$$

where the periodicity usually considered in this kind of flow has been removed so as to observe maximum stretching. We find the comparison with the literature [9, 10, 15, 27] quite satisfactory.

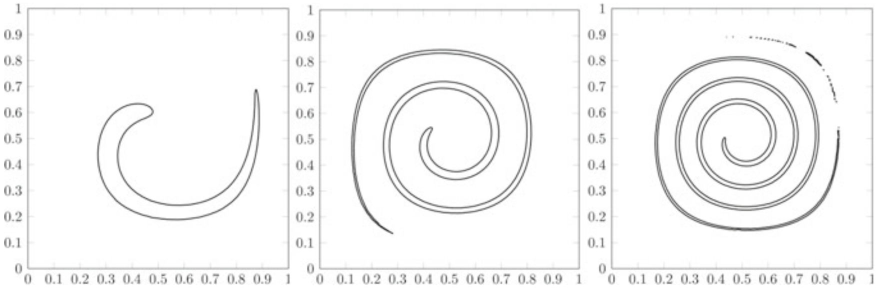


Fig. 5 Evolution of single vortex at times $T = 1, 3, 5$ using mesh M_2, with $\Delta t = 5 \times 10^{-3}$, and $N^{mk} = 1.5 \times 10^6$

4.1.2 Complex Flows

A further step is to show the ability of the code under complex flows. To that effect, we follow the rising bubble case proposed by Hysing et al. [17] in a joint effort to *quantitatively* define the solution of Newtonian rising bubbles. In Fig. 6 we plot the evolution of the relevant variables for the mesh M_2 with $N^{mk} = 5 \times 10^5$, density ratio $\rho_2/\rho_1 = 10^{-3}$, and viscosity ratio $\mu_2/\mu_1 = 10^{-2}$, when $Re = 35$, $Fr = 1$ and $We = 125$. The evolution of the shape at time instants $t = \{1, 2, 2.5, 3\}$ is represented in Fig. 7.

4.2 *Non-Newtonian Fluids*

In this section, we present results for a Newtonian droplet rising in a FENE fluid with maximum extensibility parameter $b = 75$, using the variance-reduction approach pointed out in Sect. 3.2.2, along with the QMSL-PLS method of [2] and the numerical technique outlined in Sect. 3.3 for the efficient solution of the FENE cubic equation.

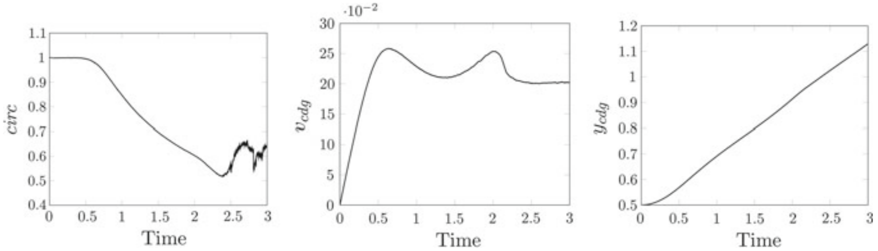


Fig. 6 From *left* to *right*, circularity, velocity v_{cdg}, and center of gravity y_{cdg} in the evolution of a rising bubble with $Re = 35$, $We = 125$ and high density and viscosity ratios

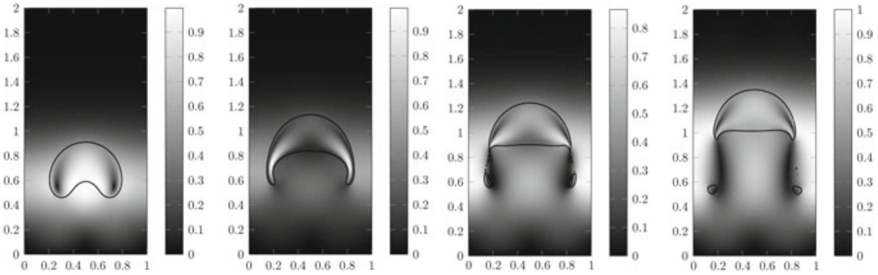


Fig. 7 From *left* to *right*, rising bubble at times $T = 1.0, 2.0, 2.5, 3.0$ with $Re = 35$, $We = 125$ and high density and viscosity ratios

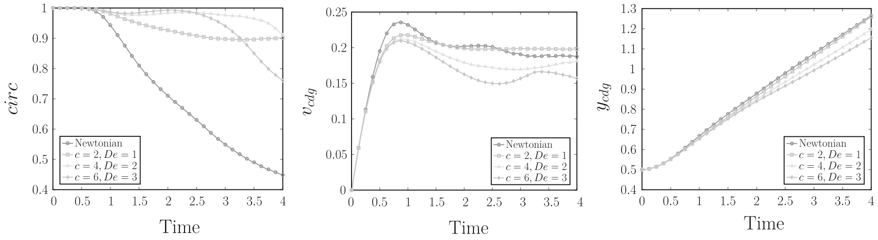


Fig. 8 From *left* to *right*, circularity, velocity v_{cdg}, and center of gravity y_{cdg} in the evolution of a rising bubble with $Re = 35$, $We = 125$, low density and viscosity ratios (10^{-1}) and increasing viscoelastic effects

We study the case of a multi-phase flow ruled by the dimensionless parameters $Re = 35$, $We = 125$, $Fr = 1$, $\rho_2/\rho_1 = 10^{-1} = \mu_2/\mu_1$, and consider a viscoelastic, ambient fluid further defined by the following numbers: $c = 0$ (Newtonian); $c = 2$, $De = 1$; $c = 4$, $De = 2$; and $c = 6$, $De = 3$. The results for circularity, vertical component of the center of mass velocity, and position of the center of mass, are depicted in Fig. 8, for all four cases.

In addition, Fig. 9 collects the streamline pattern observed for these different configurations at time $t = 4$, observing typical viscoelastic effects such as the "negative-wake" and a cusped tail when the dimensionless polymer concentration and Deborah numbers are high enough.

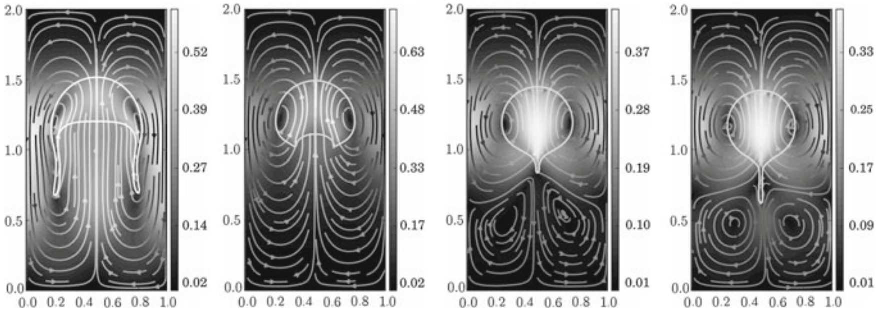


Fig. 9 From *left* to *right*, streamlines at time $t = 4$ for a Newtonian bubble rising in a FENE fluid with $b = 75$, when $c = 0$ (Newtonian); $c = 2$, $De = 1$; $c = 4$, $De = 2$; and $c = 6$, $De = 3$

5 Conclusions

In this work, we have tried to highlight the potential of a multi-scale, micro-macro approach to solve complex, multi-phase flows of Newtonian and non-Newtonian fluids via a Semi-Lagrangian, Particle Level-Set method. To this end, the macroscopic and microscopic equations have been presented and solved in a Finite Element context, and a set of numerical techniques such as variance-reduced, Brownian Dynamics simulations and the efficient computation of the cubic roots for the FENE fluid have been introduced.

Results for Newtonian and non-Newtonian flows using the FENE kinetic model under several configurations, ranging from imposed flows, to complex, multi-phase, viscous(elastic) flows have been included to highlight the versatility, robustness and accuracy of the method. Notwithstanding the limitations of the current implementation, its ability to deal with high density and viscosity ratios, or to replicate experimentally-observed viscoelastic effects are features that make us strongly believe in the promising future of such a multi-scale approach. Additional efforts to palliate some of its present drawbacks are on the way.

Acknowledgments Financial support from project MAT2011-24,834 from Ministerio de Ciencia e Innovación is gratefully acknowledged.

References

1. Adami, S., Hu, X., Adams, N.A.: A transport-velocity formulation for smoothed particle hydrodynamics. J. Comput. Phys. **241**, 292–307 (2013)
2. Bermejo, R., Prieto, J.L.: A semi-Lagrangian particle level set finite element method for interface problems. SIAM J. Sci. Comput. **35**(4), A1815–A1846 (2013)
3. Bird, R.B., Curtiss, C.F., Armstrong, R.C., Hassager, O.: Dynamics of Polymeric Liquids, Vol. 2, Kinetic Theory, second edn. Wiley-Interscience (1987)

4. Bonito, A., Picasso, M., Laso, M.: Numerical simulation of 3D viscoelastic flows with free surfaces. J. Comput. Phys. **215**, 691–716 (2006)
5. Carpio, J., Prieto, J.: An anisotropic, fully adaptive algorithm for the solution of convection dominated equations with semi-Lagrangian schemes. Comput. Methods Appl. Mech. Engrg. **273**, 77–99 (2014)
6. Carpio, J., Prieto, J.L., Bermejo, R.: Anisotropic "Goal-Oriented" mesh adaptivity for elliptic problems. SIAM J. Sci. Comput. **35**(2), A861–A885 (2013)
7. Cormenzana, J., Ledda, A., Laso, M., Debbaut, B.: Calculation of free surface flows using CONNFFESSIT. J. Rheol. **45**(1), 237–258 (2001)
8. Coussot, P.: Yield stress fluid flows: A review of experimental data. J. Non-Newtonian Fluid Mech. **211**, 31–49 (2014)
9. Enright, D., Fedkiw, R., Ferziger, J., Mitchell, I.: A Hybrid Particle Level Set Method for Improved Interface Capturing. J. Comput. Phys. **183**(1), 83–116 (2002)
10. Enright, D., Losasso, F., Fedkiw, R.: A Fast and Accurate Semi-Lagrangian Particle Level Set Method. Computers and Structures **83**, 479–490 (2005)
11. Foteinopoulou, K., Laso, M.: Numerical simulation of bubble dynamics in a Phan-Thien-Tanner liquid: Non-linear shape and size oscillatory response under periodic pressure. Ultrasonics **50**, 758–776 (2010)
12. Grande, E., Laso, M., Picasso, M.: Calculation of variable-topology free surface flows using CONNFFESSIT. J. Non-Newtonian Fluid Mech. **113**, 127–145 (2003)
13. Halin, P., Lielens, G., Keunings, R., Legat, V.: The Lagrangian Particle Method for macroscopic and micro-macro viscoelastic flow computations. J. Non-Newtonian Fluid Mech. **79**, 387–403 (1998)
14. Herrchen, M., Öttinger, H.C.: A detailed comparison of various fene dumbbell models. J. Non-Newtonian Fluid Mech. **68**, 17–42 (1997)
15. Hieber, S.E., Koumoutsakos, P.: A Lagrangian particle level set method. J. Comput. Phys. **210**, 342–367 (2005)
16. Hulsen, M.A., van Heel, A.P.G., van den Brule, B.H.A.A.: Simulation of viscoelastic flows using Brownian Configuration Fields. J. Non-Newtonian Fluid Mech. **70**, 79–101 (1997)
17. Hysing, S., Turek, S., Kuzmin, D., Parolini, N., Burman, E., Ganesan, S., Tobiska, L.: Quantitative benchmark computations of two-dimensional bubble dynamics (2008). http://dx.doi.org/10.1002/fld.1934. Published online by Doi on November 17th, 2008
18. Ilg, P., Öttinger, H.C., Kröger, M.: Systematic time-scale-bridging molecular dynamics applied to flowing polymer melts. Phys. Rev. E **79**, 011,802 (2009)
19. Marchandise, E., Geuzaine, P., Chevaugeon, N., Remacle, J.F.: A stabilized finite element method using a discontinuous level set approach for the computation of bubble dynamics. J. Comput. Phys. **225**, 949–974 (2007)
20. M.B. Liu, G.L.: Smoothed particle hydrodynamics (sph): an overview and recent developments. Arch. Comput. Methods Engrg. **17**, 25–76 (2010)
21. McKee, S., Tomé, M., Ferreira, V., Cuminato, J., Castelo, A., Sousa, F., Mangiavacchi, N.: The MAC method. Computers & Fluids **37**, 907–930 (2008)
22. Oishi, C., Martins, F., Tom, M., Alves, M.: Numerical simulation of drop impact and jet buckling problems using the eXtended PomPom model. J. Non-Newtonian Fluid Mech. **169–170**, 91–103 (2012)
23. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer (2002)
24. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
25. Öttinger, H.C.: Stochastic Processes in Polymeric Fluids: Tools and Examples for Developing Simulation Algorithms. Springer, Berlin (1996)
26. Pasquali, M., Scriven, L.E.: Free surface flows of polymer solutions. J. Non-Newtonian Fluid Mech. **108**, 363–409 (2002)
27. Pietro, D.A.D., Forte, S.L., Parolini, N.: Mass preserving finite element implementations of the level set method. Applied Numer. Math. **56**, 1179–1195 (2006)

28. Pillapakkam, S.B., Singh, P.: A Level-Set Method for Computing Solutions to Viscoelastic Two-Phase Flow. J. Comput. Phys. **174**, 552–578 (2001)
29. Pillapakkam, S.B., Singh, P., Blackmore, D., Aubry, N.: Transient and steady state of a rising bubble in a viscoelastic fluid. J. Fluid Mech. **589**, 215–252 (2007)
30. Prieto, J.L., Bermejo, R., Laso, M.: A semi-Lagrangian micro-macro method for viscoelastic flow calculations. J. Non-Newtonian Fluid Mech. **165**, 120–135 (2010)
31. Prieto, J.L., Ilg, P., Bermejo, R., Laso, M.: Stochastic semi-Lagrangian micro-macro calculations of Liquid Crystalline solutions in complex flows. J. Non-Newtonian Fluid Mech. **165**, 185–195 (2010)
32. Sethian, J.A.: Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press (1999)
33. Siegel, A., Smith, K., Felker, K., Romano, P., Forget, B., Beckman, P.: Improved cache performance in Monte Carlo transport calculations using energy banding. Computer Physics Communications (2014, In press). doi:10.1016/j.cpc.2013.10.008
34. Somasi, M., Khomami, B., Woo, N., Hur, J., Shaqfeh, E.: Brownian dynamics simulations of bead-rod and bead-spring chains: numerical algorithms and coarse-graining issues. J. Non-Newtonian Fluid Mech. **108**, 227–255 (2002)
35. Sussman, M., Fatemi, M., Smereka, P., Osher, S.: An Improved Level Set Method for Incompressible Two-Phase Flows. Computers and Fluids **27**, 663–680 (1998)
36. Tornberg, A.K., Engquist, B.: A finite element based level-set method for multiphase flow applications. Comput. Visual Sci. **3**, 93–101 (2000)
37. Völtz, C., Maeda, Y., Tabe, Y., Yokoyama, H.: Director-configurational transitions around microbubbles of hydrostatically regulated size in liquid crystals. Phys. Rev. Lett. **97**, 227,801 (2006)
38. Wang, Z., Yang, J., Stern, F.: A new volume-of-fluid method with a constructed distance function on general structured grids. J. Comput. Phys. **231**, 3703–3722 (2012)
39. Zainali, A., Tofighi, N., Shadloo, M., Yildiz, M.: Numerical investigation of Newtonian and non-Newtonian multiphase flows using ISPH method. Comput. Methods. Appl. Mech. Engrg. **254**, 99–113 (2013)

Numerical Modeling of Flow-Driven Piezoelectric Energy Harvesting Devices

S. Ravi and A. Zilian

Abstract The present work proposes uniform and simultaneous computational analysis of smart, low power energy harvesting devices targeting flow-induced vibrations in order to enable reliable sensitivity, robustness and efficiency studies of the associated nonlinear system involving fluid, structure, piezo-ceramics and electric circuit. The article introduces a monolithic approach that provides simultaneous modeling and analysis of the coupled energy harvester, which involves surface-coupled fluid-structure interaction, volume-coupled piezoelectric mechanics and a controlling energy harvesting circuit for applications in energy harvesting. A space-time finite element approximation is used for the numerical solution of the governing equations of the flow-driven piezoelectric energy harvesting device. This method enables modeling of different types of structures (plate, shells) with varying cross sections and material compositions, and different types of simple and advanced harvesting circuits.

1 Introduction

Energy harvesting is the process of generating usable electrical energy acquired from various ambient energy sources such as solar, thermal, fluid and mechanical vibrations that surround a system. A steady increase in the growth of wireless and portable electronic devices has led to the development of sophisticated low-power micro-electromechanical devices (MEMS) such as sensors and actuators. The portable nature of these devices necessitates their ability to carry their own power supply. The aim of energy harvesting is to scavenge energy from the environment to power these electronic devices. Such harvesting methods provide significant incentives to replace batteries as power source for providing electrical energy because

S. Ravi · A. Zilian (✉)
University of Luxembourg, Esch-sur-Alzette, Luxembourg
e-mail: andreas.zilian@uni.lu

S. Ravi
e-mail: srivathsan.ravi@uni.lu

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_15

of the limited lifespans of batteries, and persistent stagnation in the technological development of batteries over the years. Many applications related to wireless sensor networks and low power miniature sensors require them to be fully embedded in the structure and placed in remote locations. Conventional power sources like batteries are not an option for applications where the devices need to have their own power supply for an indefinite period of time and are only periodically maintained.

There are many methods to obtain useful electrical energy from the ambient vibration energy that usually goes untapped. Research interest towards developing energy-harvesting devices (EHDs) has grown rapidly over the past few years, and many methods have been proposed to make use of the ambient source to generate electrical power. Some of these methods include electrostatic generation, electromagnetic induction, dielectric elastomers, and piezoelectric materials. Energy harvesting from piezoelectric materials have gained significant attention, as is evident from the number of literature published every year in this field, due to their ability to convert mechanical energy from cyclic straining directly into useful electrical energy. For reviews on various forms of piezoelectric energy harvesting refer to [1, 2].

The present work proposes uniform and simultaneous computational analysis of smart, low power energy harvesting devices targeting flow-induced vibrations in order to enable reliable sensitivity, robustness and efficiency studies of the associated nonlinear system involving fluid, structure, piezo-ceramics and electric circuit. The article introduces a monolithic approach that provides simultaneous modeling and analysis of the coupled energy harvester, which involves surface-coupled fluid-structure interaction, volume-coupled piezoelectric mechanics and a controlling energy harvesting circuit for applications in energy harvesting. A Space-time finite element approximation is used for the numerical solution of the governing equations of the flow-driven piezoelectric energy harvesting device. This method enables modeling of different types of structures (plate, shells) with varying cross sections and material compositions, and different types of simple and advanced harvesting circuits. It should be noted that it is a common practice in modeling of piezoelectric energy harvesters to consider a simple resistor element as a harvesting circuit.

The outline of this article is as follows. The remainder of Sect. 1 introduces the concept of piezoelectric energy harvesters and provides a brief review of various types of modeling approaches to the problem of piezoelectric energy harvesting from base excitations. Section 2 gives a brief overview of modeling approaches for flow-driven piezoelectric energy harvesters. Section 3 starts with the modeling assumptions of the present study and proceeds to establish in the detail the strong form of the governing equations of the multi-physics problem. The coupling conditions are also explained in Sect. 3. In Sect. 4, the weak form of the governing equations are derived and Sect. 5 explains the nature of space-time interpolation with an illustration. The theoretical concepts established are then applied to the problem of piezoelectric energy harvesting from a piezoelectric bimorph subjected to base excitations and presented as a case study in Sect. 6.

1.1 Harvesting Mechanical Vibrations

The prevalence of mechanical vibrations has attracted significant research interest in vibration-based energy harvesting methods. Power generation from ambient mechanical vibrations usually constitutes the conversion of ambient mechanical vibration into useful electrical energy with the help of an EHD, to power other devices with low power requirements. Piezoelectric transduction offers many advantages over other power generating methods due to it's low form factor, high energy density, ease of integration into other systems, and its unique ability to convert cyclic straining of the material into electrical energy.

Piezoelectric materials exhibit accumulation of electric charges in response to mechanical strains which is known as *direct piezoelectric effect*. The piezoelectric effect is a reversible process, where the materials exhibit change in their shape on application of an electric field known as *inverse piezoelectric effect*. Prototypical piezoelectric EHDs are cantilevers with a seismic mass and are attached to another substrate layer. They can be employed in various modes based on the electric field orientation and the polarization direction. Utilization of a proper coupling mode is one of the ways to increase the amount of energy harvested from the piezoelectric material. Two coupling mode exist viz., the -31 mode and the -33 mode respectively. The former is characterized by the straining of the material in the direction perpendicular to the poling direction and the latter by the straining in the same direction as the poling direction. The cantilever with a seismic mass configuration facilitates a lower resonant frequency in the first bending mode, making it easy to match the resonant frequency of the structure to the ambient vibrations to obtain maximum power output. Such systems are capable of producing power output ranging from a few μW to a few mW.

The performance of these piezoelectric devices depends on various factors like the type of piezoelectric material used, size of the harvesting device, mass distribution, shape of the structure, and vibration modes to name a few. The impact of different geometries on the power density of vibration energy harvesters was studied in [3]. Coupling coefficients, strain distribution, and vibration frequency were perceived as the three limiting factors in the field of piezoelectric power scavenging, and alternative geometries were proposed to address each of these limiting factors. The real world application space was deemed too limited for testing the design considerations to improve coupling co-efficients. The strain distribution in the geometry is improved by varying the width of a beam type structure for the full utilization of straining along the length. Experimentally a 30 % increase in power was observed for trapezoidal beam compared to cantilever beams. An experimental comparison of several types of active composite actuators for power generation was carried out in [4]. The study compares a type of macro-fiber composite called MFCs made of piezoelectric fiber composites (PFCs) and interdigitated electrodes to two other actuators called Quick Pack consisting of a monolithic piezo-layer with standard electrodes and another actuator called Quick Pack IDE with interdigitated electrodes. They were all attached to the same beams and excited at their first twelve natural frequencies.

The results showed that the conventional Quick Pack with standard electrodes was able to harvest significant energy generating 137 μW at 64 Hz while the Quick Pack IDE and MFC produced 29 and 12 μW respectively. It was concluded from the study that although the MFCs' fibrous structure itself was not detrimental to the harvesting capacity, it is the low capacitance due the interdigitated electrodes that deteriorates the power output. Hence the MFCs were found impractical to real-world applications even though they had higher coupling coefficients. It is evident that all design considerations are towards the maximization of the power output from the harvester as the scope of application widens with increase in power generation.

1.2 Models of Piezoelectric Energy Harvesting Devices

Over the years many mathematical models have been proposed for the modeling of piezoelectric energy harvesters ranging from simple SDOF (single degree of freedom) models with closed form solutions for the voltage output and vibration characteristics to more sophisticated analytical and numerical methods to address various aspects of modeling. Many of the early works employed simple SDOF models to predict the voltage response of piezoelectric energy harvesting devices driven by harmonic base excitations. Piezoelectric energy harvesters are usually attached to an external circuit that transforms the harvested energy into usable form. In [5], an equivalent circuit model was proposed to account for the harvesting circuit along with the modeling of the energy harvester. The method discussed the representation of the energy harvester electrically, and then combined with the electrical representation of the harvesting circuit and modeled together in SPICE simulator. This method facilitated the representation of non-linear circuit components but suffered the drawback of simplification necessary for the harvester to be represented electrically. This prevented accounting for any change in harvester's properties during operating conditions. In another model developed in [6], a coupled FEM-circuit method was presented to account for the modeling of electrical circuits where the energy harvester was modeled using finite elements and the coupled to the electrical part modeled using a SPICE simulator. This method had comparable advantages to the model in [5] but is computationally expensive and does not provide a realistic representation of the strongly-coupled physics.

One of the frequently addressed issues in the modeling of piezoelectric energy harvesters is capturing the effect of an attached harvesting circuit. This investigation has led to piezoelectric materials being used in passive shunt damping applications as well. Earlier studies modeled this impact as a viscous damping on the harvester which was a reasonable approximation only in the case of electromagnetic generators as pointed out in [7]. The physics of the piezoelectric system is much more complex, and the impact of a harvesting circuit seems too complex to be modeled as viscous damping. Since it is common understanding that maximum power is harvested at resonance, incorrect modeling of damping will lead to inaccurate result in predicting the frequency of the system. It was shown that the load-resistance dependent variation of the resonance frequency and amplification of the motion at open-circuit frequency are indicators for the need for better representation of the effect of harvesting circuits. In one of the most important works in the field, [7, 8] presents a mathematical model with distributed parameter solution based on Euler-Bernoulli assumptions, and comparisons with several SDOF models are made to point out the inaccuracies in popular SDOF models. Several flaws in SDOF models, ranging from the neglect of base rotary motion up to the simplified modeling of damping induced by piezoelectric coupling, are addressed in this work and correction factors are introduced, where necessary, to the SDOF models. In their work, the harmonic base excitation case is considered a particular solution of the general base excitation which includes superimposed rotary motion of the base as well. The study points out that the inertia due to rigid body motion was neglected in most of the SDOF models and hence a correction factor to this effect is proposed. The relative motion transmissibility function derived from the ratio of tip deflection to base deflection is used to form a non-dimensional basis for comparison of the model to SDOF models. It is shown that the error percentage as a function of dimensionless frequency was as high as 35 % in SDOF model.

Only few piezoelectric energy harvesting applications, where the geometry of the energy harvesting device is simple, lend themselves to analytical solutions [9]. Most of the piezoelectric energy harvesting applications are complex, and numerical methods are needed to obtain the electromechanical response of such systems. A pioneering work on the finite element modeling of piezoelectric materials was presented by [10], where mechanical displacements and electrical potential were used as unknowns and both direct and inverse piezoelectric effect were included in the formulation. Since then numerous piezoelectric finite elements have been developed including beam, plate, shell and solid elements. Readers are encouraged to refer to [11] for a detailed review of different finite element models used to model vibration based piezoelectric energy harvesters. Many of the reviewed finite elements use displacements and electric potential as unknowns as suggested by [10] with a linear approximation of the electric potential through the thickness of the element. However it was shown in [12] that the electric potential has a second order component in bending.

Mixed and hybrid finite element formulations are presented in [13–16]. These formulations contain additional unknown fields besides mechanical displacements and electrical potential which reduces locking phenomena and makes the elements less susceptible to mesh distortion. The most general formulation is presented by [13] and contains six independent unknown fields which are displacements, strains and stresses for the mechanical part and electric potential, electric field and dielectric displacements for the electric part. Further mixed formulations with three and four unknown fields are derived from the six field formulation. A six field formulation is used by [15, 16] with additional enhancements for strain and electric field which further reduce locking.

2 Flow Driven Piezoelectric Energy Harvesters

A glance at recent surveys [1, 2] on piezoelectric energy harvesting indicates that much of the research work in this field is focused on harvesting energy from vibrations due to base excitations or the excitations of the structure to which the harvester is attached. This focus may be attributed to various sources of ambient energy in an urban environment, but alternative sources have to be identified in case of EHD devices placed in remote locations. One potential energy source in such locations is the kinetic energy of fluids, i.e. wind or water, which cater to those requirements. In order to extract energy from fluid flows, the kinetic energy of the fluid must first be transformed into straining energy of the harvester, which is then converted into electrical energy and utilized with appropriate harvesting circuits. A key idea in fluid-driven piezoelectric energy harvesting is to utilize the flow energy through controlled aero- or hydro-elasticity phenomena. Traditionally, the idea is to avoid dangerous fluid-structure interactions. In flow-driven piezoelectric energy harvesting, potentially harmful fluctuations are harnessed to provide power supply to small-scale energy harvesting devices. However, the research on piezoelectric energy harvesters placed directly in the fluid flow is fairly limited. There are different mechanisms to convert the flow-energy into cyclic straining of the energy harvester.

One of the ways of harvesting energy from fluid flow is *instability induced excitation* caused by fluid-intrinsic physical properties. The self-exciting flow instability produces oscillating forces even if the structure is stationary (e.g. Kármán vortex street). A further amplification of the exciting force is possible for fluid-structure feedback. The concept of energy harvesting eel, where the cyclic straining of the harvester was achieved by water flow utilizing Kármán vortex sheets was introduced in [17]. This is one of the first works to study energy harvesting from fluid flow. The fluctuations of "eel" shaped polymer beams placed in the wake of a bluff body was investigated in this research. Tests were conducted on different membranes ranging from 0.1 to 0.7 mm placed in water channel running at speeds ranging between 0.05 and 0.8 ms^{-1}. Two different widths of the bluff bodies viz., 5.08 and 3.81 cm were used to create the vortex sheets. It was shown that the membranes exhibited lock-in behavior to he shedding of the bluff body when they oscillate with the same frequency as the undisturbed wake behind the body.The relationship indicating conditions for locking were derived from Euler-Bernoulli beam theory. It was suggested that the eels should have small stiffness so as not to dampen the oscillations. Though the literature provides PIV (particle image velocimetry) images to support their predictions, it fails to provide any insights into the electrical output and the type of coupling existing between the multi-physics domains. A similar study on energy harvesting eels placed in the wake of a bluff body was carried out in [18].

In addition to vibrations produced by Kármán vortex sheets, *movement-induced excitation*—caused by fluctuating flow forces resulting from movements of the vibrating structural part—also provides a way to utilize flow energy for EHDs. Small deviations from the equilibrium position of the structure induce a redistribution of impacting fluid forces, which further increases initial disturbances. This gives rise to ongoing transfer of flow energy to the structural oscillator and is called dynamic instability. Flutter response of a piezoelectrically damped cantilever pipe utilizing such flow instabilities was studied by [6]. Energy harvesting from fluid flows with attached electrical circuit was studied with a more comprehensive model by [19]. The study considered the case of harvesting energy in the wake of a circular cylinder at high Reynold's number. This study involved a combination of experimental and analytical model. This model also considered the three-way interactions of the fluid-flow, structure and the harvesting circuit at the same time in contrast to the most previous studies with loosely-coupled approach for emulating the real-life scenario. SDOF analytical model was chosen to represent the piezoelectric structure, and the coupling between the circuit and the harvester was considered only under the open-circuit condition which makes the model ineligible in predicting energy output for a finite circuit resistance.

3 Model of a Flow-Driven Piezoelectric EHD

This section introduces the strong form of the equations that govern the fluid flow, the mechanical state and the electrical state of the flow-driven piezoelectric energy harvester. The coupled system consists of a piezoelectric structure placed in fluid flow, potentially in the wake of an arbitrarily shaped body, and connected to an electric circuit as shown in Fig. 1. The solid consists of a substrate structure sandwiched between piezoelectric patches. The individual piezoelectric patches are assumed to be covered with continuous electrodes with one voltage output per patch. The electrodes are connected to a harvesting circuit. The harvesting circuit is assumed to consist of a simple resistance across the electrodes covering the patches, which, as mentioned earlier, is a common practice in modeling piezoelectric energy harvesting devices.

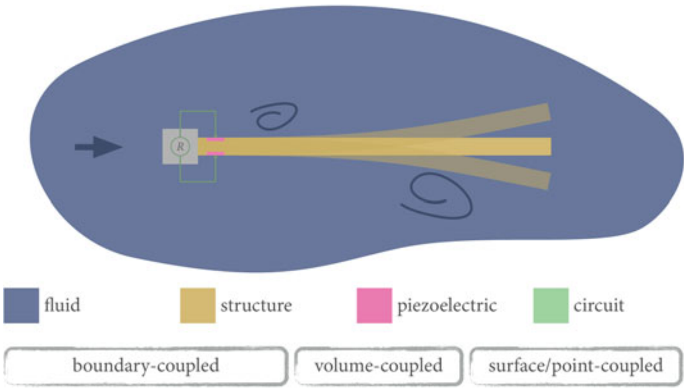


Fig. 1 An illustration of the multi-physics flow-driven piezoelectric EHD based on a cantilever setup

The choice of the solution strategy to a physical problem usually drives the modeling assumptions. The article proposes a monolithic solution strategy to the coupled problem of flow-driven energy harvesting which is a strongly-coupled modeling approach. To this extent, a native coupling between the fluid and the structural domain is achieved by $3D$ modeling of the piezoelectric thin structure formulated in terms of the structural velocity. This modeling approach also enables a straightforward application of constitutive models to piezoelectric coupling. In the present study, the flow is modeled as incompressible and viscous. Turbulence effects are not considered. Both the substrate structure and the piezoelectric material obey linear material laws.

The fluid flow is modeled with the incompressible Navier-Stokes equations, consisting of momentum and mass conservation equations, and is described using an *Eulerian framework* in the current configuration where the space-time domain $Q = \Omega \times I$ within the time interval $I = (t_\mathrm{a}, t_\mathrm{b})$. The equations of the fluid domain are time-dependent and accommodates moving boundaries resulting from structural deformations. The behavior of the piezoelectric structure, and the circuit, within the time interval $I = (t_\mathrm{a}, t_\mathrm{b})$ and occupying the space-time domain $Q_0 = \Omega_0 \times I$, is described using a *Lagrangian* description in *reference* configuration. The subscript "0" refers to the reference configuration.

3.1 Fluid

The incompressible Navier-Stokes equations describing the fluid flow are

$$\rho(\dot{\mathbf{v}} + \mathbf{v} \cdot \nabla\mathbf{v}) - \nabla \cdot \mathsf{T} - \mathbf{f} = \mathbf{0} \quad \text{in} \quad Q \tag{1}$$

and

$$\nabla \cdot \mathbf{v} = 0 \quad \text{in} \quad Q, \tag{2}$$

where $\mathbf{v}$ is the velocity of the fluid, $\mathbf{f}$ is the external body force, and ρ is the density of the fluid. The Cauchy stress tensor T is given by

$$\mathsf{T} + p\mathsf{I} - 2\mu\mathsf{D}(\mathbf{v}) = \mathsf{0} \quad \text{in} \quad Q, \tag{3}$$

where μ is the kinematic viscosity, p is the hydrostatic pressure. The Cauchy stress tensor depends linearly on the strain rate tensor D given by

$$\mathsf{D}(\mathbf{v}) - \frac{1}{2}\left(\nabla\mathbf{v} + (\nabla\mathbf{v})^\top\right) = \mathsf{0} \quad \text{in} \quad Q \tag{4}$$

with the boundary conditions being,

$$\mathbf{v} - \bar{\mathbf{v}} = 0 \quad \text{on} \quad P^{\mathrm{v}} \quad \text{and} \tag{5a}$$

$$\mathbf{t} - \bar{\mathbf{t}} = 0 \quad \text{on} \quad P^{\mathrm{t}}, \tag{5b}$$

where P^{v} in (5a) is the boundary on which velocity $\bar{\mathbf{v}}$ is imposed as a Dirichlet boundary condition, and P^{t} in (5b) is the boundary on which traction $\bar{\mathbf{t}}$ is imposed as a Neumann boundary conditions. The initial condition specifies a divergence free velocity field at time $t = 0$

$$\mathbf{v}(t = 0) = \mathbf{v}_{\mathrm{i}} \quad \text{with} \quad \nabla \cdot \mathbf{v}_{\mathrm{i}} = 0 \quad \text{on} \quad \Omega. \tag{6}$$

3.2 Piezoelectric Structure

The elastodynamic behavior of the piezoelectric structure is modeled based on the assumptions that the deformations are large and the material behavior is linear. The governing equations of the mechanical part of the electro-mechanically coupled piezoelectric structure are as follows

$$\rho_0 \dot{\mathbf{v}} - \boldsymbol{\nabla}_0 \cdot (\mathsf{FS}) - \mathbf{f}_0 = \mathbf{0} \quad \text{in} \quad Q_0, \tag{7}$$

$$\dot{\mathsf{E}} - \frac{1}{2}\left(\boldsymbol{\nabla}_0 \mathbf{v} + (\boldsymbol{\nabla}_0 \mathbf{v})^\top + (\boldsymbol{\nabla}_0 \mathbf{u})^\top \boldsymbol{\nabla}_0 \mathbf{v} + (\boldsymbol{\nabla}_0 \mathbf{v})^\top \boldsymbol{\nabla}_0 \mathbf{u}\right) = \mathsf{0} \quad \text{in} \quad Q_0, \tag{8}$$

$$\dot{\mathsf{E}} - \left[s^{\tilde{\mathsf{D}}}\right] \dot{\mathsf{S}} - [g]^\top \dot{\tilde{\mathsf{D}}}_0 = \mathsf{0} \quad \text{in} \quad Q_0, \tag{9}$$

where (7) is the momentum balance equation, (8) gives the non-linear kinematic relation, and (9) depicts the coupled constitutive relation in rate form for the direct piezoelectric effect. S is the second Piola-Kirchoff tensor, $\dot{\mathsf{E}}$ is the strain rate tensor, $\left[s^{\tilde{\mathsf{D}}}\right]$ is the compliance matrix measured at constant electric displacement, $[g]$ is the piezoelectric coefficient, and $\tilde{\mathsf{D}}$ is the dielectric displacement of the piezoelectric structure.

Velocity $\bar{\mathbf{v}}$ is imposed as a Dirichlet boundary condition on P_0^{v} as

$$\mathbf{v} - \bar{\mathbf{v}} = 0 \quad \text{on} \quad P_0^{\mathrm{v}}\ , \tag{10}$$

whereas traction $\bar{\mathbf{t}}$ is imposed as a Neumann boundary condition on P_0^{t}

$$\mathbf{t} - \bar{\mathbf{t}} = 0 \quad \text{on} \quad P_0^{t}. \tag{11}$$

The electromechanical behavior of the piezoelectric structure is described by Gauss' law which relates the distribution of electric charge to the electric field. A quasi-electrostatic approach is deemed adequate because the phase velocities of the acoustic waves are orders of magnitude less than the velocities of electromagnetic waves. The Gauss' law is given by

$$\nabla_0 \cdot \tilde{\mathsf{D}}_0 = 0 \quad \text{in} \quad \mathcal{Q}_0. \tag{12}$$

The electrical field rate $\dot{\tilde{\mathsf{E}}}$, is related to the electrical potential rate ψ, by the relation

$$\dot{\tilde{\mathsf{E}}}_0 + \nabla_0 \psi = \mathbf{0} \quad \text{in} \quad \mathcal{Q}_0. \tag{13}$$

and the inverse piezoelectric constitutive equation in rate form is given by

$$\dot{\tilde{\mathsf{E}}} + [g]\, \dot{\mathsf{S}} - \left[\varepsilon^{\mathsf{S}}\right]^{-1} \dot{\tilde{\mathsf{D}}}_0 = \mathbf{0} \quad \text{in} \quad \mathcal{Q}_0, \tag{14}$$

where the permittivity matrix, $\left[\varepsilon^{\mathsf{S}}\right]^{-1}$, is measured at constant stress.

It is common in actuation and sensing applications of piezoelectric materials to impose electric potential and charge as Dirichlet and Neumann boundary conditions respectively. In the case of energy harvesting applications, however, both the electric potential and the electric charge are considered as unknowns. Most piezoelectric materials are manufactured with electrodes completely covering their top and bottom surfaces. Thus, a single potential output and charge output can be defined for individual piezoelectric patches. To this effect, appropriate Dirichlet and Neumann boundaries can be defined and a single potential and charge output can be assigned to them.

Electric potential rate $\bar{\psi}$ is given as a Dirichlet boundary condition on P_0^{ψ} as

$$\psi - \bar{\psi} = 0 \quad \text{on} \quad P_0^{\psi}, \tag{15}$$

and electric charge $\bar{q}$ is given as a Neumann boundary condition on P_0^{q} as

$$q - \bar{q} = 0 \quad \text{on} \quad P_0^{\mathrm{q}}. \tag{16}$$

The electric potential rate $\bar{\psi}$, representing one of the two electrical variables defining individual piezoelectric patches, is further expressed as a single electrical potential output $\Phi_{\mathrm{p}}(t)$ of each piezoelectric patch.

It is pertinent to mention at this point, that the electrical field variables can be considered as analogous to the mechanical field variables. The electrical field variables charge and potential rate can be included in our intellection of generalized force and generalized structural velocity respectively. In the case of mixed hybrid model followed in this article, the analogy between electrical and mechanical field variables can be further extended to include electric field and electric displacement in our intellection of mechanical strain and mechanical stress respectively. This intellection provides for a clear comprehension of these quantities in a finite element framework.

3.3 *Circuit*

As mentioned earlier, individual piezoelectric patches are assumed to be covered with continuous electrodes. Free charges are localized on the electrode surface, and each electrode surface gives rise to a single voltage output. A harvesting circuit is attached to the electrodes, and the governing equations of the circuit are

$$I - \dot{Q} = 0 \quad \text{in} \quad I \tag{17}$$

and

$$\Delta\Phi - R \cdot I = 0 \quad \text{in} \quad I \tag{18}$$

where (17) is the charge conservation law. I is the current flowing through the circuit, and Q is the electrical charge flowing though the circuit. Equation (18) is the Ohm's law relating potential difference, $\Delta\Phi$ and the current flowing through the circuit. R is the resistor element. $\Delta\Phi$ is the potential difference existing across the piezoelectric patches covering the substructure, and its value varies depending on the connection (*series* or *parallel*) between the patches.

3.4 *Coupling Conditions*

Interface conditions determine how the different domains of the multi-physics system are coupled with each other, and depending on the interface conditions the modeling of the coupled domains can be either loosely coupled or strongly coupled. Since the research aims to have a strongly-coupled model of the flow-driven piezoelectric energy harvester, suitable interface conditions must be provided to represent the coupling between the fluid domain and the piezoelectric structural domain, and also the coupling between the electrical circuit and the harvester.

3.4.1 Fluid-Structure Interface

To complete the governing equations for the fluid-structure coupling consisting of a moving fluid-domain and vibrating elastic piezoelectric structure, coupling conditions have to be imposed on the interface $P^{\mathrm{C}} = P_0^{\mathrm{S}} \cap P^{\mathrm{F}}$ where the superscripts S

and F denotes solid interface and fluid interface respectively. Geometrical continuity (or mass conservation) at the interface is achieved with the condition

$$\mathbf{v}^{\mathrm{F}} - \mathbf{v}^{\mathrm{S}} = \mathbf{0} \quad \text{on} \quad P_0^{\mathrm{C}}. \tag{19}$$

This leads to the momentum conservation enforced on the interface using the condition given by

$$\mathbf{t}_0 + \frac{d\Gamma_{\mathrm{t}}}{d\Gamma_0}\mathbf{t}^{\mathrm{F}} = \mathbf{0} \quad \text{on} \quad P_0^{\mathrm{C}}. \tag{20}$$

The above relation demands equal tractions along the deforming fluid-structure interface [20].

3.4.2 Circuit-Structure Interface

It is frequently assumed in many literature that the vibration characteristics of the energy harvesting device is independent of the electric circuit. However, as described in length in the previous sections, this assumption can lead to incorrect prediction of the harvester output. Piezoelectric sensors not connected to any circuit are usually modeled as current source in parallel with the capacitance of the piezoelectric material or a voltage source in series with the piezoelectric capacitance where, for the calculation of current source, the electric field $\tilde{\mathsf{E}}$ is assumed as zero for short-circuit conditions, and the dielectric displacement $\tilde{\mathsf{D}}$ is assumed zero for open-circuit calculations. But this condition is no more true in the case of an electric circuit attached directly to the piezoelectric structure. The circuit imposes a relation between the current flowing through the circuit and the voltage developed in the harvester due to the vibrations. This is given by the relation

$$\Phi_{\mathrm{p}}(t) = \Phi_{\mathrm{R}}(t), \tag{21}$$

where the voltage generated by the harvester $\Phi_{\mathrm{p}}(t)$ is assumed, a priori, equal to the voltage across the resistor element $\Phi_{\mathrm{R}}(t)$. The relation between this potential, and the current flowing through the resistor is given by Eq. (18). This relation is indicative of the strong-coupling between the circuitry and the structural domain, and provides a way to understand how an external circuit might impact the power generation capability of a harvesting device.

4 Weak Form of the Governing Equations

In contrast to traditional finite element methods for elastodynamics where the solution is discretized in space, and solved in time domain using finite difference methods for ODE's leading a to semi-discrete formulation, the space-time finite element method

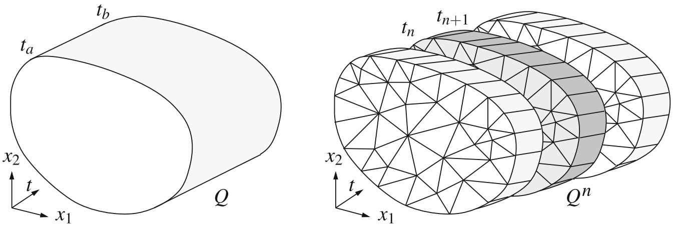


Fig. 2 Discretization of a domain with space-time finite elements using time slabs Q_{n}, see [21]

facilitates consistent discretization of both the space and the time domain leading to a uniform discretization of the governing equations in their weighted residual form. The underlying concept of the space-time finite element method is to include the temporal axis where a space-time domain Q is divided into N time slabs as $Q_{\mathrm{n}} = \Omega_{\mathrm{n}} \times [t_{\mathrm{n}}, t_{\mathrm{n+1}}]$ as shown in Fig. 2.

The time integration is performed using a time-discontinuous Galerkin method. The information flow, as in conventional finite difference methods, is in the direction of positive time, and the discontinuous Galerkin method (DG method) leads to a system in which the solution to a time-slab $Q_{\mathrm{n}} = \Omega_{\mathrm{n}} \times [t_{\mathrm{n}}, t_{\mathrm{n+1}}]$ depends on the solution of the previous time at slab t_{n}^- [22]. Hence the discontinuous approximation of the unknown fields in time leads to additional jump terms in the weak form which are derived in this section. The time integration scheme is implicit, A-Stable and third order accurate for linear interpolation in time [22]. For more detailed information on space-time finite elements readers are referred to [22–24].

4.1 Fluid

The Galerkin weighted residual form of the governing equations from (1) to (5b) describing the fluid flow are given as follows

$$\int_{\mathrm{Q_n}} \delta \mathbf{v} \cdot \rho \left(\dot{\mathbf{v}} + \mathbf{v} \cdot \nabla \mathbf{v} \right) \mathrm{d}Q + 2\mu \int_{\mathrm{Q_n}} \mathsf{D}(\delta \mathbf{v}) : \mathsf{D}(\mathbf{v}) \mathrm{d}Q \tag{22a}$$

$$- \int_{\mathrm{Q_n}} \left(\nabla \cdot \delta \mathbf{v} \right) p \mathrm{d}Q - \int_{\mathrm{Q_n}} \delta \mathbf{v} \cdot \mathbf{f} \mathrm{d}Q \tag{22b}$$

$$+ \int_{\mathrm{Q_n}} \delta p \left(\nabla \cdot \mathbf{v} \right) \mathrm{d}Q \tag{22c}$$

$$+ \int_{\Omega_{\mathrm{n}}} \delta \mathbf{v} \left(t_{\mathrm{n}}^{+} \right) \cdot \rho \Big(\mathbf{v} \left(t_{\mathrm{n}}^{+} \right) - \mathbf{v} \left(t_{\mathrm{n}}^{-} \right) \Big) \mathrm{d}\Omega_{\mathrm{t}} \tag{22d}$$

$$- \int_{\mathrm{P_n^t}} \delta \mathbf{v} \cdot \bar{\mathbf{t}} \mathrm{d}P = 0 \qquad \forall \delta \mathbf{v}, \delta p \tag{22e}$$

where (22a) represents the weighted form of the momentum conservation (1). Equation (3) describing the fluid constitutive equation, and Eq. (4) describing the fluid kinematics are satisfied exactly in Eq. (22a). Equation (22c) is the integral form of the mass conservation Eq. (2), weighted with the fluid pressure. The Neumann boundary condition (5b) is considered in a weak sense in Eq. (22e), while the Dirichlet condition is applied exactly a priori. Equation (22d) contains the jump terms required due to the time differentiation of the velocity using discontinuous Galerkin method, ensuring the transfer of kinetic energy from the end of previous time slab t_{n}^{-} to the beginning of the current time slab at t_{n}^{+} at time t_{n}.

4.2 Structure

Unlike traditional displacement finite elements, velocity based mixed-hybrid finite elements for the structure allows for the native coupling at the fluid structure interface without having to resort to enforcement of continuity using Lagrange multipliers. This coupling strategy is followed in the model presented by this article, and hence the structural discretization is done using velocity based finite element method. The weighted residual form of the governing equations from (7) to (11) of the elastodynamics problem, but without piezoelectric coupling terms, and depicting the behavior of the elastic structure is given as,

$$\int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathbf{v}\rho_0\dot{\mathbf{v}}\mathrm{d}Q_0 + \int_{\mathrm{Q}_{0,\mathrm{n}}} \dot{\mathsf{E}}\left(\delta\mathbf{v},\mathbf{u}\right) : \mathsf{S}\mathrm{d}Q_0 \tag{23a}$$

$$- \int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathbf{v}\cdot\mathbf{f}_0\mathrm{d}Q_0 \tag{23b}$$

$$+ \int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathsf{S} \colon \Big([s]\dot{\mathsf{S}} - \dot{\mathsf{E}}(\mathbf{v},\mathbf{u})\Big)\mathrm{d}Q_0 \tag{23c}$$

$$+ \int_{\Omega_0} \delta\mathbf{v}\cdot\Big(\rho_0\big(\mathbf{v}\left(t_{\mathrm{n}}^{+}\right) - \mathbf{v}\left(t_{\mathrm{n}}^{-}\right)\big)\Big)\mathrm{d}\mathit{\Omega}_0 \tag{23d}$$

$$+ \int_{\Omega_0} \delta\mathsf{S} \colon \Big([s]\big(\mathsf{S}\left(t_{\mathrm{n}}^{+}\right) - \mathsf{S}\left(t_{\mathrm{n}}^{-}\right)\big)\Big)\mathrm{d}\mathit{\Omega}_0 \tag{23e}$$

$$- \int_{\mathrm{P}_0^{\mathrm{t}}} \delta\mathbf{v}\cdot\bar{\mathbf{t}}_0\mathrm{d}P_0 = 0 \qquad \forall\delta\mathbf{v},\delta\mathbf{S} \tag{23f}$$

where (23a) and (23b) are the integral forms of (7), weighted with velocity. The constitutive law given by (9), but without the coupling term, is solved in a weak sense on the element level leading to a mixed hybrid method. The jump terms in (23d) and (23e) allow for the consistent transfer of kinetic energy and internal mechanical energy between the time slabs t_{n}^{-} and t_{n}^{+}. The boundary tractions are considered in (23f). The displacement state **u** can be computed by the integration of the structural velocity **v**.

4.3 Piezoelectric Material

The integral form of (7)–(11) including the piezoelectric coupling terms is given as

$$\int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathbf{v}\cdot\rho_0\dot{\mathbf{v}}\mathrm{d}Q_0 + \int_{\mathrm{Q}_{0,\mathrm{n}}} \dot{\mathsf{E}}\left(\delta\mathbf{v},\mathbf{u}\right) : \mathsf{S}\mathrm{d}Q_0 \tag{24a}$$

$$- \int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathbf{v}\cdot\mathbf{f}_0\mathrm{d}Q_0 \tag{24b}$$

$$- \int_{\mathrm{Q}_{0,\mathrm{n}}} \dot{\tilde{\mathsf{E}}}_0(\delta\psi)\cdot\tilde{\mathsf{D}}_0\mathrm{d}Q_0 \tag{24c}$$

$$+ \int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\mathsf{S} \colon \left(\left[s^{\tilde{\mathsf{D}}}\right]\dot{\mathsf{S}} + [g]^{\top}\dot{\tilde{\mathsf{D}}}_0 - \dot{\mathsf{E}}(\mathbf{v},\mathbf{u})\right)\mathrm{d}Q_0 \tag{24d}$$

$$+ \int_{\mathrm{Q}_{0,\mathrm{n}}} \delta\tilde{\mathsf{D}}_0\cdot\left(-[g]^{\top}\dot{\mathsf{S}} + \left[\varepsilon^{\mathsf{S}}\right]^{-1}\dot{\tilde{\mathsf{D}}}_0 - \dot{\tilde{\mathsf{E}}}_0(\psi)\right)\mathrm{d}Q_0 \tag{24e}$$

$$+ \int_{\Omega_0} \delta\mathbf{v}\cdot\left(\rho_0\big(\mathbf{v}\left(t_{\mathrm{n}}^{+}\right) - \mathbf{v}\left(t_{\mathrm{n}}^{-}\right)\big)\right)\mathrm{d}\Omega_0 \tag{24f}$$

$$+ \int_{\Omega_0} \delta\mathsf{S}\colon \left(\left[s^{\tilde{\mathsf{D}}}\right]\left(\mathsf{S}(t_{\mathrm{n}}^{+}) - \mathsf{S}(t_{\mathrm{n}}^{-})\right)\right)\mathrm{d}\Omega_0 \tag{24g}$$

$$+ \int_{\Omega_0} \delta\mathsf{S}\colon \left([g]^{\top}\big(\tilde{\mathsf{D}}_0(t_{\mathrm{n}}^{+}) - \tilde{\mathsf{D}}_0(t_{\mathrm{n}}^{-})\big)\right)\mathrm{d}\Omega_0 \tag{24h}$$

$$+ \int_{\Omega_0} \delta\tilde{\mathsf{D}}_0\cdot\left(-[g]\big(\mathsf{S}(t_{\mathrm{n}}^{+}) - \mathsf{S}(t_{\mathrm{n}}^{-})\big)\right)\mathrm{d}\Omega_0 \tag{24i}$$

$$+ \int_{\Omega_0} \delta\tilde{\mathsf{D}}_0\cdot\left(\left[\varepsilon^{\mathsf{S}}\right]^{-1}\big(\tilde{\mathsf{D}}_0(t_{\mathrm{n}}^{+}) - \tilde{\mathsf{D}}_0(t_{\mathrm{n}}^{-})\big)\right)\mathrm{d}\Omega_0 \tag{24j}$$

$$- \int_{\mathrm{P}_0^{\mathrm{t}}} \delta\mathbf{v}\cdot\bar{\mathbf{t}}_0\mathrm{d}P_0 - \int_{P_0^{\psi}} \delta q\bar{\psi}dP_0 = 0. \qquad \forall\delta\mathbf{v},\delta\psi,\delta\mathsf{S},\delta\tilde{\mathsf{D}}_0 \tag{24k}$$

As seen in the case of the purely elastic problem, the constitutive Eqs. (9) and (14) are solved in a weak sense in (24c) and (24d) on element level leading to a mixed hybrid form with mechanical stress and electric displacement as element level unknowns.They are spatially discontinuous at element edges and can be condensed on the element level. Potential rate, ψ is the additional global unknown field analogous to the velocity field.

4.4 Circuit

The integral form of the electrodes that cover the piezoelectric patches, and the harvester circuit attached to the electrodes are given as

$$- \int_{\mathrm{P}_0^{\mathrm{E}}} \delta q \psi \mathrm{d} P_0 - \int_{\mathrm{P}_0^{\mathrm{E}}} \delta \psi q \mathrm{d} P_0 + \int_{\mathrm{P}_0^{\mathrm{E}}} \delta q \dot{\Phi}_{\mathrm{p}} \mathrm{d} P_0 \qquad \forall \delta \psi, \delta q \tag{25a}$$

$$- \int_{\mathrm{I}} \delta \Phi_{\mathrm{p}} \big(\frac{\Phi}{R} - \int_{\mathrm{P}_0^{\mathrm{E}}} \dot{q} \mathrm{d} P \big) \mathrm{d} t = 0 \qquad \forall \delta \Phi_{\mathrm{p}} \tag{25b}$$

where (25a) takes into account the charges localized on the continuous electrodes covering the patch, and the relation between the boundary charges and the single potential output of a piezoelectric patch is given in (25b). These terms together naturally enforce the equipotential condition of the electrodes explained in Sect. 3.2 and (21).

5 Discretization with Space-Time Finite Elements

This subsection details the procedure involved in discretization of the domains constituting the problem setup using different types of finite elements. As a first step, an illustration of the different domains constituting the coupled system and the corresponding unknown fields is shown in Fig. 3.

5.1 Elements and Space-Time Interpolation

An eight node hexahedral element as depicted in Fig. 4 is chosen to discretize the fluid, structural and the piezoelectric weak forms spatially. Several works [20, 21, 24, 25] detail the application of the space-time finite element method for fluid-structure interaction. Since the electric charges are collected only on the surface of the electrodes, a four node quadrilateral element as shown in Fig. 5 is chosen to

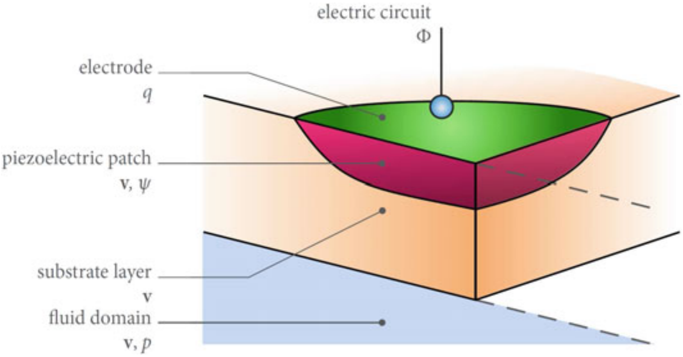


Fig. 3 An illustration of the coupled multi-physics domains with the associated unknown fields

Fig. 4 An eight-node hexahedral element and its node numbering sequence

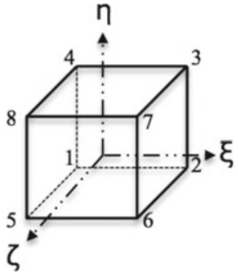
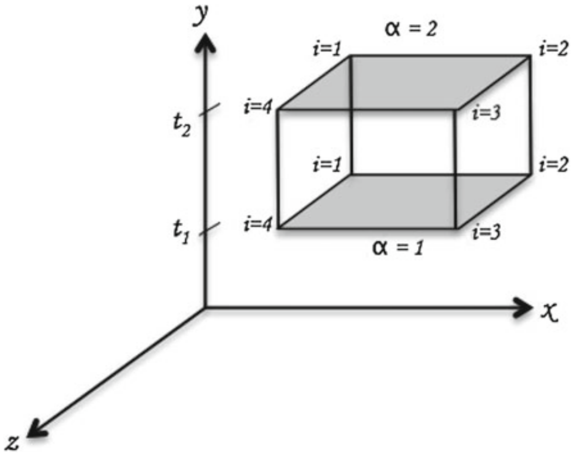


Fig. 5 A four-node quadrilateral space-time element



discretize the electrodes covering the piezoelectric patch. The quadrilateral element contains the electric charge (q) and the potential rate (ψ) as its degrees if freedom (dof). The electrode cover over the piezoelectric patch is assumed to be continuous. Thus each piezoelectric patch gives rise to a single potential output. This equipotential condition is enforced naturally in the formulation by coupling the nodal electrical potential rate d.o.f.s of the quadrilateral element to the virtual node containing the electrical potential degree of freedom (Φ). The potential output of the harvester is also the potential across the resistive element constituting the harvester circuit. This is explained in detail later in this section. The hexahedral element allows for a straightforward coupling of different domains (fluid, structure, and the piezoelectric material) without resorting to simplification of the problem under consideration. This approach also enables a strongly-coupled representation of the multi-physics problem. Locking phenomenon, which is encountered in modeling of thin structures using three dimensional elements is mitigated by adopting a mixed formulation. The thickness of the electrode is sufficiently smaller than the thickness of the piezoelectric patch, that the electrode layer is discretized using a four-node quadrilateral element.

The various assumed field variables within a generic finite element are discretized as

$$\mathbf{v} = \mathsf{N}^{\alpha}_{\mathbf{v}}\mathbf{v}_{\mathrm{m}} \tag{26a}$$

$$\psi = \mathsf{N}^{\alpha}_{\psi}\boldsymbol{\psi}_{\mathrm{e}} \tag{26b}$$

$$\mathsf{S} = \mathsf{N}^{\alpha}_{\mathsf{S}}\mathsf{S}_{\mathrm{m}} \tag{26c}$$

$$\tilde{\mathsf{D}} = \mathsf{N}^{\alpha}_{\tilde{D}}\tilde{\mathsf{D}}_{\mathrm{e}} \tag{26d}$$

$$q = \mathsf{N}^{\alpha}_{\mathrm{q}}\mathbf{q}_{\mathrm{e}} \tag{26e}$$

$$\Phi = \mathsf{N}^{\alpha}_{\Phi}\Phi_{\mathrm{e}} \tag{26f}$$

where $\mathsf{N}^{\alpha}_{\mathbf{v}}$ in (26a) is the velocity interpolation matrix, and $\mathbf{v}_{\mathrm{m}}$ is the vector of nodal velocity d.o.f.s. $\mathsf{N}^{\alpha}_{\psi}$ in (26b) is the electric potential rate interpolation matrix, and $\boldsymbol{\psi}_{\mathrm{e}}$ is vector of nodal electric potential rate dofs. $\mathsf{N}^{\alpha}_{\mathsf{S}}$ in (26c) is the mechanical stress shape function matrix, and S_{m} is the vector of stress co-efficients. $\mathsf{N}^{\alpha}_{\tilde{\mathsf{D}}}$ in (26d) is the electric displacement shape function matrix, and $\tilde{\mathsf{D}}_{\mathrm{e}}$ is the vector of electric displacement coefficients. $\mathsf{N}^{\alpha}_{\mathrm{q}}$ in (26e) is the charge interpolation matrix, and $\mathbf{q}_{\mathrm{e}}$ is the vector of nodal charge d.o.f.s. $\mathsf{N}^{\alpha}_{\Phi}$ in (26f) is the electric potential interpolation matrix, and Φ_{e} is the vector of nodal electric potential d.o.f.s. The subscripts "e" and "m" refer to mechanical and electrical quantities respectively. The superscript "α" refers to the fact that the fields are interpolated in space and time. The velocity and electric potential rate satisfy the continuity requirements. The assumed mechanical stress and dielectric displacement are not expressed in terms of nodal values, but through unique shape functions and can be independent of the ones in other elements.

It is pertinent at this point to explain the nature of space-time interpolation. For the sake of clarity and brevity, the four-node space-time quadrilateral element shown in Fig. 5 is taken as an example to briefly explain the derivation of space-time interpolation functions. For a more extensive study, readers are referred to [22, 24]. The extension of the concept to an eight-node hexahedral element is fairly straightforward. Typical shape functions used in spatial finite elements have an additional temporal component in space-time finite elements. For the four-node quadrilateral element, the spatial and temporal components of the typical space-time shape functions are given by

$$N^{\alpha}_{\mathrm{i}} = N_{\mathrm{i}}T^{\alpha} = N_{\mathrm{i}}(\xi,\eta)T^{\alpha}(\theta), \tag{27}$$

where "$i = 1, 2, \ldots n_{nodes}$" refers to the number of nodes and "$\alpha = 1, 2, \ldots$" refers to the temporal division of the time slab. $\xi, \eta \in [-1, +1]$ are the natural spatial co-ordinates, and $\theta \in [-1, +1]$ is the natural temporal co-ordinate.

Temporal shape function can be explicitly defined as opposed to the spatial shape functions which usually depends on the spatial dimensions. The temporal shape function is expressed as

$$T^{1}(\theta) = \frac{1}{2}(1-\theta), \quad T^{2}(\theta) = \frac{1}{2}(1+\theta). \tag{28}$$

Time derivative of any unknown field can be readily obtained by taking the derivative of (28) and is expressed as

$$T^1_{,\theta} = -\frac{1}{2}, \quad T^2_{,\theta} = +\frac{1}{2}. \tag{29}$$

The spatial co-ordinates x and y can be interpolated in space and time as given by the following equations assuming linear interpolation in time

$$x = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N^\alpha_\mathrm{i} x^\alpha_\mathrm{i} = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} T^\alpha x^\alpha_\mathrm{i} \tag{30a}$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i}(T^1 x^1_\mathrm{i} + T^2 x^2_\mathrm{i}) = \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} x_\mathrm{i}(\theta) \tag{30b}$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} x_\mathrm{i} \tag{30c}$$

and

$$y = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N^\alpha_\mathrm{i} y^\alpha_\mathrm{i} = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} T^\alpha y^\alpha_\mathrm{i} \tag{31a}$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i}(T^1 y^1_\mathrm{i} + T^2 y^2_\mathrm{i}) = \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} y_\mathrm{i}(\theta) \tag{31b}$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} y_\mathrm{i}. \tag{31c}$$

Similarly, we can interpolate the temporal co-ordinate t, where $t^\alpha_\mathrm{i} = t^\alpha$ for "$i = 1, 2, 3, \ldots n$" as

$$t = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N^\alpha_\mathrm{i} T^\alpha_\mathrm{i} = \sum_{\alpha=1}^{\alpha=2}\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i} T^\alpha t^\alpha \tag{32a}$$

$$= \left(\sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} N_\mathrm{i}\right) \sum_{\alpha=1}^{\alpha=2} T^\alpha t^\alpha = T^1 t^1 + T^2 t^2 \tag{32b}$$

The derivative of the shape functions with respect to the global axis is given by the following equations

$$\begin{bmatrix} N^{\alpha}_{\mathrm{a,x}} \\ N^{\alpha}_{\mathrm{a,y}} \\ N^{\alpha}_{\mathrm{a,t}} \end{bmatrix} = \begin{bmatrix} \xi_{,\mathrm{x}} & \eta_{,\mathrm{x}} & \theta_{,\mathrm{x}} \\ \xi_{,\mathrm{y}} & \eta_{,\mathrm{y}} & \theta_{,\mathrm{y}} \\ \xi_{,\mathrm{t}} & \eta_{,\mathrm{t}} & \theta_{,\mathrm{t}} \end{bmatrix} \begin{bmatrix} N^{\alpha}_{\mathrm{a},\xi} \\ N^{\alpha}_{\mathrm{a},\eta} \\ N^{\alpha}_{\mathrm{a},\theta} \end{bmatrix} \tag{33a}$$

$$= \begin{bmatrix} x_{,\xi} & y_{,\xi} & t_{,\xi} \\ x_{,\eta} & y_{,\eta} & t_{,\eta} \\ x_{,\theta} & y_{,\theta} & t_{,\theta} \end{bmatrix}^{-1} \begin{bmatrix} N^{\alpha}_{\mathrm{a},\xi} \\ N^{\alpha}_{\mathrm{a},\eta} \\ N^{\alpha}_{\mathrm{a},\theta} \end{bmatrix}, \tag{33b}$$

where in (33b) the derivative of global time axis with respect to the local spatial axes is zero $t_{,\xi}, t_{,\eta} = 0$. Also, the structure is modeled in Lagrangian framework and hence the derivative of global axes with respect to the local time axes is zero $x_{,\theta}, y_{,\theta} = 0$ and $t_{,\theta} = \frac{\Delta t}{2}$.

With this brief introduction in place, for the eight-node hexahedral element shown in Fig. 4, the space-time interpolation function for the ith node can expressed as

$$N^{\alpha}_{\mathrm{i}} = N_{\mathrm{i}} T^{\alpha} = N_{\mathrm{i}}(\xi, \eta, \zeta) T^{\alpha}(\theta), \tag{34}$$

where $N_{\mathrm{i}}, i = 1, \ldots, n_{nodes}$, as seen earlier, is the spatial interpolation function and is given by

$$N_{\mathrm{i}} = \frac{1}{8}(1 + \xi_{\mathrm{i}}\xi)(1 + \eta_{\mathrm{i}}\eta)(1 + \zeta_{\mathrm{i}}\zeta) \tag{35}$$

in which ξ, η and $\zeta \in [-1, +1]$ are the natural spatial co-ordinates.

The assumed field variables potential rate and velocity can be interpolated as

$$\psi = [T^1[N_1, \ldots, N_8]\, T^2[N_1, \ldots, N_8]]\{\psi_1, \ldots \psi_{16}\}^{\top} \tag{36a}$$

$$= \mathsf{N}^{\alpha}_{\psi} \boldsymbol{\psi}_{\mathrm{e}} \tag{36b}$$

$$\mathbf{v} = [T^1[N_1 \mathsf{I}_3, \ldots, N_8 \mathsf{I}_3]\, T^2[N_1 \mathsf{I}_3, \ldots, N_8 \mathsf{I}_3]]\{v_1, \ldots v_{48}\}^{\top} \tag{36c}$$

$$= \mathsf{N}^{\alpha}_{\mathbf{v}} \mathbf{v}_{\mathrm{m}} \tag{36d}$$

where in (36a) and (36c), it can be seen that the are 16 potential rate dofs, and 48 velocity dofs. respectively. This is because the interpolation is performed in space and time. This is true in case of all the field variables. In (36c), I_{i} is the ith order identity matrix.

The mechanical stress and electric displacement are approximated using unique shape functions. For the mechanical stress, the stress shape function used in the Pian's hybrid element [26–28] is used. Pian's hybrid element contains 18 stress modes. For the electric displacement, the shape function employed in [13] is used. Sze's element contains 7 assumed electric displacement modes. In a space-time setting, the mechanical stress and electric displacement will have 36 and 14 assumed modes respectively. The number of assumed mechanical stress and electric displacement modes are chosen so to secure proper element rank.

The interpolation of the mechanical stress can be expressed as

$$\begin{aligned}\mathsf{S} &= [T^1[\mathsf{I}_6\ \mathsf{T}_{\mathrm{m}}\mathsf{N}_{\mathsf{S}}]\ T^2[\mathsf{I}_6\ \mathsf{T}_{\mathrm{m}}\mathsf{N}_{\mathsf{S}}]]\{S_1, \ldots, S_{36}\}^{\top} && \text{(37a)}\\ &= \mathsf{N}^{\alpha}_{\mathsf{S}}\mathsf{S}_{\mathrm{m}} && \text{(37b)}\end{aligned}$$

where T_{m} is the transformation matrix evaluated at the element origin, I is the identity matrix, and N_{S} is the stress shape function matrix and given as

$$\begin{bmatrix} 0 & 0 & 0 & \eta & 0 & 0 & \zeta & 0 & 0 & \eta\zeta & 0 & 0 \\ \xi & 0 & 0 & 0 & 0 & 0 & 0 & \zeta & 0 & 0 & \zeta\xi & 0 \\ 0 & \xi & 0 & 0 & \eta & 0 & 0 & 0 & 0 & 0 & 0 & \xi\eta \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \zeta & 0 & 0 & 0 \\ 0 & 0 & \xi & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \eta & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{38}$$

The electrical displacement shape function as given in [13], in a space-time setting, can be expressed as

$$\tilde{\mathsf{D}} = \left[T^1[\mathsf{I}_3\ \mathsf{T}_{\mathrm{e}}\mathsf{N}_{\tilde{\mathsf{D}}}]\ T^2[\mathsf{I}_3\ \mathsf{T}_{\mathrm{e}}\mathsf{N}_{\tilde{\mathsf{D}}}]\right]\{D_1, \ldots, D_{14}\}^{\top}, \tag{39}$$

where in (39), T_{e} is the Jacobian matrix evaluated at the origin of the natural coordinates, and $\mathsf{N}_{\tilde{\mathsf{D}}}$ is the electric displacement shape function as given by

$$\mathsf{N}_{\tilde{\mathsf{D}}} = \begin{bmatrix} \eta & 0 & \zeta & \eta\zeta \\ \xi & \zeta & 0 & \zeta\xi \\ 0 & \eta & \xi & \xi\eta \end{bmatrix}. \tag{40}$$

As mentioned earlier, the four-node quadrilateral space-time element shown in Fig. 5 is employed as a boundary element to discretize the charges localized on the electrode surface. For this element, the space-time interpolation function for the ith node is given in (27), where

$$N_{\mathrm{i}} = \frac{1}{4}(1 + \xi_{\mathrm{i}}\xi)(1 + \eta_{\mathrm{i}}\eta). \tag{41}$$

The electric charge can be interpolated as,

$$\begin{aligned}q &= \left[T^1[N_1, \ldots, N_4]\ T^2[N_1, \ldots N_4]\right]\{q_1, \ldots, q_8\}^{\top} && \text{(42a)}\\ &= \mathsf{N}^{\alpha}_{\mathrm{q}}\mathbf{q}_{\mathrm{e}}. && \text{(42b)}\end{aligned}$$

The electric potential is discretized only in time domain to meet the equipotential condition. It is represented by a single virtual node in space, and thus has a unit shape function in space. The space-time interpolation of the electric potential is given expressed as

$$\varPhi = \left[T^1[N_1]\ T^2[N_1]\right]\{\varPhi_1, \varPhi_2\}^\top \tag{43a}$$

$$= \mathsf{N}^\alpha_\varPhi \Phi_e \tag{43b}$$

where N_1 in (43a) is 1.

The expressions for mechanical strain rate ($\dot{\mathbf{E}}$) in (8) and electric field rate ($\dot{\tilde{\mathbf{E}}}$) in (13) can be obtained by differentiating (36d) and (36b) respectively, and expressed as

$$\dot{\mathsf{E}} = \mathsf{B}^\alpha_{\mathbf{v}} \mathbf{v}_\mathrm{m} \tag{44a}$$

$$\dot{\tilde{\mathsf{E}}} = \mathsf{B}^\alpha_\psi \psi_\mathrm{e}. \tag{44b}$$

As mentioned earlier, time derivative of any unknown field can easily be obtained by taking the derivative of time interpolation function as given in (29) and multiplying the spatial interpolation function. As an example, time derivative of velocity can be expressed as follows:

$$\dot{\mathbf{v}} = [T^1_{,\theta}[\mathsf{N}_\mathbf{v}]\ T^2_{,\theta}[\mathsf{N}_\mathbf{v}]]\{v_1, \ldots, v_{48}\}^\top \tag{45a}$$

$$= \dot{\mathsf{N}}^\alpha_\mathbf{v}\mathbf{v}_\mathrm{m}. \tag{45b}$$

Time derivative of other unknown fields can be obtained in the same way as shown above.

5.2 Monolithic Solution Strategy

At the element level, the space-time discretization of a specific time slab Q_n applied to the weak form of the coupled system (22a)–(25b), leads to a system of coupled algebraic equations:

$$\mathsf{K}_{lin}(\mathbf{v}_\mathrm{i}, \psi_\mathrm{i}, q_\mathrm{i}, \varPhi_\mathrm{i}, \mathsf{S}, \tilde{\mathsf{D}})\varDelta\mathbf{x} = \mathbf{r}, \tag{46}$$

where K_{lin} is the element level coefficient matrix, $\varDelta\mathbf{x}$ is the vector of unknowns, and $\mathbf{r}$ is the residual vector. The mechanical stress and the electric displacement are discontinuous across the elements, they can be statically condensed on the element level. The resulting element-level matrices are assembled in the usual way by nodal addition of elemental contributions and can be expressed as

$$\mathsf{K}^*_{lin}(\mathbf{v}_\mathrm{i}, \psi_\mathrm{i}, q_\mathrm{i}, \varPhi_\mathrm{i})\varDelta\mathbf{x}^* = \mathbf{r}^* \tag{47}$$

where K^*_{lin} is the global coefficient matrix, $\varDelta\mathbf{x}^*$ is the vector of global unknowns, and $\mathbf{r}^*$ is the global residual vector.

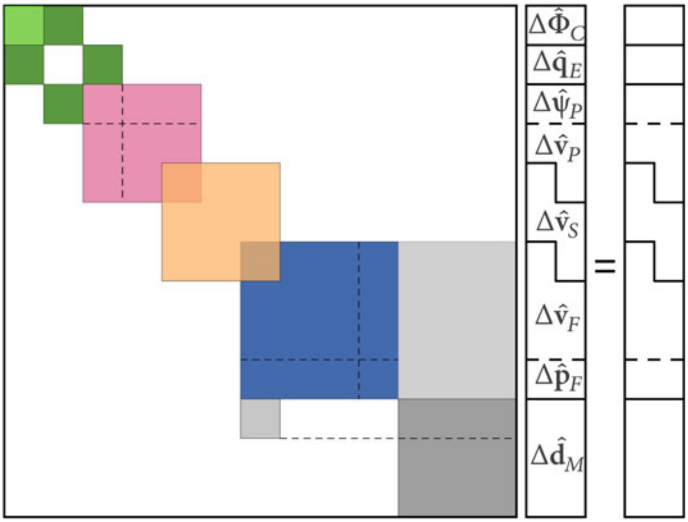


Fig. 6 Structure of the monolithic algebraic system including mesh-deformation d.o.f.s

Equation (47) is a monolithic algebraic representation of the discretized coupled multi-physics problem, and the solution to the unknown fields as shown in Fig. 6 is obtained using the Newton-Raphson iterative scheme.

6 Numerical Example

This section is devoted to the application of theoretical concepts explained in the preceding sections to the problem of piezoelectric energy harvesting from base excitations.

6.1 Problem Setup

The bimorph cantilever beam considered in this numerical example is also discussed in [8], and the basic setup is shown in Fig. 7 The substructure is sandwiched between two identical piezo patches which are fully covered with conductive electrodes. The piezo patches are polarized in the same direction thus constituting a parallel connection between the electrodes. A harvesting circuit constituting a resistor R is also attached to the electrodes. The base excitation causes longitudinal strains (x direction) in the beam which are coupled to the electric field in transverse direction (y direction) leading to a -31 coupling of the piezoelectric elements. The geometric and material properties of the piezoceramic and substrate layers are given in Table 1.

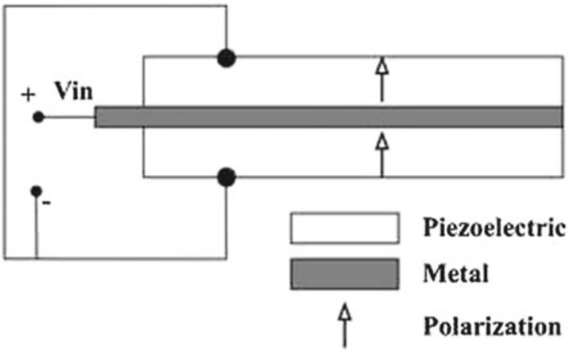


Fig. 7 Parallel configuration of piezoelectric bimorph [29]

Table 1 Geometric and material properties of the energy harvester

Quantity	Dimension	Value
Length of the beam	L (mm)	50.8
Width of the beam	b (mm)	31.8
Thickness of the piezo. patch (PZT-5A)	h_p (mm)	0.26 (each)
Thickness of the substructure	h_s (mm)	0.14
Young's modulus of the substructure (brass)	Y_s (GPa)	105
Young's modulus of PZT-5A	Y_p (GPa)	66
Mass density of the substructure (brass)	ρ_s $\left(\mathrm{Kg\,m^{-3}}\right)$	9000
Mass density of PZT-5A	ρ_p $\left(\mathrm{Kg\,m^{-3}}\right)$	7800
Piezoelectric displacement coefficient	d_{31} $\left(\mathrm{pm\,V^{-1}}\right)$	-190
Permittivity	$\varepsilon_{33}^{\mathsf{E}}$ $\left(\mathrm{F\,m^{-1}}\right)$	$1500\,\varepsilon_0$

In Table 1, the permittivity $\varepsilon_{33}^{\mathsf{E}}$ ($\varepsilon_0 = 8.854\,\mathrm{pFm^{-1}}$) is the measure at constant strain and piezoelectric voltage coefficient d_{31} is the measure used in strain-charge form of the piezoelectric constitutive equation. However, the constitutive relation given in Eqs. (9) and (14) is expressed in terms of $\dot{\mathsf{E}}$ and $\dot{\tilde{\mathsf{E}}}$ as a function of S and $\tilde{\mathsf{D}}$.

The following relations are used to transform the values in the table to fit the formulation given in (9) and (14).

$$[\varepsilon^{\mathsf{S}}] = \left[\varepsilon^{\mathsf{E}}\right] + [d]\left[c^{\tilde{\mathsf{E}}}\right][d]^{\top} \tag{48a}$$

$$[\varepsilon^{\mathsf{S}}]^{-1} = \left(\left[\varepsilon^{\mathsf{E}}\right] + [d]\left[c^{\tilde{\mathsf{E}}}\right][d]^{\top}\right)^{-1}, \tag{48b}$$

where $\left[c^{\tilde{\mathsf{E}}}\right]$ is the Young's modulus Y_p of the piezoelectric material measured at constant electric field as given in table.

The compliance matrix at constant dielectric displacement is obtained from the table values as follows

$$\left[s^{\tilde{\mathsf{D}}}\right] = \left[c^{\tilde{\mathsf{E}}}\right]^{-1} - \left([d]^{\top}[\varepsilon^{\mathsf{S}}]^{-1}[d]\right). \tag{49}$$

The piezoelectric voltage coefficient derived from displacement coefficient is given by

$$[g] = \left[\varepsilon^{\mathsf{S}}\right]^{-1}[d]. \tag{50}$$

The substructure is discretized using a three dimensional mixed-hybrid space-time *structural element*, and the piezoelectric material is discretized using a three dimensional mixed-hybrid space-time *piezoelectric element*. The electrodes on the top and the bottom surfaces of the bimorph are discretized using a spatially two-dimensional space-time face element, and the harvester circuit is attached to the electrode elements by coupling them with a common virtual node. The length dimension of the bimorph is discretized using 15 elements, the width dimension with 2 elements. The substructure is discretized using 2 elements in the height dimension and the piezoelectric material with 3 elements. Since the substructure is sandwiched between two piezoelectric patches, the total number of elements in the height dimension of the bimorph is 8.

The parallel connection of the conductive electrodes is facilitated by having the same polarization direction for the top and bottom piezo elements. Physically this means that both the top and bottom surfaces of the bimorph constitute one terminal and the electrode layers present in the upper piezo-substructure interface and lower piezo-substructure interface constitute the other terminal. The interface terminal is grounded by setting the nodes of the piezoelectric element to zero. The potential on the top surface is equal to the potential of the bottom surface, and this potential drives the harvesting circuit represented by a resistor element.

In harvesting energy from base excitations, many studies focus on the excitation of the harvester at it's fundamental resonance frequency to investigate power output characteristics of the harvester. The first fundamental short circuit ($R = 0$) frequency of the piezoelectric bimorph considered in this study is 118.5 Hz. The bimorph is excited at this frequency to observe the power output and vibration characteristics.

Figure 8 presents the evolution of electrical power and electrical potential with time at two different resistances. Since it is impractical to have exactly zero resistance under experimental conditions, a resistance of $1\,\mathrm{k}\Omega$ is chosen to represent short circuit condition. The harvester reaches a steady state voltage of 4 V at short circuit condition compared to its steady stage voltage output of 6 V at $R = 10\,\mathrm{k}\Omega$. This behavior is expected as potential builds up in the electrodes when there is infinite resistance present between the terminals, and the electrical potential drops to zero when there is zero resistance present between the terminals. Moreover, the steady state power output of 4 mW is reasonable for the given base excitation. Figure 9 presents the evolution of relative tip displacement and base excitation with time. The results suggests that when a finite resistance is present between the terminals,

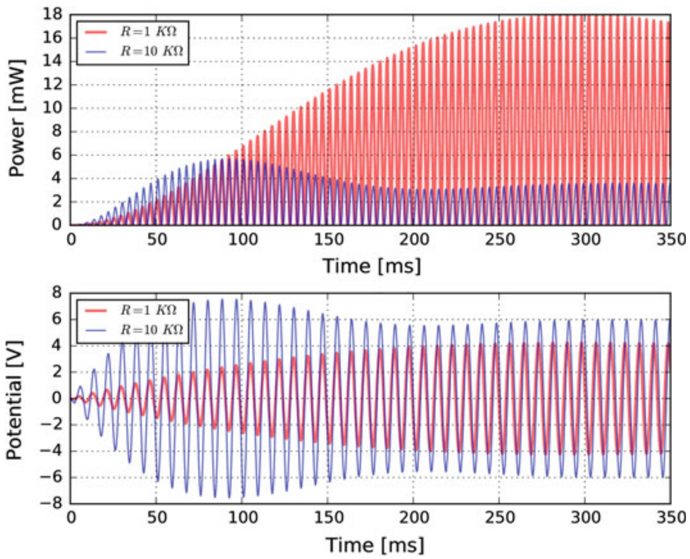


Fig. 8 Solution for electrical power (*top*) and electrical potential (*bottom*) with time

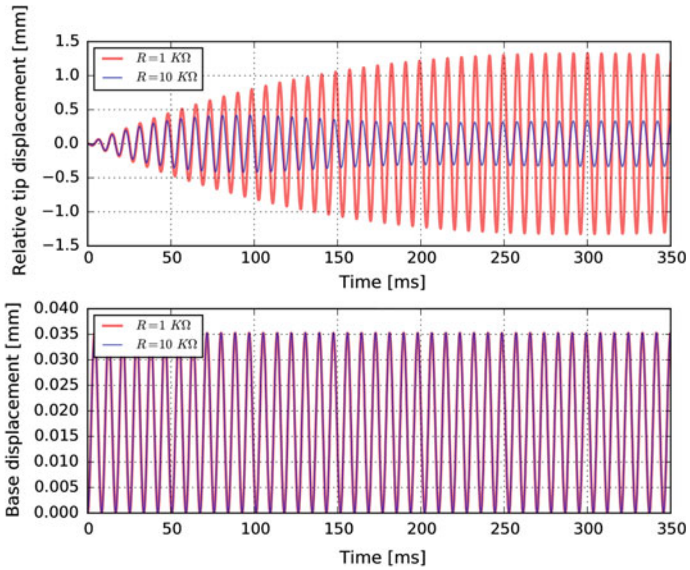


Fig. 9 Solution for tip velocity (*top*) and relative end displacement (*bottom*) with time

the displacement amplitudes are suppressed. This is due to the fact that the attached circuit has an impact on the vibration characteristics of the harvester, and this impact can be captured effectively in a strongly-coupled modeling approach as presented in this study.

Acknowledgments The authors gratefully acknowledge the support of the FNR AFR PhD grant 3996097 and the FP7 Marie Curie—Career Integration Grant 322151 by the European Commission.

References

1. S R Anton and H A Sodano. A review of power harvesting using piezoelectric materials (2003–2006). *Smart materials and structures*, 16:1–21, 2007.
2. H A Sodano, G Park, and D J Inman. Estimation of electric charge output for piezoelectric energy harvesting. *Strain*, 40:49–58, 2004.
3. J Baker, S Roundy, and P Wright. Alternative geometries for increasing power density in vibration energy scavenging for wireless sensor networks. *Proceedings of the 3rd international enery conversion conference*, pages 1–12, 2005.
4. H A Sodano, J Lloyd, and D J Inman. An experimental comparison between several active composite actuators for power generation. *Journal of Smart material structures*, 15:1211–1216, 2006.
5. Y Yang and L Tang. Equivalent circuit modeling of piezoelectric energy harvesters. *Journal of Intelligent material systems and structures*, 20:2223–2235, 2009.
6. N G Elvin and Alex A Elvin. A general equivalent circuit model for piezoelectric generators. *Journal of Intelligent material systems and structures*, 20:3–9, 2009.
7. D J Erturk, Aand Inman. On mechanical modeling of cantilevered piezoelectric vibration energy harvesters. *Journal of Intelligent material systems and structures*, 19:1311–1325, 2008.
8. A Erturk and D J Inman. An experimentally validated bimorph cantilever model for piezoelectric energy harvesting from base excitations. *Smart materials and structures*, 18:1–18, 2009.
9. O Thomas, J F Deü, and J Ducarne. Vibrations of an elastic structure with shunted piezoelectric patches: efficient finite element formulation and electromechanical coupling coefficients. *International Journal for Numerical methods in engineering*, 80:235–268, 2009.
10. H Allik and Thomas J R Hughes. Finite element method for piezoelectric vibration. *International Journal for Numerical methods in engineering*, 2:151–157, 1970.
11. A Benjeddou. Advances in piezoelectric finite element modeling of adaptive structural elements: a survey. *Computers and structures*, 76:347–363, 2000.
12. J S Yang. Equations for thick elastic plates with partially electroded piezoelectric actuators and higher order electric fields. *Smart materials and structures*, 8, 1999.
13. K Y Sze and Y S Pan. Hybrid finite element models for piezoelectric materials. *Journal of Sound and vibration*, 226:519–547, 1999.
14. C C Wu, K Y Sze, and Y Q Huang. Numerical solutions on fracture of piezoelectric materials by hybrid element. *International Journal of Solids and structures*, 38:4315–4329, 2001.
15. S Klinkel and W Wagner. A geometrically non-linear piezoelectric solid shell element based on a mixed multi-field variational formulation. *International Journal for Numerical methods in engineering*, 65:349–382, 2006.
16. S Klinkel, F Gruttmann, and W Wagner. A robust non-linear solid shell element based on a mixed variational formulation. *Computer methods in applied mechanics and engineering*, 195:179–201, 2006.
17. J J Allen and A J Smits. Energy harvesting eel. *Journal of Fluid and structures*, 15:629–640, 2001.
18. G W Taylor, J R Burns, S A Kammann, W B Powers, and T R Welsh. The energy harvesting eel: a small subsurface ocean/river power generator. *IEEE journal of oceanic engineering*, 26:539–547, 2001.
19. H D Akaydin, N Elvin, and Y Andreopoulos. Energy harvesting from highly unsteady fluid flows using piezoelectric materials. *Journal of Intelligent material systems and structures*, 21:1263–1278, 2010.

20. B Hübner and D Dinkler. A simultaneous solution procedure for strong interactions of generalized newtonian fluids and viscoelastic solids at large strains. *International Journal for Numerical methods in engineering*, 64:920–939, 2005.
21. A Zilian and A Legay. The enriched space-time finite element method (EST) for simultaneous solution of fluid-structure interaction. *International Journal for Numerical methods in engineering*, 75:305–334, 2008.
22. T J R Hughes and G M Hulbert. Space-time finite elmement methods for elastodynamics: formulations and error estimates. *Computer methods in applied mechanics and engineering*, 66:339–363, 1988.
23. G M Hulbert and T J R Hughes. Space-time finite element methods for second-order hyperbolic equations. *Computer methods in applied mechanics and engineering*, 84:327–348, 1990.
24. T E Tezduyar, S Sathe, R Keedy, and K Stein. Spacetime finite element techniques for computation of fluid-structure interactions. *Computer methods in applied mechanics and engineering*, 195:2002–2027, 2006.
25. E Walhorn, A Kölke, B Hübner, and D Dinkler. Fluid-structure coupling within a monolithic model involving free surface flows. *Computers and structures*, 83:2100–2111, 2005.
26. T H H Pian and Da-Peng Chen. Alternative ways for formulation of hybrid stress elements. *International Journal for Numerical methods in engineering*, 18:1679–1684, 1982.
27. T H H Pian and K Sumihara. Rational approach for assumed stress finite elements. *International Journal for Numerical methods in engineering*, 20:1685–1695, 1984.
28. T H H Pian. State-of-the-art development of hybrid/mixed finite element method. *Finite elements in analysis and design*, 21:5–20, 1995.
29. H J Lee, S Zhang, Y Bar-Cohen, and Stewart Sherrit. High temperature, high power piezoelectric composite transducers. *Sensors*, 14:14526–14552, 2014.

Comparison of Numerical Approaches to Bayesian Updating

Bojana Rosić, Jan Sýkora, Oliver Pajonk, Anna Kučerová and Hermann G. Matthies

Abstract This paper investigates the Bayesian process of identifying unknown model parameters given prior information and a set of noisy measurement data. There are two approaches being adopted in this research: one that uses the classical formula for measures and probability densities and one that leaves the underlying measure unchanged and updates the relevant random variable. The former is numerically tackled by a Markov chain Monte Carlo procedure based on the Metropolis-Hastings algorithm, whereas the latter is implemented via the ensemble/square root ensemble Kalman filters, as well as the functional approximation approaches in the form of the polynomial chaos based linear Bayesian filter and its corresponding square root algorithm. The study attempts to show the principal differences between full and linear Bayesian updates when a direct or a transformed version of measurements are taken into consideration. In this regard the comparison of both strategies is provided on the example of a steady state diffusion equation with nonlinear and transformed linear measurement operators.

Keywords Polynomial chaos based linear Bayesian update · Markov chain Monte Carlo · Square root algorithm · Ensemble Kalman filter

B. Rosić (✉) · H.G. Matthies
Institute of Scientific Computing, TU Braunschweig, 38106 Braunschweig, Germany
e-mail: bojana.rosic@tu-bs.de

J. Sýkora · A. Kučerová
Department of Mechanics, Faculty of Civil Engineering,
Czech Technical University in Prague, 166 29 Prague, Czech Republic

O. Pajonk
Elektrobit, Braunschweig, Germany

O. Pajonk
Schlumberger Information Solutions AS, Instituttveien 8, Kjeller, Norway

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_16

1 Introduction

In many applications one would like to predict the behaviour of a physical system with the help of a mathematical model whose parameters are not directly observable. In order to obtain the desired information about the system state or to allow for a prediction of a system response beyond observed localities, the model parameters are estimated from noisy data gathered by measuring some of the observable system responses. Most of the existing studies have tackled the issue by tuning the model parameters such that the distance between the observed and predicted system responses is minimised in a certain norm (e.g. [9, 31]). However, the resulting optimisation problem is often ill-posed—the minimised function is multimodal, non-smooth or non-differentiable—and hence a regularisation procedure [6] or a soft-computing-based method [13] is required. Yet, such a fitting-based approach provides only a one point estimate and hence omits the related uncertainties in measurements, imperfections of the numerical model as well as the preliminary knowledge about the material parameters arising from their physical occurrence. Unlike point-estimation techniques, the probabilistic concept transforms the prior expert-based probability description to a posterior via the incorporation of observations. From a Bayesian point of view this further means that the unknown parameters are taken to be uncertain and are modelled with the help of random variables (RVs)/fields (RFs), whose probability descriptions are coming from expert knowledge and the maximum entropy law [27]. This prior knowledge is then updated to a posterior distribution via Bayes's rule given in terms of conditional probabilities. In this regard, the process of assimilating more information obtained via experiments becomes well-posed. As a final outcome, the posterior distribution summarizes all available information about the model parameters such as the mean value, variance, probability of occurrence etc.

The primary computational challenge in Bayesian inference consist in extracting information from the posterior by evolving the probability measure. The Markov chain Monte Carlo (MCMC) method [10] is one of the most commonly used techniques for this kind of parameter estimation. In MCMC methods, the Markov chain is constructed such that the asymptotic distribution of the chain is the Bayesian posterior distribution. The posterior is sampled by letting the Markov chain run for a sufficiently long time. With the intention of accelerating the MCMC method some authors (e.g. [14–16, 29]) have introduced stochastic spectral methods into the computation. Expanding the prior random process into a polynomial chaos (PCE) or a Karhunen-Loève expansion (KLE), the inverse problem becomes an inference on the weights of the KLE or PCE coefficients. Another solution is to combine polynomial chaos theory with the maximum likelihood estimation and to calculate the parameter estimates in a recursive manner (see [20]), or to apply a local linearisation of the forward model to improve the acceptance probability of proposed moves [5]. However, the previously mentioned methods are all based on pure sampling procedures, or a combination of spectral approximations and MCMC. Therefore, they are slowly convergent and often computationally infeasible especially when one deals with large-scale problems.

To mitigate this issue, the authors of [18, 19, 21, 22] constructed a more efficient approach based on conditional expectation, which is an equivalent way to formulate the Bayesian update. The conditional expectation can be approximated by linear or higher order maps, which have to be found during the updating. In this way the Bayesian update (BU) is an algebraic formula, which can be computed in a purely analytical way as indicated in [18, 19, 21, 22]. In a simpler version, this idea appeared independently in [2], whereas in [23] it appears as a variant of the Kalman filter [12].

The aim of this study is to investigate the differences between full and linear Bayesian updates on a simple linear diffusion model with one uncertain parameter. For this purpose two scenarios are considered: one that features a direct—classical Bayesian update—and a second one that introduces a transformed measurement operator—a transformed Bayesian update. The transformation is studied from the mathematical and numerical point of view in order to better understand the linear Bayesian update. In addition, the paper proposes the use of the proxy modelling in the assimilation process in order to reduce the required computational time.

The paper is organised in the following way: Sect. 2 gives a short introduction into the model problem for which the Bayesian inference is presented in Sect. 3. Section 4 gives a brief overview of the available methods for the computation of a Bayesian update. The methods are analysed and compared in Sect. 5. Finally, the paper is concluded in Sect. 6.

2 Model Problem

The problem considered in this paper is a steady state heat transfer expressed by an energy balance equation

$$-\operatorname{div}(\kappa \nabla u(x)) = f(x), \quad \forall x \in \mathscr{G} \subset \mathrm{d}R^2, \tag{1}$$

in which the scalar conductivity coefficient $\kappa \in \mathscr{K}$ together with the loading $f(x)$ and initial conditions u_0 determine the system response u. In computational practice u is very often evaluated assuming that the conductivity parameter κ is known. This process is called the prediction of the system response, or the forward problem. However, the aim of this paper is to not to determine u but κ given the set of observation data z—the measurements of the system response in a few points of the physical domain of consideration.

Usually, the measurement set is mathematically formalised by an observation operator H, which relates the complete model response $u \in \mathscr{U}$ to an observation y in some vector space $\mathscr{Y}$ [21, 22]:

$$H : (\kappa, u(x)) \mapsto y(x) = H(\kappa; u(x)) \in \mathscr{Y}. \tag{2}$$

Since the modelled values y differ from the real data set z, the previous equation transforms to

$$y = z - \varepsilon = H(\kappa; u(x)), \tag{3}$$

in which the variable ε subsumes both model imperfections and the measurement error. A key aspect of this is that one may try to compute the non-observable thermal conductivity κ given z from Eq. (3). This further agrees with the inversion of the operator H, which in general may not be invertible or has a non-continuous inverse—the ill-posed problem. To ensure the existence and uniqueness of the solution, the previous issue can be resolved by, for example, taking the additional information into consideration. In a Bayesian point of view this corresponds to the prescription of a prior distribution on the model parameter κ.

3 Identification via Bayesian Regularisation

The Bayesian inference treats the problem in Eq. (3) by acquiring additional knowledge on the parameter set next to the observation data. Such a description has two ingredients, the measurable function or random variable, and the probability measure. One group of methods updates the measure—the classical Bayesian updating [10], the other group changes the function—the linear Bayesian updating [22]. In this section we show connection between these methods as well as give their short description.

The prior information on κ comes from expert knowledge about its realistic values and can be modelled in a form of a prior probability density function $p(\kappa)$ with the help of the maximum entropy approach [27]. In this manner, κ in Eq. (3) can be described by the finite-variance $\mathscr{K}$-valued RV/ RF

$$\kappa(\omega) : \Omega \to \mathscr{K} \tag{4}$$

on a probability space $S := L_2(\Omega, \mathfrak{A}, \mathbb{P})$. Here, Ω denotes the space of elementary events ω, $\mathfrak{A}$ is the σ-algebra and $\mathbb{P}$ stands for the probability measure.

Once the prior is chosen, the posterior density can be obtained with the help of classical Bayes's rule:

$$\pi_a(\kappa) := p(\kappa|z) = \frac{p(z|\kappa)}{p(z)} p(\kappa) \propto L(\kappa)p(\kappa) \tag{5}$$

given in terms of the conditional probability density functions. Here, $\pi_a(\kappa)$ and $p(\kappa)$ stand for the posterior and prior density functions of κ, and $L(\kappa)$ is the likelihood giving a measure of how good the model is explaining the data z.

The law described in Eq. (5) is not the one used further. Namely, the random variable κ is restricted to the positive cone in the vector space, and hence requires a transformation, see [22]. By defining the bijective differentiable mapping

$$T_q : \quad \mathscr{K} \to \mathscr{Q} \tag{6}$$

from the model $\mathscr{K}$ to the assimilation $\mathscr{Q}$ space, κ is transformed to a random variable

$$q = T_q(\kappa) \tag{7}$$

which lives in a vector space. As a consequence, Bayes's rule in Eq. (5) obtains the form

$$\pi_q(q) = \frac{p(z|q)}{p(z)} p_q(q), \tag{8}$$

where $p_q(q)$ and $\pi_q(q)$ are the prior and posterior density of q, respectively. Once the assimilation is performed, the back-transformation to the model space is applied such that

$$\pi_a(\kappa) = \pi_q(q) \frac{dT_q(\kappa)}{d\kappa} \tag{9}$$

holds, where $dT_q(\kappa)/d\kappa$ denotes the Radon-Nikodým derivative of the assimilation measure with respect to the original measure. In this manner the process of computing $\pi_a(\kappa)$ is equivalent to the problem of evaluating the likelihood function $L(q) = p(z|q)$. The likelihood is incorporating the information from the data into the updating process, and hence, it is shaped by the measurement density [28]. By assuming normally distributed measurements, the likelihood takes the form

$$L(q) = \exp\left(-\frac{1}{2}(d)^{\mathrm{T}} C_\varepsilon^{-1}(d)\right), \tag{10}$$

in which d denotes the difference

$$d = z - y \tag{11}$$

between the forecast $y = Y(q, u)$ and the measurement z, whereas C_ε stands for the measurement covariance.

Following this, the evaluation of Eq. (8) corresponds to the simulation of the forward problem

$$-\operatorname{div}(T_q(q)\nabla u(x)) = f(x), \tag{12}$$

and the response forecast

$$y(x, \omega) = Y(q(\omega); u(x, \omega)). \tag{13}$$

Here, Y denotes the observation operator with respect to the transformed parameter q such that

$$z = H(\kappa) = Y(q), \quad Y = H \circ T_q^{-1} \tag{14}$$

holds.

Note that in a special case when the random variable κ follows a lognormal distribution the transformation in Eq. (7) coincides with the Gaussian anamorphosis

[24, 25]. Indeed, by defining q as $\log(\kappa)$, the non-Guassian RV κ transforms to the Gaussian RV q.

The transformations previously mentioned are especially important when the second version of Bayes's rule is applied—the one that updates the measurable function. This alternative formulation of Bayes's rule can be achieved by expressing the conditional probabilities in Eq. (8) in terms of conditional expectation. Following the mathematical derivation in [21, 22], this approach boils down to a quadratic minimisation problem:

$$q_a(\omega) = P_{\mathscr{Q}_{sn}} q = \underset{\eta \in \mathscr{Q}_{sn}}{\arg\min} \; \|q - \eta\|^2_{L_2}, \tag{15}$$

in which $P_{\mathscr{Q}_{sn}}$ is the orthogonal projection operator of q onto the space of the new information $\mathscr{Q}_{sn} := \mathscr{Q} \otimes S_n$. This space of $\mathscr{Q}$-valued random variables with finite variance is defined by the triplet $S_n := L_2(\Omega, \mathfrak{S}, \mathbb{P})$, where $\mathfrak{S} := \sigma(Y)$ denotes the sub-σ-algebra generated by Y. According to the Doob-Dynkin lemma [3], one may state that $\eta := \phi \circ Y \circ q$, in which ϕ belongs to the space $L_0(\mathscr{Y}, \mathscr{Q})$ of measurable maps. Constraining the vector space $L_0(\mathscr{Y}, \mathscr{Q})$ to the subspace of linear maps $\mathscr{L}(\mathscr{Y}, \mathscr{Q})$, the minimisation problem in Eq. (15) leads to a unique solution K. This gives an affine approximation of Eq. (15)

$$q_a(\omega) = q_f(\omega) + K(z(\omega) - y_f(\omega)), \tag{16}$$

also known as a linear Bayesian posterior estimate. Here, q_f represents the prior random variable, q_a is the posterior, y_f is the forecasted measurement and K represents the very well-known Kalman gain

$$K := C_{q_f y_f} \left(C_{y_f} + C_\varepsilon\right)^{-1} \tag{17}$$

which can be easily evaluated if the appropriate covariance matrices $C_{q_f y_f}$, C_{y_f} and C_ε are known. We would like to emphasise that the Hilbert-space setting of $\mathscr{Q}$ and $\mathscr{Y}$ has made the formulation in Eq. (16) possible [12]. Therefore, the transformation in Eq. (7) was necessary.

It is interesting to note that the projection in Eq. (16) is performed over a smaller space than $\mathscr{Q}_{sn}$. An implication of this is that available information is not completely used in the process of updating. It is therefore likely that the minimisation error remains larger. However, the computation of the projection becomes simpler. Another advantage of Eq. (16) compared to Eq. (8) is that the inference in Eq. (16) is given in terms of RVs instead of conditional densities. Namely, $q_a(\omega)$, $q_f(\omega)$, $z(\omega)$ and $y_f(\omega)$ denote the RVs used to model the posterior, prior, observation and forecasted observation, respectively.

Having in mind that Y is in general nonlinear, one may alter the estimation in Eq. (16) by transforming the measurement to the linear one. In other words, one may apply the nonlinear transformation

$$z_t = T_z(z) = (T_z \circ H \circ T_q^{-1})(q) = G(q) \tag{18}$$

such that the transformed measurement z_t is linear in q. In our example in Eq. (1) one has that

$$\kappa \sim 1/z. \tag{19}$$

Furthermore, if κ follows lognormal distribution

$$\kappa \sim \exp(q), \tag{20}$$

the transformation in Eq. (18) reads

$$z_t = -\log(z) \sim q. \tag{21}$$

The last relation coincides with the Gaussian anamorphosis because the non-Gaussian z is transformed to a standard Gaussian z_t. In this example the transformation is easy to achieve as it is purely algebraic. When an algebraic transformation is not possible, one may apply empirical anamorphosis function as shown in [24]. Note that for the multivariate case, the transformation has to be applied to each of random variables individually (locally).

After the measurement transformation, Bayes's rule assimilates the measurement data z_t with the prior information q by means of the following formula:

$$p(q|z_t) = \frac{p(z_t|q)}{p(z_t)} p(q). \tag{22}$$

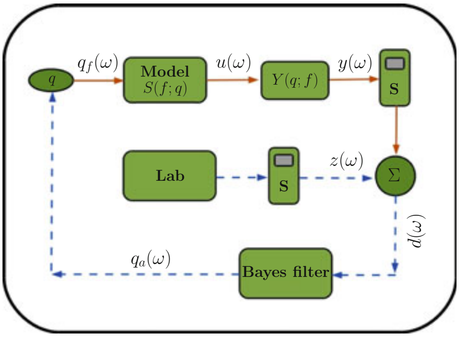
For a Gaussian measurement error, this means evaluation of the likelihood function

$$L(q) = \exp\left(-\frac{1}{2}(T_z(z) - T_z(y))^{\mathrm{T}} C_{\varepsilon_t}^{-1}(T_z(z) - T_z(y))\right), \tag{23}$$

in which the distance d between the measurement z and the forecast y is transformed, as well as the covariance function from C_ε to C_{ε_t}. Hence, by transforming observations, the measurement errors also transform. Such a transformation can lead to an overestimation of the measurement error with respect to the transformed forecast error as observed in [25]. Therefore, one often advises to compute the variance of the measurement error in the assimilation space directly from the transformed measurements.

The linear Bayes's rule corresponding to Eq. (22) is characterised by a solution belonging to $\mathscr{Q} \otimes S_n$, in which $S_n := L_2(\Omega, \mathfrak{S}, \mathbb{P})$, where $\mathfrak{S} := \sigma(G)$ denotes the sub-σ-algebra generated by G. In our case G describes the linear relation between q and z_t, and hence the linear Bayes's rule in Eq. (16) becomes optimal. However, note that the transformations introduce additional errors into the computation process. This is especially the case for the measurement errors, as already explained.

Fig. 1 Schematic representation of Bayesian approach to identification



The schematic representation of the Bayesian inference is shown in Fig. 1. The scheme describes the closed loop of one Bayesian update. The loop starts by assuming the prior distribution $q_f(\omega)$, which is then propagated through the model $S(f;q)$ and the measurement operator Y to the forecasted (predicted) measurement $y(\omega)$ read by sensor S. The prediction is then subtracted from noisy data $z(\omega)$ coming from real experiments, and the resulting difference is forwarded to the Bayesian filter, which further produces the posterior distribution $q_a(\omega)$, i.e. the updated value of $q_f(\omega)$.

4 Computational Approaches

In recent years there has been an increasing amount of literature on computational approaches related to Bayesian inference. However, this paper reviews only the research conducted on MCMC [10, 15] and linear Bayesian filters [2, 19, 21]. The main aim of this work is to contrast the linear Bayesian methods to a full MCMC approach on a numerical example.

4.1 Markov Chain Monte Carlo

Markov chain Monte Carlo is a sampling procedure used for the estimation of the posterior probability density function via Eq. (5) or Eq. (8), respectively. The method is very general as it does not require any model approximations in contrast to those further described. Instead, MCMC constructs a Markov chain with the posterior as an equilibrium distribution. The two most often used types of this algorithm are: the Gibbs sampling technique [26] and the Metropolis-Hastings algorithm [4], the one used in this paper.

```
procedure MHA($p(q)$, $g_k$, $L(q)$, $N$)
draw initial value $q^{(0)}$ from prior $p(q)$
for each $i = 1 \to N$ do
    draw $q^{(*)}$ from proposal distribution
        $g_k(q^{(*)}|q^{(i-1)})$
    evaluate the probability of acceptance
        $r = \min\{1, \frac{\pi_a(q^{(*)}) g_k(q^{(i-1)}|q^{(*)})}{\pi_a(q^{(i-1)}) g_k(q^{(*)}|q^{(i-1)})}\}$
    accept the next state with probability $r$
        $q^{(i)} = q^{(*)}$
    or reject with probability $1 - r$
        $q^{(i)} = q^{(i-1)}$
    end
end procedure
```

Algorithm 1: Metropolis-Hastings procedure

As shown in Algorithm 1, the Metropolis scheme generates a sequence of samples (states) $q^{(i)}$ whose values solely depend on the previous sample in the chain. The value of the new state $q^{(i)}$ is generated with the help of the proposal distribution $g_k(q|q^{(i-1)})$, also known as the transition kernel. In practice the transition kernel is very often chosen to be symmetric such that $g_k(q^{(*)}|q^{(i)}) = g_k(q^{(i)}|q^{(*)})$. A typical example of such a kernel is the normal distribution centered at the previous state—the random walk chain. However, other types of kernels can be used in a similar manner, e.g. independence chains, rejection sampling chains, auto-regressive chains or grid-based chains, see [30]. Once the new sample is drawn it is either accepted with the probability r or rejected with probability $1 - r$.

The advantage of the previous algorithm is that it does not need target probabilities but only ratios of target probabilities to work. In this manner the computation of the normalisation constant in Eq. (8) is avoided. However, the samples are not independent any more as they are drawn from the proposal distribution with the probability r. Even though this is the case, the obtained samples can still be used for the evaluation of integrals in a Monte Carlo fashion. However, this works only if the Markov chain is aperiodic, irreducible, and positive recurrent [32]. Under the previously mentioned regularity conditions the sampling sequence $q^{(i)}$ converges in distribution to our target posterior distribution (Theorem 3 in [30]) regardless of the starting point. However, the speed of the convergence greatly depends on the initial choice. Due to this, a few starting samples have to be excluded from the chain — a so-called “burn in period”.

Computationally, the Markov chain Monte Carlo algorithm is a very demanding procedure because one has to evaluate the system response for each new proposed sample, see Fig. 2. The evaluation starts with samples that are used as input in the deterministic model $S(f, q)$, which closely describes the system response u as a function of the external loading f and the input parameters q. Once the response is computed, the measurement operator Y is applied and the value of the observable quantity y is diagnosed by a sensor. The predicted measurements are then compared to the real data, which further results in the distance measure d entering the likelihood

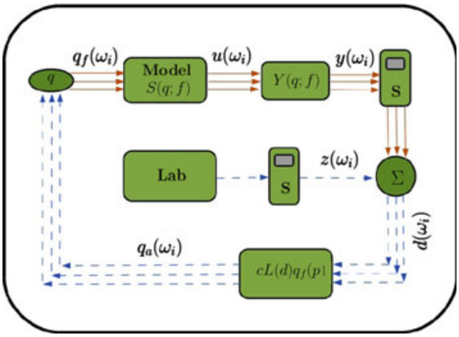


Fig. 2 The algorithmic scheme of an inverse problem solved by MCMC filtering

function, see Eq. (10). In this manner a posterior sample is obtained, and the process is repeated all over again for the next sample.

4.2 *Proxy Modelling*

To speed up the assimilation process one may introduce a proxy model for the forecasted measurement. Usually, the proxy model is made with the help of a functional approximation of random variables/fields entering the process, and a stochastic Galerkin procedure [22]. To this end, both the predicted system response u_f and observation y_f can be represented in a polynomial chaos expansion form [11, 17, 33, 35]

$$\hat{u}_f(\omega) = \sum_{\mathscr{J}} u_f^{(\alpha)} H_\alpha(\omega),$$
$$\hat{y}_f(\omega) = \sum_{\mathscr{J}} y_f^{(\alpha)} H_\alpha(\omega), \tag{24}$$

in which $H_\alpha(\omega)$ represent the generalised orthogonal polynomials and $\mathscr{J}$ stands for the set of all finite non-negative integer sequences, i.e. multi-indices α such that

$$\mathscr{J} \; : = \{\alpha = (\alpha_1, ..., \alpha_j, ...) \;\; | \;\; \alpha_j \in \mathbf{N}_0, \quad |\alpha| := \sum_{k=1}^{\infty} \alpha_j < \infty\} \tag{25}$$

holds. Due to computational reasons, only a finite subset of $\mathscr{J}$ is taken, i.e. the expansion in Eq. (24) is truncated to a finite number of terms. This results in another type of error which has to be added to the modelling error mentioned above.

To increase its efficiency, the MCMC cycle can be modified such that the forward model is substituted with a less accurate but computationally cheaper proxy model (see Fig. 3), as already reported in [14]. In this manner the forward model is not

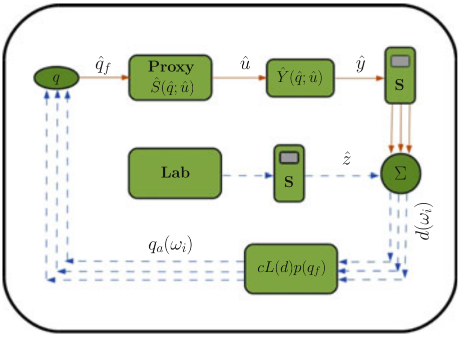


Fig. 3 The algorithmic scheme of an inverse problem solved by proxy MCMC filtering

individually solved for each MCMC sample, but apriori, see Fig. 3. After evaluating the functional approximation of the measurement, the sampling occurs and the update loop proceeds in the same manner as described previously, see Figs. 2 and 3.

4.3 Linear Bayesian Inference

The advantage of the methods described in the previous section is that they are model-independent. However, their main drawback is the slow convergence. The issue of high computational cost can be improved via recently developed Bayesian linear methods [18, 19, 21, 22] as shown in Eq. (16). Recalling that the RV $q_a(\omega)$ can be numerically represented by either sampling $q_a(\omega_i)$ or the functional approximation such as polynomial chaos expansion, one may distinguish at least two numerical approaches to the problem given in Eq. (16): the ensemble Kalman filter [7] and the polynomial chaos based update [21, 22].

4.3.1 Ensemble Kalman Filter

The simplest way to numerically estimate q_a is to sample Eq. (16) in a Monte Carlo fashion. Such a procedure starts by building ensembles of prior samples $\boldsymbol{Q}_f := [q_f(\omega_1), \ldots, q_f(\omega_Z)]$, forecasts $\boldsymbol{Y}_f := [y_f(\omega_1), \ldots, y_f(\omega_Z)]$ and measurements $\boldsymbol{Z}_m$, such that Eq. (16) can be formulated in a matrix notation as

$$\boldsymbol{Q}_a = \boldsymbol{Q}_f + \boldsymbol{K}(\boldsymbol{Z}_m - \boldsymbol{Y}_f), \tag{26}$$

in which $\boldsymbol{K}$ takes the form as in Eq. (16). Note that its corresponding covariances may be estimated from the ensemble, i.e.

$$\boldsymbol{C}_{q_f,y_f} \approx \frac{1}{Z-1}\tilde{\boldsymbol{Q}}_f\tilde{\boldsymbol{Y}}_f^T \quad \text{and} \quad \boldsymbol{C}_{y_f} \approx \frac{1}{Z-1}\tilde{\boldsymbol{Y}}_f\tilde{\boldsymbol{Y}}_f^T. \tag{27}$$

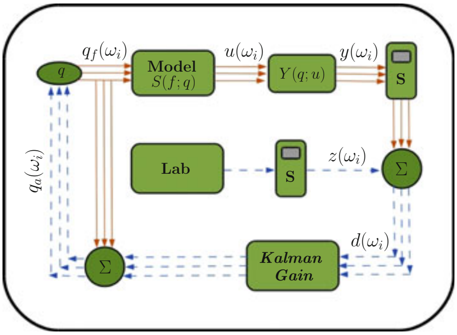


Fig. 4 The algorithmic scheme of an inverse problem solved by the ensemble Kalman filter

Here, $\tilde{\boldsymbol{Q}}_f = \boldsymbol{Q}_f - \bar{q}_f\,\boldsymbol{1}_Z^T$ and $\tilde{\boldsymbol{Y}}_f = \boldsymbol{Y}_f - \bar{y}_f\,\boldsymbol{1}_Z^T$ represent the fluctuation parts of corresponding RVs, in which $\bar{q}_f = \frac{1}{Z}\sum_{z=1}^{Z} q_f(\omega_z)$ and $\bar{y}_f = \frac{1}{Z}\sum_{z=1}^{Z} y_f(\omega_z)$ are the estimated means and $\boldsymbol{1}_Z$ is a vector of ones of size Z.

This method is a Monte Carlo method, hence it also suffers from the slow convergence with increasing Z. On the other hand, it is fairly simple to implement: all it needs are random samples, see Fig. 4. In practice the number of samples is often low, and then special care is needed when computing the covariances and the Kalman gain $\boldsymbol{K}$, see [8]. To reduce the computation time one may use the proxy model instead of a forward simulator in a similar way as it is done in the MCMC procedure. In this manner only the update formula in Eq. (16) is sampled.

4.3.2 Polynomial Chaos Based Linear Bayesian Update

To avoid the sampling procedure presented previously in a form of the ensemble Kalman filter (EnKF) algorithm, one may use the opportunity to functionally approximate the random variables (fields) in Eq. (16). In this light the linear Bayesian procedure can be reduced to a simple algebraic method. Starting from the functional representation of the prior

$$\hat{q}_f = \sum_{\alpha} q_f^{(\alpha)} H_\alpha(\omega) \tag{28}$$

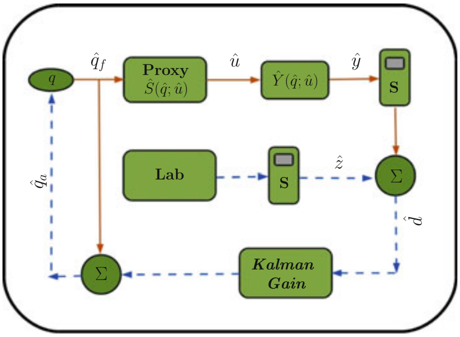
and the proxy in Eq. (24), one may discretise Eq. (16) as:

$$\hat{q}_a = \hat{q}_f + K\big(\hat{z} - \hat{\boldsymbol{y}}_f\big), \tag{29}$$

where $\hat{z} \in \mathbb{R}^{L\times Z}$ is the PCE of the measurement. Here, K in Eq. (29) is the Kalman gain evaluated in an algebraic way knowing that

$$C_{q_f,y_f} = \sum_{\alpha>0} \alpha!\, q_f^{(\alpha)} (\boldsymbol{y}_f^{(\alpha)})^T. \tag{30}$$

Fig. 5 The polynomial chaos based linear Bayesian scheme



Note that in the numerical computation $\hat{q}_a \in \mathbb{R}^Z$, $\hat{q}_f \in \mathbb{R}^Z$, $\hat{\boldsymbol{y}}_f \in \mathbb{R}^{L\times Z}$ and $\hat{\boldsymbol{z}} \in \mathbb{R}^{L\times Z}$ are PCEs with cardinality Z determined by $(L+1)$ RVs and polynomial order p. Here, the number $(L+1)$ subsumes all the RVs describing the prior and the RVs $\{\theta_i\}_{i=1}^{L}$ used to model the measurement error ε.

The previous algorithm is shown in Fig. 5 where the whole update process can be represented by only one loop.

4.3.3 Square Root Polynomial Chaos Based Linear Bayesian Update

The idea of linear Bayesian inference is allowing the computation of the posterior in a quite efficient way, however the update requires the introduction of additional RVs—corresponding to ε—into the update process. This essentially may enlarge the dimension of the stochastic space one is working with, especially in case of sequential updating (for an illustration see [18, 22]). To avoid the presence of the observation RVs and corresponding PCE in Eq. (16), one may follow the idea of a square root filter [1], as the authors already addressed in [19]. In such an algorithm the evaluation of the posterior consists of two phases:

1. the estimation of the posterior mean via

$$\mathbb{E}(\hat{q}_a) = \mathbb{E}(\hat{q}_f) + K(\mathbb{E}(\hat{\boldsymbol{z}}) - \mathbb{E}(\boldsymbol{y}_f)), \tag{31}$$

2. the prediction of the varying part $\tilde{q}_a := \hat{q}_a - \mathbb{E}(\hat{q}_a)$ via

$$\tilde{q}_a = \frac{S_{q_a}}{\sqrt{\Delta}}. \tag{32}$$

Here, $\sqrt{\Delta} = \text{diag}\,(\sqrt{\alpha}!)$ stands for the square root of the Gram matrix $\Delta{=}\mathbb{E}(H_\alpha H_\beta)$, while S_{q_a} denotes the matrix square root of the posterior covariance $C_{q_a} = S_{q_a} S_{q_a}^T$. The latter one represents a linear transformation of the prior matrix square root S_{q_f}, i.e.

$$S_{q_a} = S_{q_f} V \sqrt{I - \Sigma^T \Sigma} V^T. \tag{33}$$

The transformation essentially comes from the definition of the covariance structure $(C_{y_f} + C_\varepsilon)$ and its decomposition

$$\left(C_{y_f} + C_\varepsilon\right) = B\Lambda B^T, \tag{34}$$

where B^T rotates the simulated measurements into directions aligned with the covariance structure $(C_{y_f} + C_\varepsilon)$, while $\Lambda^{-\frac{1}{2}}$ weights them accordingly. In this light matrices V and Σ are obtained by the singular value decomposition of

$$W = \Lambda^{-\frac{1}{2}} B^T H S_{q_f}, \tag{35}$$

for more details on mathematical derivation please see [19].

Equation (34) is exactly the place where the additional information (in the form of C_ε) enters the update—C_ε specifies directions and magnitude of the uncertainty (variance) reduction induced by the observation. Thus, no additional random variables have to be included into the update. However, note that the square-root formulation is only equivalent to the standard linear filter form in case of Gaussianity. In addition, if the random variable is represented by an ensemble, the previous algorithm is of the ensemble square root type.

5 Numerical Results

To test the numerical procedures described previously, two benchmark problems are introduced: a two point boundary value problem in one spatial dimension and a heat conduction problem for a rectangular plate.

5.1 One Dimensional Heat Problem

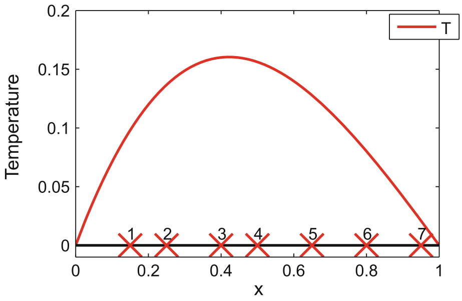
The thin metal rod of unit length is exposed to the deterministic heat source $f(x) = 5(1 - x)$ linearly dependent on the spatial coordinate x. The heat transfer in the rod is assumed to be steady, i.e. described by the one-dimensional heat equation:

$$-\operatorname{div} \kappa(\omega) u(x,\omega) = f(x,\omega), \tag{36}$$

with zero Dirichlet boundary conditions. Here, $\kappa(x,\omega)$ denotes the conductivity coefficient one is uncertain about and $u(x,\omega)$ is the temperature response.

The true value of the thermal conductivity κ_t is taken to be one realisation of a lognormal random variable described independently from the a priori distribution—so called truth. The corresponding data set-observations are then obtained with the help of the deterministic finite-difference (FD) approach. The temperature is esti-

Fig. 6 The virtual measurement: the temperature dependence on the point position x. The *red crosses* represent the sensor placements



mated on a uniform mesh of 41 points, from which $L = 7$ randomly chosen nodes are used to place the measurement sensors, see Fig. 6. Each measurement is subjected to Gaussian noise with zero mean and covariance $C_\varepsilon = \sigma_\varepsilon^2 I$ (I is the identity matrix of size L).

For the purpose of measurement prediction, the prior conductivity $\kappa_f(\omega)$ is assumed to be a lognormally distributed RV with mean 2.3 and standard deviation 0.3. Once the prior is chosen, the predicted measurement is evaluated on the uniform mesh of 21 spatial points with the help of the stochastic Galerkin approach, see Eq. (36). Note that the spatial mesh is taken to be different than the one chosen for the computation of the "virtual truth". This is done in order to avoid the "inverse crime" problem [34].

As depicted in Fig. 7a, the "virtual truth" is taken to lie in

- C_1: high probability ($\kappa_t = 2$),
- C_2: 2σ ($\kappa_t = 1.7$),
- and C_3: low probability ($\kappa_t = 1.4$)

regions of the prior. While the first case scenario (i.e. when $\kappa_t = 2$) represents a reliable assumption of the prior, the other two case scenarios are describing situations in which one cannot have precise expert knowledge on the value of the parameter κ.

Following this, Fig. 7b compares the predicted measurement and the observation including the measurement noise for each of the previously mentioned scenarios. These results argue how much of an impact the prior distribution has on the distance between the measured and observed data sets. Namely, if the prior is such that the truth lies in the high probability region then the distance is small, and vice versa.

5.1.1 Nonlinear Measurement

The experimental set up as described previously introduces the temperature observations into the identification process. However, one may note that this kind of measurement is inconsistent with the statement made in Sect. 3 where the linear dependence between the measurement and the parameter has been assumed. According to Fig. 8, this hypothesis does not hold. Even though this is the case, the objectives of the

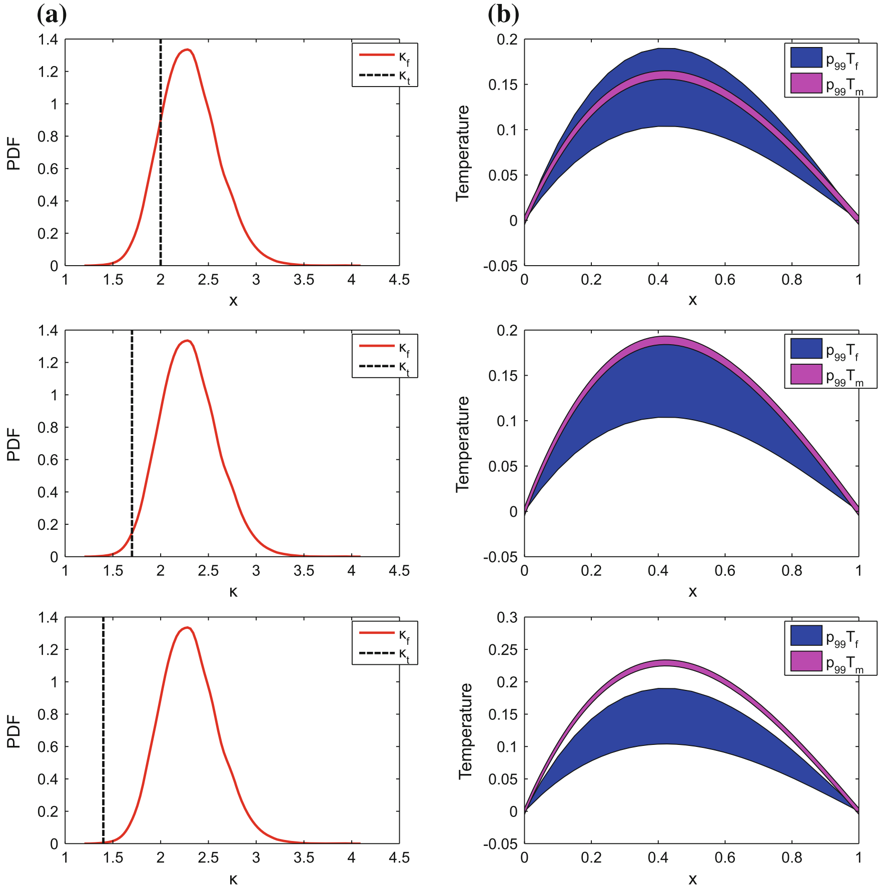


Fig. 7 **a** Three case scenarios for the true value of the conductivity coefficient κ, **b** the predicted versus the virtual measurement: 99 % probability regions of the predicted temperature T_f and 99 % probability regions of the measurement T_m

following numerical computations are to determine whether and up to which degree the truth can be identified from the nonlinear measurement data.

The six identification procedures—one random variable based (reduced) linear Bayesian update (RLBU),[1] full linear Bayesian update (FLBU), the square root update (SQRT), ensemble Kalman filter (EnKF) with 1000 samples, square root ensemble Kalman filter (EnKFS) with 1000 samples, and the full Bayesian MCMC update with 10^5 samples—are used to estimate the value of the parameter κ given seven temperature observations. The update process is performed only once using the complete measurement data. The results obtained, as shown in Tables 1 and 2, indicate that the MCMC procedure is the only one which can identify the truth in all

[1] Random variables describing the measurement error are not taken into consideration—they are projected out during the update process.

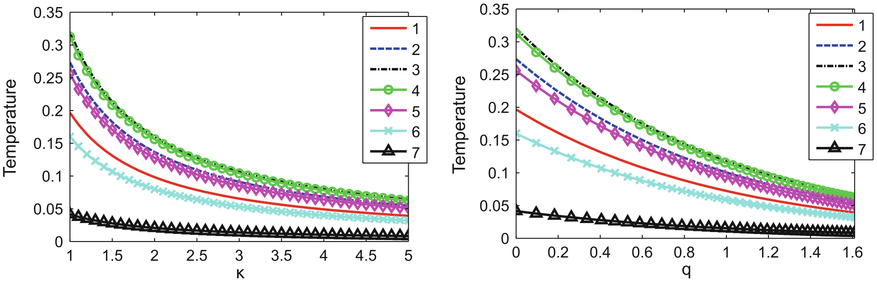


Fig. 8 The dependence of temperature on the parameter set κ (*left figure*) and its transformation q (*right figure*)

Table 1 Comparison of modes and standard deviations for the posterior κ obtained by different update procedures

Parameter method		Mode			Std		
	Case:	C_1	C_2	C_3	C_1	C_2	C_3
κ	Truth	2	1.7	1.4	0	0	0
	MCMC	1.9979	1.6952	1.4102	0.0131	0.0123	0.0163
	RLBU	1.9934	1.6338	1.2294	0.0117	0.0096	0.0072
	FLBU	1.9976	1.6373	1.2314	0.0184	0.0151	0.0113
	SQRT	1.9814	1.6236	1.2224	0.0274	0.0225	0.0169
	EnKF	1.9994	1.6328	1.2341	0.0262	0.0202	0.0158
	EnKFS	1.9859	1.6237	1.2311	0.0262	0.0207	0.0160

Table 2 Comparison of modes and standard deviations for the posterior q obtained by different update procedures

Parameter method		Mode			Std		
	Case:	C_1	C_2	C_3	C_1	C_2	C_3
q	Truth	0.6931	0.5306	0.3365	0	0	0
	MCMC	0.6921	0.5278	0.3437	0.0066	0.0069	0.0098
	RLBU	0.6899	0.4909	0.2065	0.0058	0.0058	0.0058
	FLBU	0.6928	0.4935	0.2095	0.0092	0.0092	0.0092
	SQRT	0.6839	0.4849	0.2004	0.0134	0.0134	0.0134
	EnKF	0.6910	0.4928	0.2064	0.0119	0.0126	0.0124
	EnKFS	0.6858	0.4892	0.2010	0.0121	0.0126	0.0120

three case scenarios. In contrast to this, the linear approximants are able to estimate the truth only in the first case scenario although with an overestimated standard deviation. The overestimation appears to be stronger in case of the square root posterior, as well as posteriors obtained from the ensemble data (EnKF-kind of procedures). Since the square root estimation is not equivalent to the linear Bayesian; and since the ensemble Kalman filter estimates strongly depend on the chosen seed (here 1000

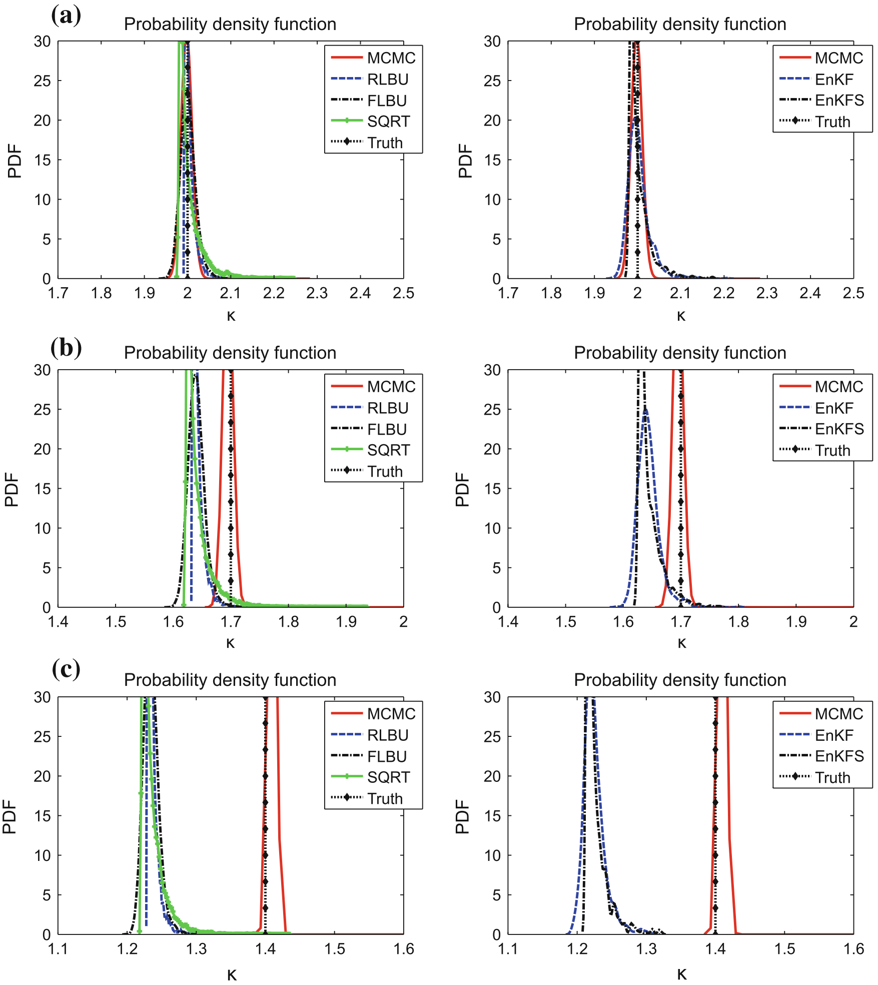


Fig. 9 Comparison of posterior probability density functions describing κ. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

samples), this finding was expected. Contrary to the expectations, the one random variable linear Bayesian update is underestimating the posterior variance. It seems that the underestimation happens due to constraints put on the basis on which the posterior is projected. Namely, one random variable linear Bayesian update is neglecting (projecting out) the additional random variables coming from the measurement data in the process of updating.

These findings are consistent with the plots of posterior probability density functions given in Figs. 9 and 10, where one may clearly notice that the MCMC procedure

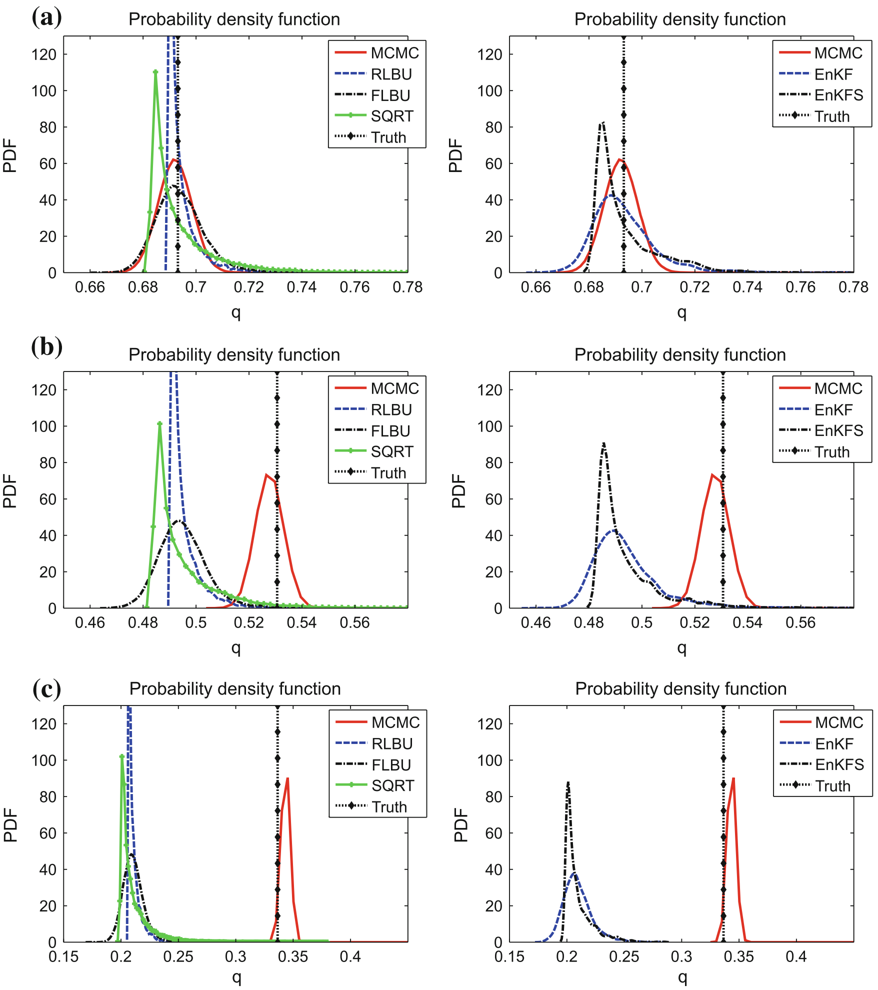
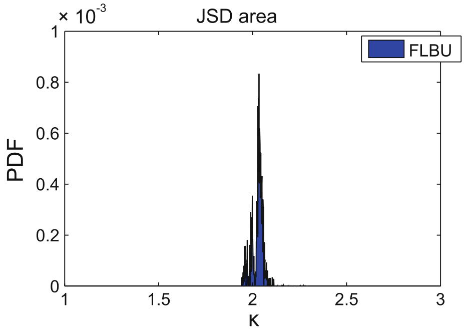


Fig. 10 Comparison of posterior probability density functions describing q. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

outperforms other numerical techniques. Moreover, the linear Bayesian methods "overshoot" when the truth is assumed to be $\kappa_t = 1.7$ or $\kappa_t = 1.4$. This issue can be explained in part by a nonlinearity of the measurement operator. However, there are other possible explanations, such as the prior which is "wrongly" assumed.

To determine the similarity between the MCMC (π_1) and other posterior distributions (π_i), the Jensen-Shannon divergence (JSD):

Fig. 11 The JSD area measuring the distance of the FLBU and MCMC posteriors describing κ



$$\mathrm{JSD}(\pi_i \| \pi_1) = \frac{1}{2} D(\pi_i \parallel \frac{1}{2}(\pi_i + \pi_1)) + \frac{1}{2} D(\pi_1 \parallel \frac{1}{2}(\pi_i + \pi_1)) \tag{37}$$

is estimated. This metric is a smooth and symmetrised version of the Kullback-Leibler divergence $\mathrm{D}(\pi_i \| \pi_1)$ defined as

$$D(\pi_i \| \pi_1) = \int_{-\infty}^{\infty} \ln \left(\frac{\rho_i(x)}{\rho_1(x)} \right) \rho_i(x)\, \mathrm{d}x, \tag{38}$$

where ρ_i and ρ_1 represent the densities of the quantity obtained by one of the Bayesian linear approximation methods and MCMC methods, respectively. In the first case scenario of the truth, the JSD value between the MCMC and FLBU probability distributions is equal to the total area under a curve in Fig. 11. This value corresponds to the error obtained by accumulation of sampling errors and the error caused by a nonlinearity of the measurement operator.

5.1.2 Linear Measurement

To get an adequate understanding of the conclusions drawn in the previous section, one has to consider the experiments in which the measurement operator is linear. Since the relationship between the parameter and the observation is explicitly known

$$\kappa \sim \exp(q) \sim 1/z, \tag{39}$$

one may linearise the measurement operator via the following transformation

$$\log\kappa \sim q \sim \log(1/z). \tag{40}$$

In such a case the estimation parameter q linearly depends on $\log(1/z)$. Following this, the numerical analysis is repeated as in the previous section, only this time

Table 3 Comparison of modes and standard deviations for the posterior κ obtained by different update procedures

Parameter method		Mode			Std		
		C_1	C_2	C_3	C_1	C_2	C_3
κ	Truth	2	1.7	1.4	0	0	0
	MCMC	1.9982	1.7109	1.4878	0.0342	0.0269	0.0209
	RLBU	2.0020	1.7044	1.4062	0.0054	0.0039	0.0034
	FLBU	2.0023	1.7046	1.4064	0.0332	0.0281	0.0232
	SQRT	2.0023	1.7046	1.4064	0.0339	0.0288	0.0241
	EnKF	2.0024	1.7038	1.4046	0.0340	0.0294	0.0237
	EnKFS	2.0022	1.7043	1.4054	0.0339	0.0290	0.0245

Table 4 Comparison of modes and standard deviations for the posterior κ obtained by different update procedures

Parameter method		Mode			Std		
		C_1	C_2	C_3	C_1	C_2	C_3
q	Truth	0.6931	0.5306	0.3365	0	0	0
	MCMC	0.6921	0.5369	0.3972	0.0171	0.0155	0.0209
	RLBU	0.6942	0.5341	0.3407	0.0023	0.0027	0.0022
	FLBU	0.6942	0.5341	0.3407	0.0165	0.0165	0.0165
	SQRT	0.6942	0.5341	0.3407	0.0167	0.0174	0.0168
	EnKF	0.6948	0.5339	0.3410	0.0167	0.0175	0.0170
	EnKFS	0.6940	0.5346	0.3410	0.0166	0.0173	0.0167

with the new version of the measurement. Note that the nonlinear transformation is applied on the measurement data solely without the measurement error.

This study produced results which confirm the findings of a great deal of previous works in this field, see [18, 19, 21, 22]. Namely, as results in Tables 3 and 4 show, the methods based on the linear Bayesian formula are able to identify the truth in all three case scenarios without strong overestimations of variance. However, this is not the case for the one random variable based linear Bayesian update. This method underestimates the posterior variance similarly to the case study already discussed in the previous section. Therefore, the use of the one random variable based linear Bayesian update is not advised in practice. Furthermore, the MCMC procedure shows a slightly different behaviour than in the nonlinear case. The nonlinear transformation of the predicted measurement and observation data in a polynomial chaos form have resulted in a poor posterior estimation in the worst case scenario, for which the truth takes the value in the low probability region of the prior. The problem appears due to large numerical errors caused by both sampling and transformation. The previous findings are in agreement with the probability density plots shown in Figs. 12 and 13, where one may clearly observe the described behaviour of the linear procedures. The

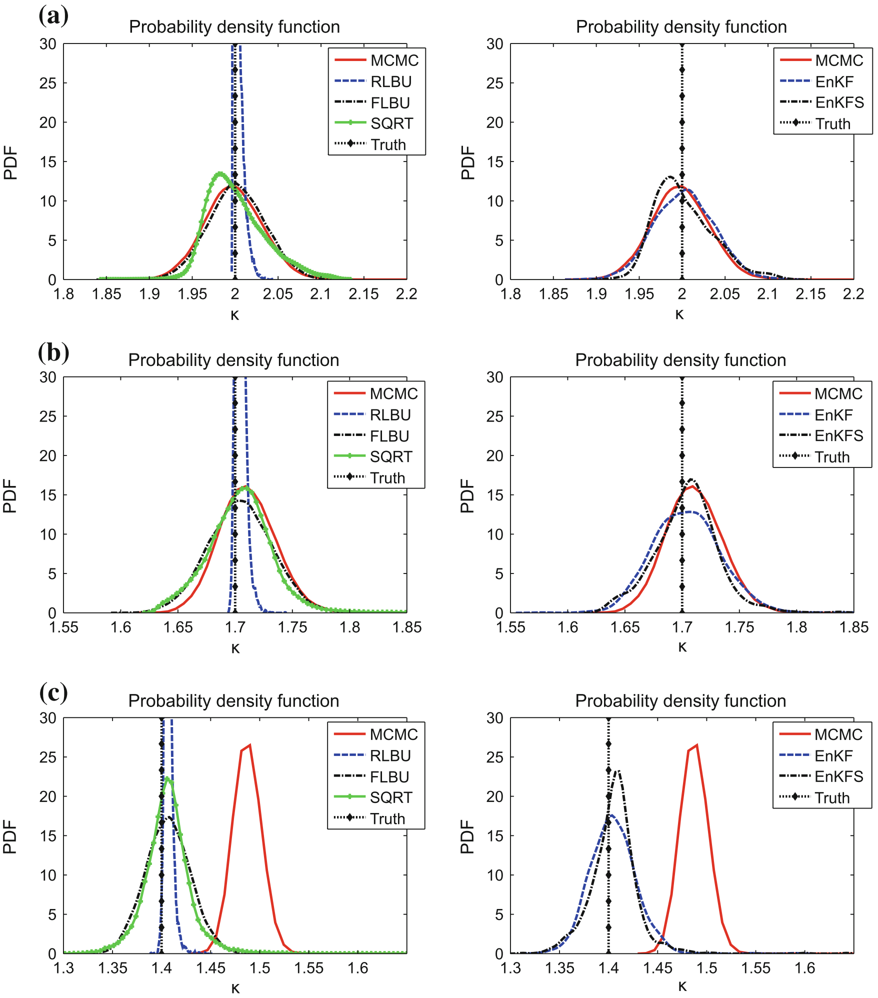


Fig. 12 Comparison of posterior probability density functions describing κ. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

difference in the posterior estimates of the linear Bayesian update and full MCMC procedure for the first case scenario can be seen in Fig. 14, where the JSD area is plotted.

By having run several MCMC computations for the worst case scenario, we came to the conclusion that the posterior distribution converges to the linear Bayes's posterior with the increase of the polynomial order of the proxy model. Increasing the polynomial order from four to six, the MCMC posterior distribution approaches the one obtained by the full linear Bayesian update of fourth order, as shown in Fig. 15.

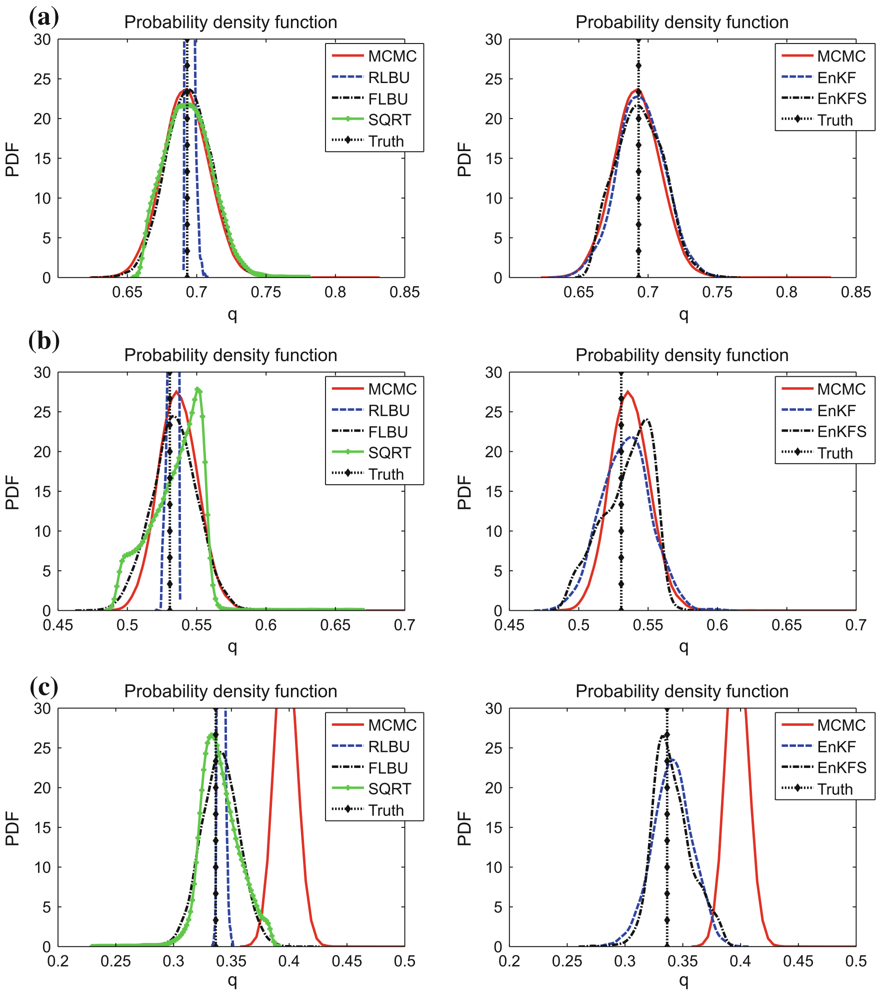
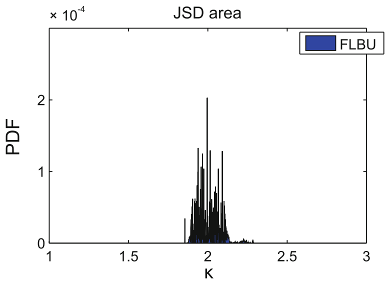


Fig. 13 Comparison of posterior probability density functions describing q. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

Fig. 14 The JSD area measuring the distance of the FLBU and MCMC posterior describing κ



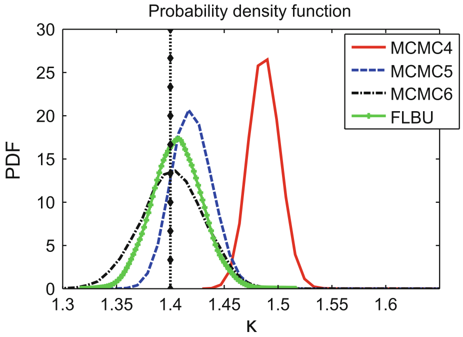


Fig. 15 The improvement of posterior distribution obtained by the MCMC procedure with the increase of the polynomial order from 4 (MCMC4) to 6 (MCMC6). The reference is the full linear Bayesian update of polynomial order 4 (FLBU)

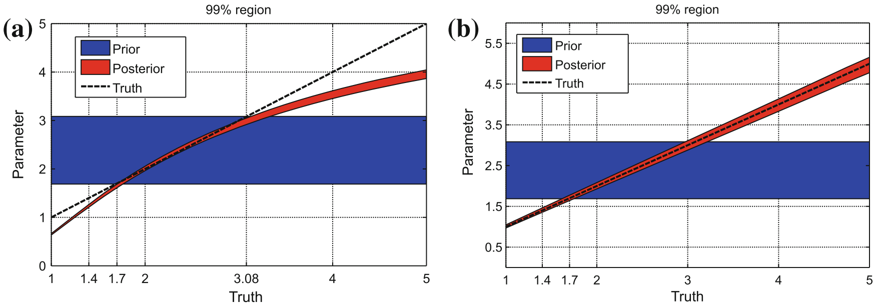


Fig. 16 Comparison of the upper 99 % bounds of posterior and prior of the parameter κ for: **a** nonlinear measurement: **b** linear measurement

This further means that the modelling error can have a huge influence on the MCMC result in the low probability regions of the prior.

Having the previous results in mind, the update procedure is repeated for all possible "truth scenarios"—for the values of κ_t between one and five—and the prior as described in the beginning of this section. As shown in Fig. 16, the truth is inside the 99 % region of the posterior (red area) in case of linear measurement. However, the same line crosses the 99 % region of the posterior (red area) in case of nonlinear measurement. For the former scenario, the posterior contains the truth only in a small region around the 2.3 value (prior mean), where the truth line appears to be the tangent. On the other hand, in the linear case the posterior better estimates the truth although the variance can be over- or underestimated. This leads to the conclusion that only the measurement operators with slight nonlinearities can be handled with the linear Bayes procedure.

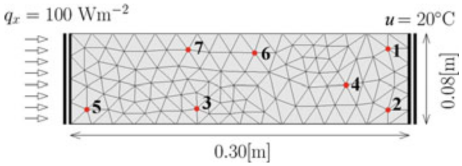


Fig. 17 Experimental setup

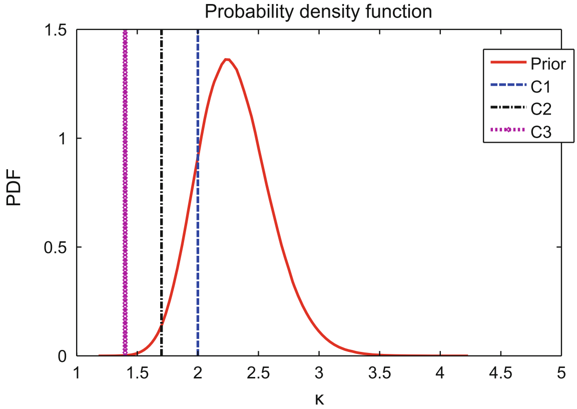


Fig. 18 The prior distribution and three case scenarios for the truth C_1–C_3

5.2 *Two Dimensional Heat Problem*

In order to improve the previous qualitative analysis, the steady diffusion problem is also examined on a two-dimensional rectangular domain, see Fig. 17. The boundary conditions consist of a heat flux $q = 100\,\mathrm{W\,m^{-2}}$ prescribed on the left boundary and a constant temperature of 20 °C imposed on the right boundary. The computational domain is discretised with the help of 204 irregular finite elements and $N = 124$ nodes.

For the virtual truth, the thermal conductivity κ_t is taken to be one realisation of a lognormal random variable described independently from the a priori distribution. The temperature is evaluated with the help of a deterministic finite element (FE) method, but only the values in 7 randomly chosen points (FE nodes highlighted by red dots in Fig. 17) are taken into consideration. For reasons of simplicity, these points (sensors) are not optimally placed, even though this can be achieved with the help of the optimisation theory. The measured data are additionally disturbed by Gaussian noise with zero mean and covariance $C_\varepsilon = \sigma_\varepsilon^2 I$ in order to simulate realistic data.

For this study, the prior thermal conductivity κ_f is designed with the help of the maximum entropy approach which takes all available information about the conductivity parameter in the process of model selection. Thanks to the positive definiteness of κ, the prior is taken to be a lognormal random variable with the mean $\bar{\kappa}_f = 2.3$ and standard deviation $\sigma_f = 0.3$. In a similar manner as before, the value of κ_t is adopted such that the truth places in the one, two or three sigma region of the prior, see Fig. 18. The reasons for this are the same as described before in Sect. 5.1.

5.3 Forward Problem

For prediction purposes the stochastic diffusion problem described by uncertain conductivity coefficient is solved with the help of the stochastic Galerkin method [22]. These results are verified with the help of a pure Monte Carlo approach with one million samples, see Fig. 19 for its convergence in mean and variance. For further investigation only one surrogate model is selected through a validation process: that is the polynomial chaos expansion of order 4. Compared to the MC reference solution this approximation results in 0.3029e-4 for the relative error in the mean, and 0.0011 for the relative error in variance. The mean value and variance of such an approximated solution are shown in Fig. 20. As expected, the mean is a linear and the variance is a nonlinear function of the coordinates.

5.4 Identification

As in the previous example, the experimental analysis is run by measuring the temperature, i.e. the nonlinear function of the conductivity parameter. For comparison purposes, several computational strategies as described in Sect. 5.1 are implemented and tested: the one random variable based linear Bayesian update (RLBU), full linear Bayesian update (FLBU), the square root update (SQRT), ensemble Kalman filter (EnKF) with 1000 samples, square root ensemble Kalman filter (EnKFS) with 1000 samples and the full Bayesian MCMC update with 10^5 samples. The last procedure is declared as the reference solution. Its convergence with respect to the number of samples can be seen in Fig. 21, where the relative errors of the mean and variance

$$\varepsilon_m = \frac{\|\bar{\kappa}_a^N - \bar{\kappa}_a^R\|}{\|\bar{\kappa}_a^R\|}, \quad \varepsilon_v = \frac{\|\mathrm{var}\,\kappa_a^N - \mathrm{var}\,\kappa_a^R\|}{\|\mathrm{var}\,\kappa_a^R\|}, \tag{41}$$

are plotted, respectively. Here, $\bar{\kappa}_a^N$, var κ_a^N stand for the mean and variance of the posterior distribution obtained with N samples, whereas $\bar{\kappa}_a^R$, var κ_a^R denote the mean and variance of the posterior distribution as a result of 10^5 runs. According to these plots the relative errors are slowly converging with the number of samples, as expected. This results in an accuracy of ca. 1e-12 for the mean conductivity and 1e-8 for the conductivity variance.

A comparison of identification results in Fig. 22 reveals that the MCMC procedure is the only one able to identify the truth in all three assumed cases. The methods based on the linear approximation behave well in cases when the truth lies in the high probability region, otherwise an “overshooting” occurs. This lends weight to the previously given argument that the error of the linear approximant strongly depends on the nonlinearity of the measurement operator, as well as on the prior assumption.

The previous results are also supported by the plots of posterior 99 % confidence intervals in Fig. 23. Initially, before the measurements are carried out, the 99 % con-

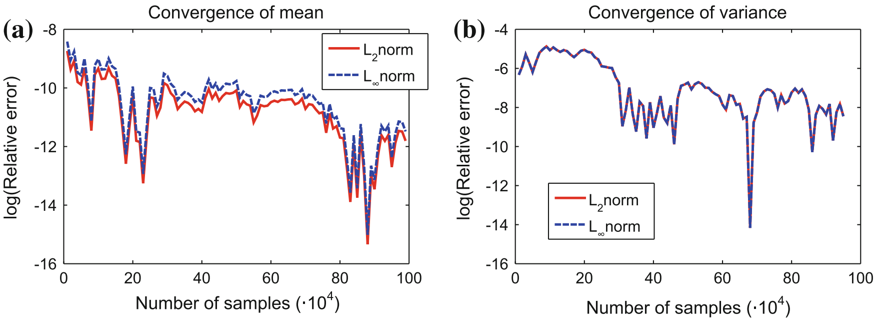
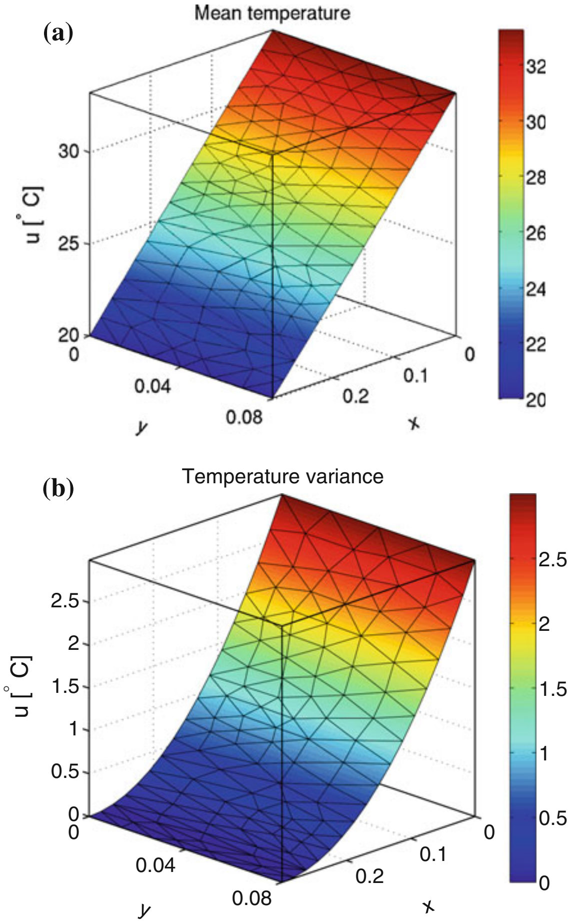


Fig. 19 Convergence of the **a** mean and **b** variance of the Monte Carlo method with 10^6 samples

Fig. 20 The second order statistics obtained with the help of the stochastic Galerkin method. **a** The mean value of temperature, **b** the temperature variance



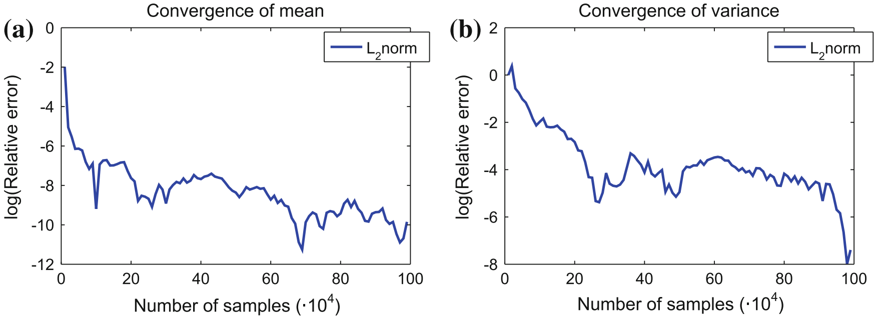


Fig. 21 The Markov chain Monte Carlo convergence of posterior κ_a for the Gaussian measurement error described by a standard deviation 0.3. **a** Mean convergence, **b** convergence in variance

fidence interval is assumed to be broad in order to "catch" the truth. With every new successive measurement the probability region narrows down such that the interval becomes almost deterministic after seven performed measurements. Even though the truth is assumed to be deterministic, the posterior 99 % confidence interval does not disappear due to the measurement and model errors, as well as the error due to the nonlinearity of the measurement operator. Additionally, the results, as seen in Fig. 23, indicate that the full linear Bayesian update (FLBU), as well as the EnKF almost match the MCMC results in the first scenario, whereas the one random variable linear Bayesian update (RLBU) underestimates the posterior variance. On the other hand, the square root filter and EnKF square root filter deliver similar results with slightly shifted median.

The issues previously described can be resolved at the expense of improving the prior description. This can be done by moving the mean of the prior distribution towards the truth. Similar to the Kalman way of updating, we may obtain more information about the prior mean using the existing measurement data. Once this has been done, one may alternate the old prior with the newly obtained mean value and continue the estimation as described previously. Even though the method just described represents an oversimplification (especially in case of nonlinear functions), it does work for slightly nonlinear functions (measurement operators). Such a case is depicted in Fig. 24, where the process of identifying the truth in 2σ region is illustrated. Here, the red line denotes the prior distribution, while the dashed black line is the newly adopted prior (same statistics besides mean). From the resulting plots of the posterior one may see that this time all procedures return similar results. This point emphasizes the importance of the prior assumption. Also, one may note in Fig. 25 that linear Bayesian updates are better than the MCMC update when the number of measurement points is small.

As mentioned earlier, the estimation of the conductivity coefficient greatly depends on the number of the measurement points, as well as on the measurement (model) errors. The analysis and simulation of the updating procedures for different levels of the measurement errors are shown in Fig. 26. From these figures, it is appar-

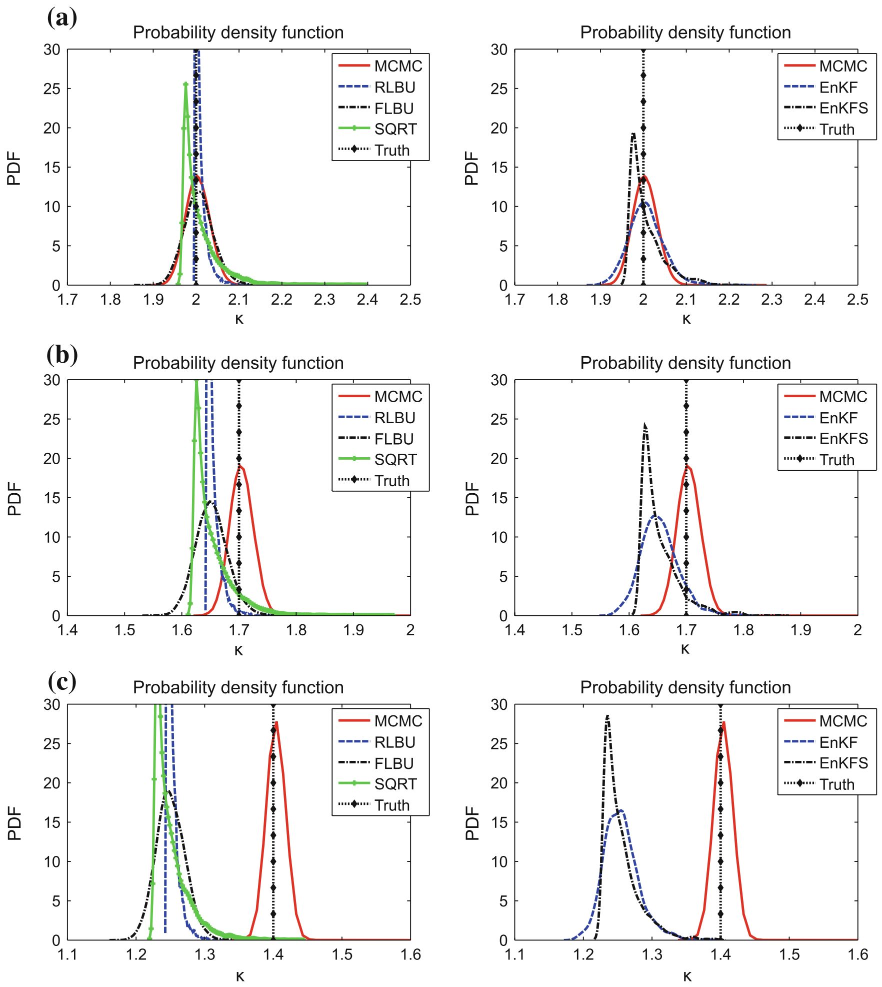


Fig. 22 Comparison of posterior probability density functions describing κ. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

ent that the posterior estimate better complies the truth for smaller values of error. The same is valid for second two case scenarios, see Fig. 27. However, these plots also reveal that the bigger measurement error can regularise the estimation process such that the 99 % confidence interval of posterior includes the truth value.

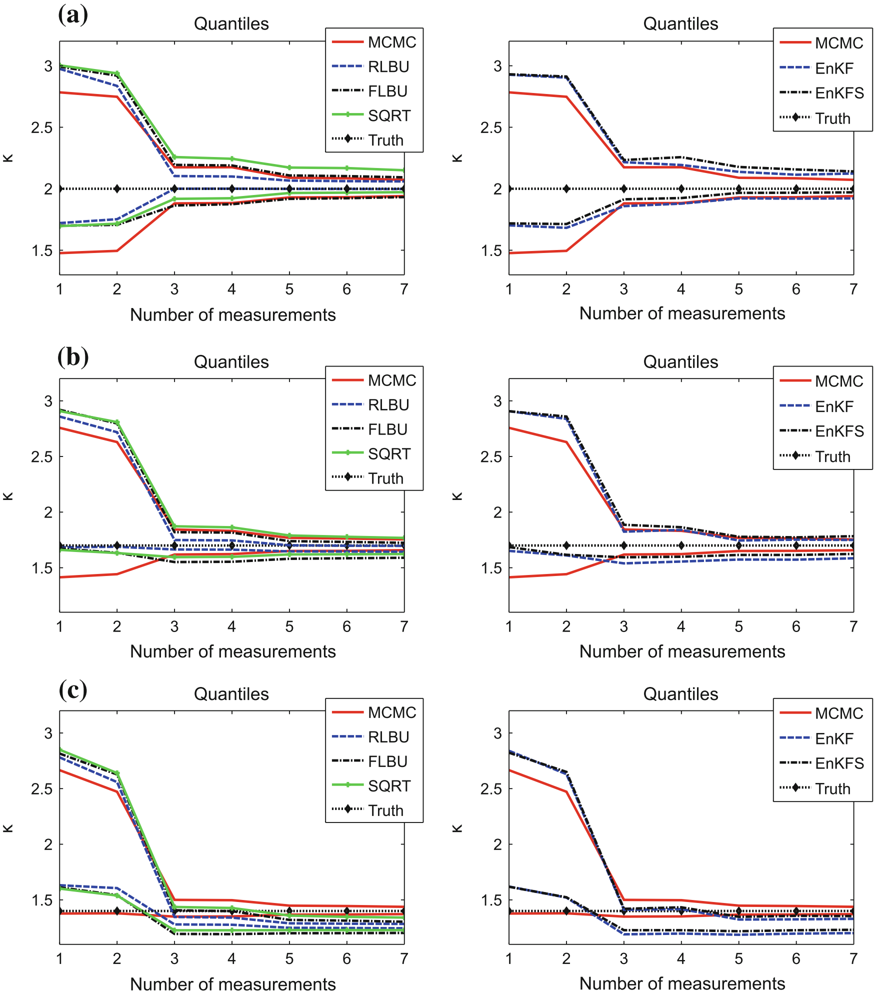


Fig. 23 Comparison of posterior probability density functions describing κ for different number of measurement points. **a** The truth lies in the one sigma region of the prior, **b** the truth lies in the two sigma region of the prior, **c** the truth lies in the three sigma region of the prior

6 Conclusions

This contribution aimed to present and compare different numerical approaches to Bayesian estimations of non-observable model parameters from noisy measurement data. The model parameter stands for the thermal conductivity and is represented by a random variable with a non-Gaussian prior distribution. The numerical findings suggest that Markov chain Monte Carlo sampling of the posterior distribution is a reliable way of computing the Bayesian update. However, the model simulation—

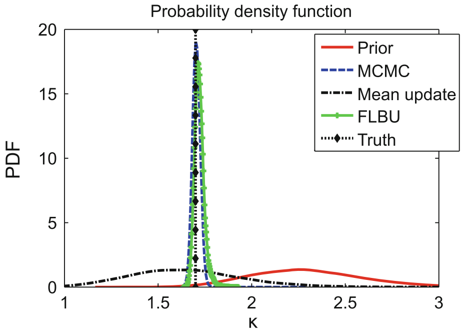


Fig. 24 The process of updating by first moving the prior mean

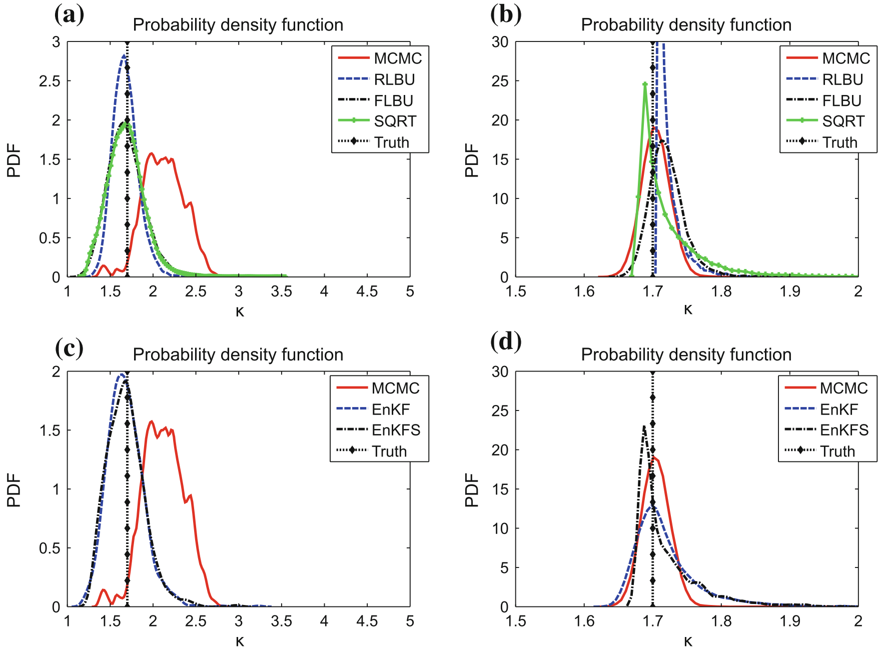


Fig. 25 Comparison of probability density functions of the posterior of κ for different numbers of measurement points after moving the prior. **a** Two measurements, **b** seven measurements, **c** two measurements, **d** seven measurements

often time-consuming—has to be evaluated for each sample in the chain, which makes the whole procedure computationally expensive. To overcome this problem, the stochastic Galerkin method is employed in order to construct a polynomial chaos based approximation of the model response. This is then used within the sampling procedure, instead of full model simulations. From these numerical findings it is evident that the proxy model may jeopardize the accuracy of the Markov chain Monte

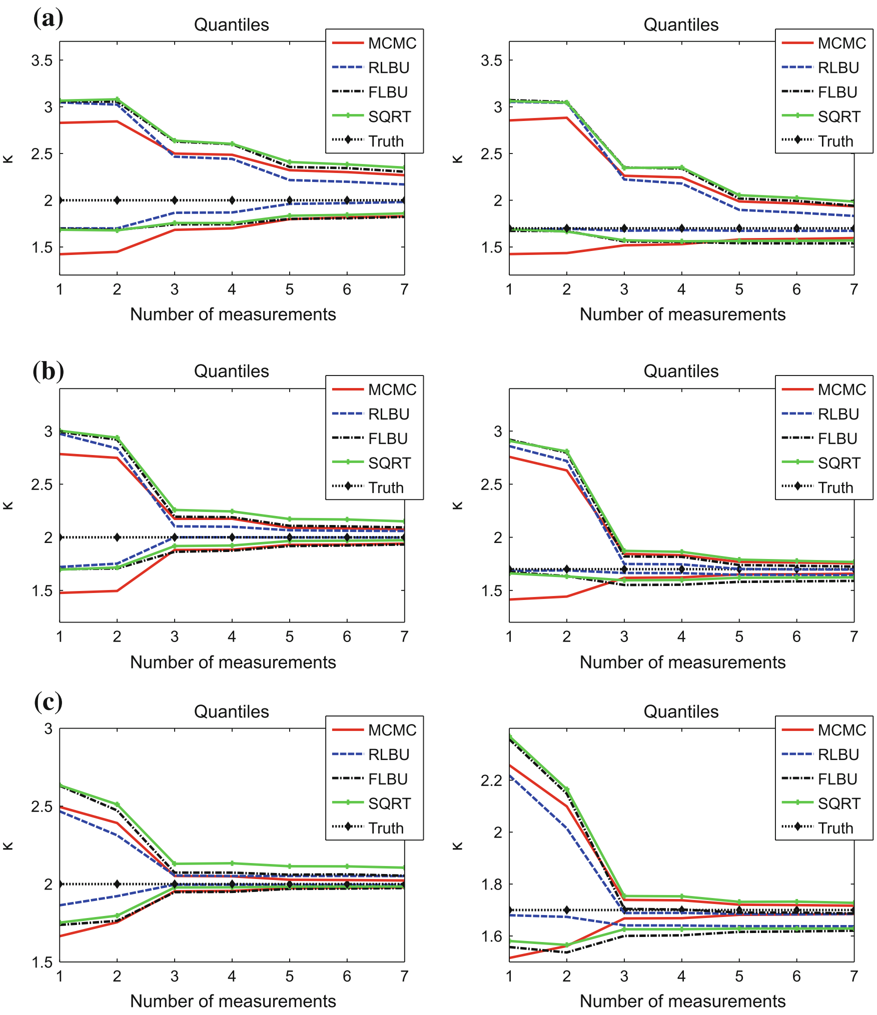
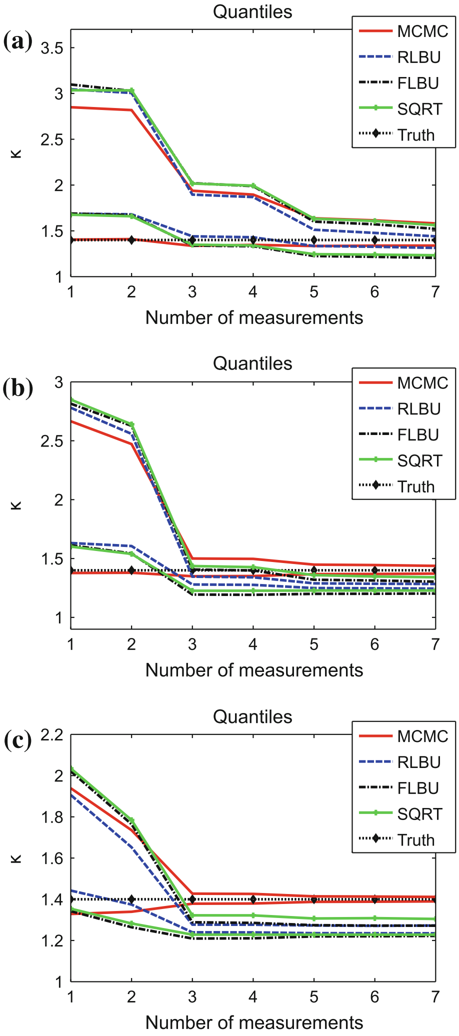


Fig. 26 The influence of the measurement error on the posterior quantiles: *left* is the truth in C_1 scenario and *right* in C_2 scenario. **a** $\sigma_\epsilon = 1$, **b** $\sigma_\epsilon = 0.3$, **c** $\sigma_\epsilon = 0.1$

Carlo procedure due to a combination of both approximation and sampling errors. For reasons of efficiency the updating procedure is recasted into an alternative linear Bayesian form thereby enabling a direct algebraic way of computing the posterior distribution. While the sampling version of the linear filter—also known as ensemble Kalman filter—needs a considerably smaller number of samples than the MCMC procedure, this approach is severely underestimating or overestimating the residual uncertainty. On the other hand, the polynomial chaos linear Bayesian methods do not seem to suffer from any of previously mentioned issues. They deliver more reliable results than EnKF procedures and are better than proxy MCMC in case of linear

Fig. 27 The influence of the measurement error on the posterior quantiles for case scenario C_3. **a** $\sigma_\epsilon = 1$, **b** $\sigma_\epsilon = 0.3$, **c** $\sigma_\epsilon = 0.1$



measurements. However, LBU may suffer from larger residual errors when applied in nonlinear cases.

While the initial findings are promising, further research is necessary. Therefore, future analysis will be needed to validate the mentioned numerical behaviour of the presented computational procedures for the more complex diffusion problem when the conductivity parameter is modelled in a form of random field. In addition, the adaption of the polynomial chaos based linear Bayes filter has to be made in order to handle the assimilation from the noisy nonlinear measurements.

Acknowledgments This outcome has been achieved with the financial support of the German Research Foundation (DFG), Czech Science Foundation, projects No. 105/12/1146 and No. 15-07299S, and the German Academic Exchange Service, project DAAD BeWuF ID 50755591.

References

1. A. Andrews. A square root formulation of the Kalman covariance equations. *AIAA Journal*, 22(6):1165–1166, June 1968.
2. E. D. Blanchard. *Polynomial chaos approaches to parameter estimation and control design for mechanical systems with uncertain parameters*. PhD thesis, Department of Mechanical Engineering, VirginiaTech University, 2010.
3. A. Bobrowski. *Functional analysis for probability and stochastic processes: an introduction*. Cambridge University Press, 2005.
4. S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.
5. J. A. Christen and C. Fox. MCMC using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795 – 810, 2005.
6. H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Dordrecht: Kluwer, 2000.
7. G. Evensen. *Data Assimilation, The Ensemble Kalman Filter*. Springer, Berlin, June 2007.
8. G. Evensen. The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, 29:82–104, 2009.
9. M. Friswell and J. E. Mottershead. *Finite element model updating in structural dynamics*. Kluwer Academic Publishers, Dordrecht, Neitherlands, 1995.
10. D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian Inference*. Chapman and Hall/CRC, 2006.
11. R. Ghanem and P. D. Spanos. *Stochastic finite elements—A spectral approach*. Springer-Verlag, New York, 1991.
12. R. E. Kálmán. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering (Ser D)*, 82:35–45, 1960.
13. A. Kučerová. *Identification of nonlinear mechanical model parameters based on softcomputing methods*. PhD thesis, Ecole Normale Supérieure de Cachan, Laboratoire de Mécanique et Technologie, 2007.
14. A. Kučerová, J. Sýkora, B. Rosić, and H. G. Matthies. Acceleration of uncertainty updating in the description of transport processes in heterogeneous materials. *Journal of Computational and Applied Mathematics*, 236(18):4862–4872, 2012.
15. Y. M. Marzouk, H. N. Najm, and L. A. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, June 2007.
16. Y. M. Marzouk and D. Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.
17. H. G. Matthies. Stochastic finite elements: computational approaches to stochastic partial differential equations. *Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)*, 88(11):849–873, 2008.
18. O. Pajonk, B. Rosić, A. Litvinenko, and H. G. Matthies. A deterministic filter for non-Gaussian Bayesian estimation. *Physica D: Nonlinear Phenomena*, 241(7):775–788, 2012.
19. O. Pajonk, B. Rosić, and H. G. Matthies. Sampling-free linear Bayesian updating of model state and parameters using a square root approach. *Computers and Geosciences*, 55:70–83, 2012.
20. B. L. Pence, H. K. Fathy, and J. L. Stein. A maximum likelihood approach to recursive polynomial chaos parameter estimation. In *Proceedings of American Control Conference (ACC)*, pages 2144–2151, 2010.

21. B. Rosić, A. Kučerová, J. Sýkora, O. Pajonk, A. Litvinenko, and H. G. Matthies. Parameter identification in a probabilistic setting. *Engineering Structures*, 50:179–196, 2013.
22. B. Rosić, O. Pajonk, A. Litvinenko, and H. G. Matthies. Sampling-free linear Bayesian update of polynomial chaos represenations. *Journal of Computational Physics*, 231(17):5761–5787, 2012.
23. G. Saad and R. Ghanem. Characterization of reservoir simulation models using a polynomial chaos-based ensemble Kalman filter. *Water Resources Research*, 45(W04417):–, 2009.
24. A. Schöniger, W. Nowak, and H.-J. Hendricks Franssen. Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. *Water Resources Research*, 48(4):n/a–n/a, 2012.
25. E. Simon and L. Bertino. Application of the Gaussian anamorphosis to assimilation in a 3-d coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. *Ocean Science*, 5(4):495–510, 2009.
26. A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3–23, 1993.
27. C. Soize. Construction of probability distributions in high dimensions using the maximum entropy principle. Applications to stochastic processes, random fields and random matrices. *International Journal for Numerical Methods in Engineering*, 76(10):1583–1611, 2008.
28. L. Stone, R. L. Streit, and T. L. Corwin. *Bayesian multiple target tracking*. Artech House Inc, 2014.
29. H. A. Tchelepi, H. Bazargan, and M. A. Christie. Efficient Markov chain Monte Carlo sampling using polynomial chaos expansion. In *Proceedings of the SPE Reservoir Simulation Symposium*, The Woodlands, Texas, United States, 2013. online.
30. L. Tiernay. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.
31. A. N. Tikhonov. Regularisation of incorrectly posed problems. *Soviet Mathematics Doklady*, 4(4):1624–1627, 1963.
32. R. L. Tweedie. Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes and their Applications*, 3(4):385 – 403, 1975.
33. N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60:897–936, 1938.
34. A. Wirgin. The inverse crime. Technical report, arXiv:math-ph/0401050, 2004.
35. D. Xiu and G. E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24:619–644, 2002.

Two Models for Hydraulic Cylinders in Flexible Multibody Simulations

Antti Ylinen, Jari Mäkinen and Reijo Kouhia

Abstract In modelling hydraulic cylinders interaction between the structural response and the hydraulic system needs to be taken into account. In this chapter two approaches for modelling flexible multibody systems coupled with hydraulic actuators i.e. cylinders are presented and compared. These models are the truss-element-like cylinder and bending flexible cylinder models. The bending flexible cylinder element is a super-element combining the geometrically exact Reissner-beam element, the C^1-continuous slide-spring element needed for the telescopc movement and the hydraulic fluid field. Both models are embeded with a friction model based on a bristle approach. The models are implemented in a finite element enviroment. In time the coupled stiff differential equation system is integrated using the L-stable Rosenbrock method.

1 Introduction

The range of applications for hydraulic driven working machines is vast compassing very robust excavators and precise robots such as used in the maintenance of the ITER fusion reactor. In hydraulic systems, the energy is transferred via pressurized fluid instead of using mechanical components such as gears and levers, which allows for a more flexible layout for the components since the fluid flows through pipes or

A. Ylinen (✉)
FS Dynamics Finland Oy Ab, Hermiankatu 1, FI-33720 Tampere, Finland
e-mail: antti.ylinen@fsdynamics.fi

J. Mäkinen
Department of Civil Engineering, Tampere University of Technology,
P.O. Box 600, FI-33101 Tampere, Finland
e-mail: jari.m.makinen@tut.fi

R. Kouhia
Department of Mechanical Engineering and Industrial Systems,
Tampere University of Technology, P.O. Box 589, FI-33101 Tampere, Finland
e-mail: reijo.kouhia@tut.fi

A. Ibrahimbegovic (ed.), *Computational Methods for Solids and Fluids*,
Computational Methods in Applied Sciences 41,
DOI 10.1007/978-3-319-27996-1_17

flexible hoses. The hydraulic energy is then converted to mechanical energy using actuators.

The hydraulic pump producing the flow rate to the system, is typically run by an electric motor or a diesel engine. The flow rate is then impelled through the hydraulic control system to an actuator, where the hydraulic energy is transferred back into mechanical energy. Thus, three major subsystems in hydraulically driven systems can be identified: (i) the hydraulic control system, (ii) the hydraulic actuator, and (iii) the mechanical system. These systems are coupled, since the state of the mechanical system is dependent on the state of the hydraulic actuator and hydraulic control system, and vice versa.

In mobile working machines, such as excavators or personnel lifting gear, the boom movements are typically driven by hydraulic cylinders. Hydraulic cylinders are linear actuators, where the movement is in the direction of the cylinder itself, and they can extend or contract in length. The other type of hydraulic actuator is a hydraulic motor, which produces a rotary movement. The movement in both of these actuators is achieved by pumping pressurized fluid into the actuator, which creates a reactive force that accelerates the actuator and, finally, the mechanical system.

In this chapter some modeling features of the linear actuator, i.e. the hydraulic cylinder, are addressed. In addition, the hydraulic cylinder elements are constructed to be used in multibody similations compatible with a framework of finite elements.

Traditionally, linear movement can be expressed using constraint equations. However, such equations only provide for a rigid connection between two points, whereas hydraulic cylinder elements also include the flexibility of the hydraulic oil. Moreover, the hydraulic cylinder elements can be used to study the chamber pressures and the flow rates in and out of the cylinder chambers and the hydraulic cylinder element provides an interface for incorporating the hydraulic control system into the simulations.

In this paper the effect of bending flexibility of the hydraulic cylinder is studied. Hydraulic cylinders have to be designed to resist buckling under maximum operating pressures. The maximum operating pressure thus defines the maximum axial load which the cylinder has to carry without buckling. In normal operating conditions, the cylinder should only carry a load through axial compression, whereas in accident conditions the cylinder can also be exposed to bending. In addition, if the cylinder is not mounted in a vertical position, the gravitational forces always place bending stresses on the cylinder. This effect has been studied in [19].

Since hydraulic cylinders are operated with hydraulic fluid, the cylinder has to be sealed. The seal then introduces friction forces opposing the cylinder movement. Therefore, a novel friction model into the cylinder element formulations is included, and the effects of the apparent friction force on the response of the multibody system are studied Ascertaining the correct stresses in the system can have an impact on any fatigue assessment of the system, for instance. One objective of this study is to define the features to be modeled when dealing with multibody simulations. Since dynamic simulations are involved, it is useful to have as few degrees of freedom in the system as possible, which means that it is not particularly useful to account for all the properties in the simulation models.

2 Equations of Motion

The equations of motion can be derived by utilizing the principle of virtual work, which can be written as, see [7]

$$\delta W = \int_{\mathscr{B}} \delta \mathbf{u} \cdot \mathbf{b} \, \mathrm{d}v + \int_{\partial \mathscr{B}} \delta \mathbf{u} \cdot \mathbf{t}_\sigma \, \mathrm{d}a_\sigma - \int_{\mathscr{B}} \delta \varepsilon : \sigma \, \mathrm{d}v - \int_{\mathscr{B}} \delta \mathbf{u} \cdot \rho \ddot{\mathbf{u}} \, \mathrm{d}v = 0, \tag{1}$$

where the first term is related to the body forces, with the body force and displacement vectors denoted as $\mathbf{b}$ and $\mathbf{u}$, respectively. The second term relates to the tractions $\mathbf{t}_\sigma$ acting on the surfaces of the body $\mathscr{B}$. The third term is due to the internal forces with Cauchy stress σ and the work conjugate strain measure ε, called the Almansi strain, and finally the last term is due to the virtual work of the acceleration forces written with the aid of D'Alembert's principle. The density of the body $\mathscr{B}$ is given as ρ.

In the finite element method the displacement field is interpolated as $\mathbf{u} = \mathbf{N}\mathbf{q}$, where the matrix $\mathbf{N}$ contains the interpolation functions and the $\mathbf{q}$ the unknown parameters, the generalized coordinates. In the Galerkin approach the same interpolation is also used for the virtual displacements $\delta\mathbf{u} = \mathbf{N}\delta\mathbf{q}$, the virtual work equation can be written as

$$\delta W = \delta \mathbf{q} \cdot (\mathbf{f}_{\text{ext}} - \mathbf{f}_{\text{int}} - \mathbf{f}_{\text{acc}}) = 0. \tag{2}$$

Since the equation above has to hold for all virtual displacements $\delta\mathbf{q}$, and by noting that the last term in the equation, $\mathbf{f}_{\text{acc}}$, corresponds to $\mathbf{f}_{\text{acc}} = \mathbf{M}\ddot{\mathbf{q}}$, it can be written

$$\mathbf{M}\ddot{\mathbf{q}} = \mathbf{f}_{\text{ext}} - \mathbf{f}_{\text{int}} \tag{3}$$

where $\mathbf{M}$ is the mass matrix of the system. The mass matrix is the product of the integration of the last term in (1) when the discretization is performed using the interpolation functions. The external forces, $\mathbf{f}_{\text{ext}}$, and the internal forces, $\mathbf{f}_{\text{int}}$, can depend on the current state and velocities of the system. In the following section the equations of motion for the coupled system are given.

3 Coupled Multibody System

In the previous section the equation of motion for the mechanical system was presented. Here, the coupled multibody system is introduced, consisting of three major components: the mechanical system, the hydraulic cylinder and the control system, see Fig. 1 for an illustration of a lifting boom. The generalized coordinates of the mechanical system is denoted with $\mathbf{q}$, the hydraulic cylinder state variables with $\mathbf{z}$

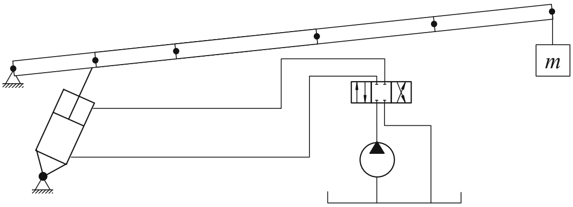


Fig. 1 General coupled multibody system with three major parts: Mechanical system (the boom), hydraulic cylinder and hydraulic control system (4/3-directional valve)

and the control system variables with **c**. The complete equations of motion for the three systems, where the constraint equations are excluded, can be written as

$$\begin{cases} \mathbf{M}\ddot{\mathbf{q}} = \mathbf{g}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{z}, t) \\ \dot{\mathbf{z}} = \mathbf{f}_{\mathrm{cyl}}(\mathbf{z}, \mathbf{q}, \dot{\mathbf{q}}, \mathbf{c}, t) \\ \dot{\mathbf{c}} = \mathbf{f}_{\mathrm{hyd}}(\mathbf{c}, \mathbf{z}, t), \end{cases} \tag{4}$$

where the first equation of motion is for the mechanical system, the second for the hydraulic cylinder and the third for the hydraulic control system. For the mechanical system, **M** is the mass matrix and function **g** is the sum of the external, internal and complementary inertial forces, see [5] and (3). The complementary inertial forces are related to the beam element formulations and include gyroscopic forces, see [5, 10].

The evolution laws for the hydraulic cylinder and for the hydraulic control system are denoted as $\mathbf{f}_{\mathrm{cyl}}$ and $\mathbf{f}_{\mathrm{hyd}}$ respectively. Descriptions for these evolution laws are given in detail in Sect. 4.1.1. Coupling between the control system and the mechanical system is through the hydraulic cylinder, which is visible on the right hand sides of (4) where the state of the hydraulic cylinder variables depends on the state of the control system variables **c** and the mechanical system variables **q** and $\dot{\mathbf{q}}$.

For the solution of the coupled system the linearization of the equations of motion (4) is needed. Using the following subscripts m, c and h for the mechanical system, hydraulic cylinder and control system respectively, the linearization can be written as

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta\ddot{\mathbf{q}} \\ \Delta\ddot{\mathbf{z}} \\ \Delta\ddot{\mathbf{c}} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{\mathrm{mm}} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_{\mathrm{cm}} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \Delta\dot{\mathbf{q}} \\ \Delta\dot{\mathbf{z}} \\ \Delta\dot{\mathbf{c}} \end{bmatrix}$$
$$+ \begin{bmatrix} \mathbf{K}_{\mathrm{mm}} & \mathbf{J}_{\mathrm{mc}} & \mathbf{0} \\ \mathbf{J}_{\mathrm{cm}} & \mathbf{J}_{\mathrm{cc}} & \mathbf{J}_{\mathrm{ch}} \\ \mathbf{0} & \mathbf{J}_{\mathrm{hc}} & \mathbf{J}_{\mathrm{hh}} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{q} \\ \Delta\mathbf{z} \\ \Delta\mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{r}^* \\ \mathbf{s}^* \\ \mathbf{t}^* \end{bmatrix}, \tag{5}$$

where the right hand sides are the residual vectors defined as

$$\mathbf{r}^* = \mathbf{g}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{z}, t) - \mathbf{M}\ddot{\mathbf{q}} \tag{6}$$
$$\mathbf{s}^* = -\dot{\mathbf{z}} + \mathbf{f}_{\mathrm{cyl}}(\mathbf{z}, \mathbf{q}, \dot{\mathbf{q}}, t) \tag{7}$$
$$\mathbf{t}^* = -\dot{\mathbf{c}} + \mathbf{f}_{\mathrm{hyd}}(\mathbf{c}, \mathbf{z}, t) \tag{8}$$

for the mechanical system, hydraulic cylinder and the hydraulic control system.

The mass matrix for the system in (5), $\mathbf{M}$, is due to the last term in (1) when the discretization is performed. For the mechanical system the damping matrix is $\mathbf{C}_{\mathrm{mm}}$ and the stiffness matrix $\mathbf{K}_{\mathrm{mm}}$. The damping matrix and the stiffness matrix appear in the differentiation of the vector $\mathbf{r}^*$ in (6) with respect to $\mathbf{q}$ and $\dot{\mathbf{q}}$. The stiffness matrix is then

$$\mathbf{K}_{\mathrm{mm}} = -\frac{\partial \mathbf{g}}{\partial \mathbf{q}} + \frac{\partial (\mathbf{M}\ddot{\mathbf{q}})}{\partial \mathbf{q}}, \tag{9}$$

where the latter term arises from the derivation of the inertial forces. This gyroscopic term appears when the mass matrix is configuration (or displacement) dependent. The damping matrix is then defined as

$$\mathbf{C}_{\mathrm{mm}} = -\frac{\partial \mathbf{g}}{\partial \dot{\mathbf{q}}}. \tag{10}$$

The Jacobian matrices for the hydraulic cylinder and the control system are written by differentiating the residual vectors given in (7)–(8) as follows

$$\mathbf{J}_{\mathrm{cc}} = -\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \mathbf{z}}, \quad \mathbf{J}_{\mathrm{hh}} = -\frac{\partial \mathbf{f}_{\mathrm{hyd}}}{\partial \mathbf{c}}. \tag{11}$$

The off-diagonal terms in the linearized equation of motion (5) are due to the coupling of the three systems. Explicit forms of the coupling matrices are given in Sect. 4.

As stated earlier, the coupling between the mechanical system and the hydraulic control system is through the hydraulic cylinder. Two coupling pairs can be identified: the mechanical system and hydraulic cylinder, and the hydraulic cylinder and control system. The first of the couplings is handled with three matrices defined as

$$\mathbf{C}_{\mathrm{cm}} = -\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \dot{\mathbf{q}}}, \quad \mathbf{J}_{\mathrm{cm}} = -\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \mathbf{q}}, \quad \mathbf{J}_{\mathrm{mc}} = -\frac{\partial \mathbf{g}}{\partial \mathbf{z}}. \tag{12}$$

When the hydraulic control system is excluded from the simulation model only these coupling terms are active. In addition the control system Jacobian in $(11)_2$ is also excluded because in the linearized equations of motion in (5) only the first two equations are utilized.

If the hydraulic control system is included, two more coupling matrices are included into the system through the last equation in (5). These coupling terms are then written as

$$\mathbf{J}_{\mathrm{ch}} = -\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \mathbf{c}}, \quad \mathbf{J}_{\mathrm{hc}} = -\frac{\partial \mathbf{f}_{\mathrm{hyd}}}{\partial \mathbf{z}}. \tag{13}$$

The main interest for the hydraulic cylinder modeling is in defining the evolution law $\mathbf{f}_{\mathrm{cyl}}(\mathbf{z}, \mathbf{q}, \dot{\mathbf{q}}, t)$ and finding the appropriate coupling matrices with the mechanical system.

4 Finite Element Formulations

In this section the finite elements used in modeling the hydraulic cylinder are described. The basic element is the geometrically exact beam with total Lagrangian formulation described in [10]. In addition to the conventional beam modelling, a special spring element is developed in [13]. These two elements are then incorporated to create the C^1-continuous slide-spring element to model sliding joints.

This section describes derivation of two different type of hydraulic cylinder elements, namely the truss element cylinder (TC) and the bending flexible cylinder (BF). The truss element cylinder is, as the name suggests, rigid in bend, whereas the bending flexible cylinder can capture the bending flexibility of the hydraulic cylinder.

In deriving the hydraulic cylinder elements, two different descriptions for each cylinder model, one for a dynamic simulation and the other for a quasi-static analysis is given in [19]. In the literature regarding hydraulic cylinders, typically, only dynamic simulations are considered, see, for instance, [2, 3]. Application of the quasi-static analysis allows the initial state to be computed without a dynamic analysis, thus reducing the computational requirements. For further details see [19].

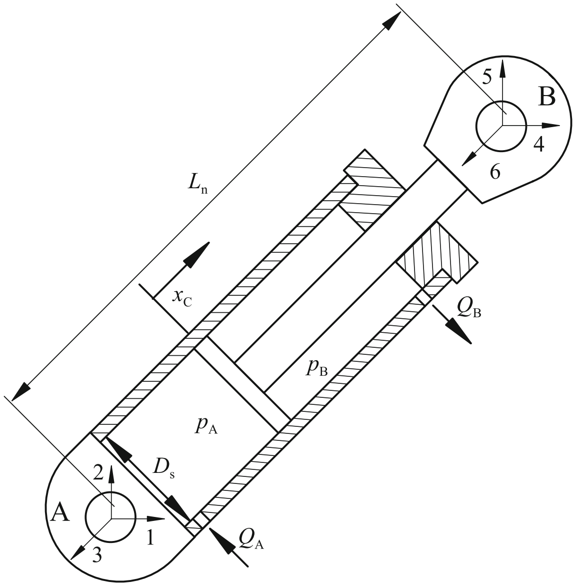
4.1 Truss-Element Cylinder

The truss-element-like cylinder (TC) model is closely related to the length-controlled rod element introduced in [11, 12]. However, the length-controlled rod element is a purely mechanical element, whereas the truss-element cylinder couples the hydraulic and mechanical systems. A similar hydraulic cylinder model was first introduced by Cardona and Géradin [3] where a truss element cylinder with a friction model was derived. A similar approach has been adopted by Bauchau and Liu [2], where models for the hydraulic components are given along with a model for a hydraulic cylinder with no friction and without explicit formulas for the tangent operators. The basic concept of modeling a hydraulic cylinder has also been given by Viersma [17], where the model is the same as in the more recent approaches. In this section an element with an improved friction model is described. In addition, the tangent operators and are presented in detail. For more datails and expressions for the mass matrix see [19, 20].

The truss cylinder has two nodes with only the translational degrees of freedom in each node, see Fig. 2. As can be seen from the figure a new variable x_c is introduced for the position of the cylinder piston. The piston position is related to the initial length L_0, and the current length of the cylinder L_n,

$$x_\mathrm{c} = L_\mathrm{n} - L_0. \tag{14}$$

Fig. 2 The hydraulic cylinder with main dimensions and variables



The current length of the cylinder is expressed as $L_\mathrm{n} = \sqrt{\mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}}$. The piston position x_c is a relative displacement, so in the initial state $x_\mathrm{c} = 0$, and the current nodal coordinates can be gathered into vector $\mathbf{x}$ as

$$\mathbf{x} = \begin{bmatrix} x_1 & y_1 & z_1 & x_2 & y_2 & z_2 \end{bmatrix}^\mathrm{T} \tag{15}$$

and the symmetric matrix $\mathbf{A}$ consisting of the 3×3 identity matrices as follows

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix}. \tag{16}$$

See also Fig. 3 for the definition of the displacements. The current and initial coordinates are related as $\mathbf{x} = \mathbf{X} + \mathbf{u}$ where $\mathbf{u}$ contains the 6 nodal displacements. Thus for the variation it can be written $\delta\mathbf{x} = \delta\mathbf{u}$ since the initial state remains constant.

To define the internal force vector for the cylinder element, the equilibrium of the cylinder piston using the chamber pressures can be written as follows, see [17, 19, 20] and Fig. 4

$$F_\mathrm{c} = p_\mathrm{A} A_\mathrm{A} - p_\mathrm{B} A_\mathrm{B} - F_\mu \tag{17}$$

where p_A and p_B are the chamber pressures in chambers A and B, respectively. The corresponding piston areas are denoted A_A and A_B.

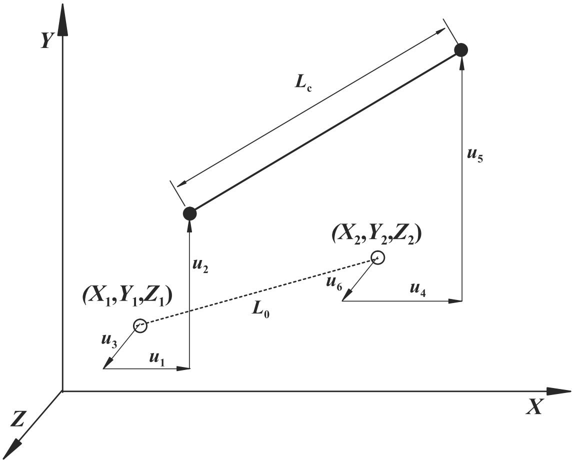


Fig. 3 Cylinder element with nodal displacements. L_c is the current length of the element

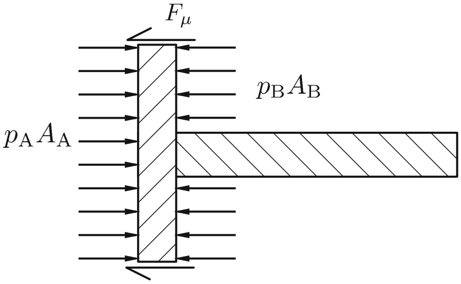


Fig. 4 Free body diagram of the truss element cylinder piston where the pressure forces are acting on the piston along with the friction force. It is assumed that the piston is moving to the right and therefore the friction force points to the left

The last term arises from the friction model. In this work the LuGre model is used [18], and the friction force is given as

$$F_\mu = k_0 z + k_1 \dot{z} + F_v v_t, \tag{18}$$

where k_0, k_1 are the stiffness coefficient and the damping coefficient related to the model. The average bristle deflection is denoted as z, whose evolution equation is given in the form

$$\dot{z} = v_t - \frac{|v_t|}{g(v_t)} z, \tag{19}$$

where v_t is the sliding velocity [18]. With the last term in (18), the viscous friction can be captured. The parametrization of function g then defines the LuGre model as

$$g(v_t) = \frac{1}{k_0} \left(F_C + (F_{st} - F_C) \exp\left(-\frac{v_t^2}{v_{Str}^2} \right) \right), \tag{20}$$

where F_C and F_st are the Coulomb friction and static friction respectively. The Stribeck velocity is denoted as v_Str.

The magnitude of the internal force is given in (17). To define the force direction, the unit vector in the direction of the cylinder is given as

$$\mathbf{n}_\mathrm{c} = \frac{\mathbf{x}_\mathrm{B} - \mathbf{x}_\mathrm{A}}{\|\mathbf{x}_\mathrm{B} - \mathbf{x}_\mathrm{A}\|}. \tag{21}$$

Using the matrix $\mathbf{A}$ and the denominator in (21) as the length of the cylinder element, the internal force vector can be written in the form

$$\mathbf{f}_\mathrm{int} = F_\mathrm{c} \begin{bmatrix} \mathbf{n}_\mathrm{c} \\ -\mathbf{n}_\mathrm{c} \end{bmatrix} = -\frac{F_\mathrm{c}}{L_\mathrm{n}} \mathbf{A}\mathbf{x}. \tag{22}$$

The internal force of the truss-element cylinder is in the direction of the element, and cannot capture the bending effects.

4.1.1 Cylinder Variables

In this section the differential equation, or the evolution law for the cylinder state variables is described.

First the cylinder variables is chosen. The technique presented in [17] is used to derive the diffrential equation for the hydraulic system. Since the LuGre friction model is chosen, the bristle deflection z is used as a state variable. In the case of hydraulic cylinder the bristle can be treated as the sealing ring deformation, see Fig. 5.

Now, the state equation for this cylinder model is written as

$$\dot{\mathbf{z}} = \begin{bmatrix} \dot{p}_\mathrm{A} \\ \dot{p}_\mathrm{B} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} B \dfrac{Q_\mathrm{A} - \dot{x}_\mathrm{c} A_\mathrm{A}}{V_\mathrm{A} + x_\mathrm{c} A_\mathrm{A}} \\ B \dfrac{-Q_\mathrm{B} + \dot{x}_\mathrm{c} A_\mathrm{B}}{V_\mathrm{B} - x_\mathrm{c} A_\mathrm{B}} \\ \dot{x}_\mathrm{c} - \dfrac{|\dot{x}_\mathrm{c}|}{g(\dot{x}_\mathrm{c})} z \end{bmatrix} \tag{23}$$

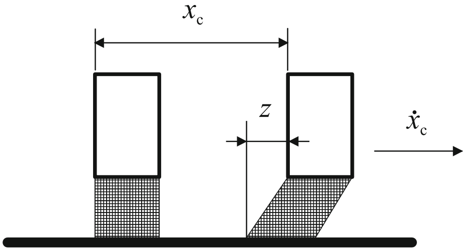


Fig. 5 The bristle deflection in cylinder application. The bristle can be treated as the piston sealing ring

So, the above equation explicitly defines the second equation in (4). The flow rates into the A and B chambers are denoted as Q_{A} and Q_{B}, and the positive flow directions are show in Fig. 2. The volumes V_{A} and V_{B} correspond to the volumes of the cylinder chambers when $x_{\mathrm{c}} = 0$.[1] The bulk modulus B is assumed to be constant, and $\dot{x}_{\mathrm{c}}$ is the piston velocity.

4.1.2 Dynamic Simulation

In this section derivation of the tangent matrices for the dynamic simulation model of the truss element cylinder is given. Only the two first equations in the set given in (4) is treated, thus the coupling with the hydraulic control system is not considered here.

Variations of the Internal Force

To shorten the equations, the variation of the internal force is written in two parts as follows

$$\delta \mathbf{f}_{\mathrm{int}} = \underbrace{\frac{\partial \mathbf{f}_{\mathrm{int}}}{\partial \mathbf{x}} \delta \mathbf{x} + \frac{\partial \mathbf{f}_{\mathrm{int}}}{\partial \dot{\mathbf{x}}} \delta \dot{\mathbf{X}}}_{\text{Part 1}} + \underbrace{\frac{\partial \mathbf{f}_{\mathrm{int}}}{\partial \mathbf{z}} \delta \mathbf{z}}_{\text{Part 2}} \tag{24}$$

where the first part yields to matrices given implicitly in (9)–(10), whereas the second term leads to a coupling matrix, see (13). The off-diagonal terms of the linearized equations of motion are referred as coupling matrices. The stiffness matrix related to the acceleration forces will be derived later.

Part 1: Taking the variation of (22) with the first term in (24) gives

$$\delta \mathbf{f}_{\mathrm{int}} = -\frac{F_{\mathrm{c}}}{L_{\mathrm{n}}} \mathbf{A} \delta \mathbf{x} - F_{\mathrm{c}} \mathbf{A} \mathbf{x} \delta \left(\frac{1}{L_{\mathrm{n}}} \right) - \frac{1}{L_{\mathrm{n}}} \mathbf{A} \mathbf{x} \delta F_{\mathrm{c}}. \tag{25}$$

The variation for the current length of the rod is given by $\delta(L_{\mathrm{n}}^2) = 2\mathbf{x}^T \mathbf{A} \delta \mathbf{x}$, and the variation for the inverse can be written as

$$\delta \left(\frac{1}{L_{\mathrm{n}}} \right) = -\frac{1}{L_{\mathrm{n}}^2} \delta L_{\mathrm{n}} = -\frac{1}{L_{\mathrm{n}}^3} \mathbf{x}^{\mathrm{T}} \mathbf{A} \delta \mathbf{x}. \tag{26}$$

Finally, the variation of the cylinder force δF_{c} with respect to the mechanical variables is needed. From (17) to (18) it is noticed that the only variable dependent on the mechanical system is the friction force F_{μ}, resulting in

[1] Since the piston position x_{c} is a relative quantity, the initial volume includes the dead volume as well as the volume of the piston displacement resulting from the initial stroke. The initial volumes are as given in Fig. 2.

$$\delta F_{\mathrm{c}} = -\delta F_{\mu}. \tag{27}$$

The variation of the friction force F_{μ} is taken from the (18)

$$\begin{aligned}\delta F_{\mu} &= (F_{\mathrm{v}} + k_1 H(z, \dot{x}_{\mathrm{c}}))\, \delta \dot{x}_{\mathrm{c}} \\ &= c_{\mu} \delta \dot{x}_{\mathrm{c}},\end{aligned} \tag{28}$$

where the function $H(z, \dot{x}_{\mathrm{c}})$ has the form

$$H(z, \dot{x}_{\mathrm{c}}) = 1 - \frac{z}{g(\dot{x}_{\mathrm{c}})} \operatorname{sign}(\dot{x}_{\mathrm{c}}) + \frac{g'(\dot{x}_{\mathrm{c}})}{g^2(\dot{x}_{\mathrm{c}})} |\dot{x}_{\mathrm{c}}| z \tag{29}$$

$$g'(\dot{x}_{\mathrm{c}}) = -\frac{2\dot{x}_{\mathrm{c}}}{k_0 v_{\mathrm{Str}}^2} (F_{\mathrm{st}} - F_{\mathrm{C}}) \exp\left(-\frac{\dot{x}_{\mathrm{c}}^2}{v_{\mathrm{Str}}^2}\right). \tag{30}$$

To write the variation for the cylinder piston velocity, time derivative of the cylinder piston position x_{c} is

$$\dot{x}_{\mathrm{c}} = \frac{1}{L_{\mathrm{n}}} \mathbf{x}^{\mathrm{T}} \mathbf{A} \dot{\mathbf{x}}, \tag{31}$$

resulting in

$$\begin{aligned}\delta \dot{x}_{\mathrm{c}} &= \left(\frac{1}{L_{\mathrm{n}}} \dot{\mathbf{x}}^{\mathrm{T}} \mathbf{A} - \frac{1}{L_{\mathrm{n}}^3} \mathbf{x}^{\mathrm{T}} \mathbf{A} \dot{\mathbf{x}} \mathbf{x}^{\mathrm{T}} \mathbf{A}\right) \delta \mathbf{x} + \frac{1}{L_{\mathrm{n}}} \mathbf{x}^{\mathrm{T}} \mathbf{A} \delta \dot{\mathbf{x}} \\ &= \mathbf{B}_{\mathrm{c}} \delta \mathbf{x} + \frac{1}{L_{\mathrm{n}}} \mathbf{x}^{\mathrm{T}} \mathbf{A} \delta \dot{\mathbf{x}}.\end{aligned} \tag{32}$$

Now it is possible to write the variation of the internal force for the cylinder element which includes the stiffness matrix and the damping matrix by collecting the equations above and substituting them to (25)

$$\begin{aligned}\delta \mathbf{f}_{\mathrm{int}} &= \left(-\frac{F_{\mathrm{c}}}{L_{\mathrm{n}}} \mathbf{A} + \frac{F_{\mathrm{c}}}{L_{\mathrm{n}}} \mathbf{A} \mathbf{x} \mathbf{x}^{\mathrm{T}} \mathbf{A} - \frac{c_{\mu}}{L_{\mathrm{n}}} \mathbf{A} \mathbf{x} \mathbf{B}_{\mathrm{c}}\right) \delta \mathbf{x} + \left(-\frac{c_{\mu}}{L_{\mathrm{n}}^2} \mathbf{A} \mathbf{x} \mathbf{x}^{\mathrm{T}} \mathbf{A}\right) \delta \dot{\mathbf{x}} \\ &= (\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)\, \delta \mathbf{x} + \mathbf{C}_{\mathrm{mm}} \delta \dot{\mathbf{x}} \\ &= \mathbf{K}_{\mathrm{mm}} \delta \mathbf{x} + \mathbf{C}_{\mathrm{mm}} \delta \dot{\mathbf{x}}.\end{aligned} \tag{33}$$

It is worth noticing that the damping matrix $\mathbf{C}_{\mathrm{mm}}$ and the geometric stiffness matrices $\mathbf{k}_1$ and $\mathbf{k}_2$ are symmetric, whereas the material stiffness matrix $\mathbf{k}_3$ is not symmetrical, see [19, 20]. The asymmetry arises from the variation of the piston velocity in (32). If the velocity of the cylinder nodes $\dot{\mathbf{x}}$ is in the direction of $\mathbf{x}$, the matrix $\mathbf{k}_3$ is also symmetric.

Part 2: Considering the part 2 in (24) the variation of the internal force with respect to the cylinder variables gives[2]

$$\begin{aligned}\delta F_{\mathrm{c}} &= \left[A_{\mathrm{A}} \; -A_{\mathrm{B}} \; \left(k_0 - k_1 \frac{|\dot{x}_{\mathrm{c}}|}{g(\dot{x}_{\mathrm{c}})} \right) \right] \delta \mathbf{z} \\ &= \mathbf{b}_{\mathrm{mc}} \delta \mathbf{z}. \end{aligned} \tag{34}$$

The second part of the variation in (24) can be expressed explicitly as

$$\delta \mathbf{f}_{\mathrm{int}} = -\frac{1}{L_{\mathrm{n}}} \mathbf{A} \mathbf{x} \mathbf{b}_{\mathrm{mc}} \delta \mathbf{z} = \mathbf{J}_{\mathrm{mc}} \delta \mathbf{z}. \tag{35}$$

Variations of the Cylinder State Equation

In a similar way the variationof the cylinder state equation is splitted into two parts

$$\delta \mathbf{f}_{\mathrm{cyl}} = \underbrace{\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \mathbf{x}} \delta \mathbf{x} + \frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \dot{\mathbf{x}}} \delta \dot{\mathbf{x}}}_{\text{Part 1}} + \underbrace{\frac{\partial \mathbf{f}_{\mathrm{cyl}}}{\partial \mathbf{z}} \delta \mathbf{z}}_{\text{Part 2}}, \tag{36}$$

where the first part yields the coupling terms $\mathbf{J}_{\mathrm{cm}}$ and $\mathbf{C}_{\mathrm{cm}}$, whereas the second part is the cylinder Jacobian denoted as $\mathbf{J}_{\mathrm{cc}}$, see (12).

Part 1: Variation of the cylinder state Eq. (23) with respect to the mechanical variables x_{c} and $\dot{x}_{\mathrm{c}}$ is

$$\begin{aligned}\delta \mathbf{f}_{\mathrm{cyl}} &= \begin{bmatrix} -B A_{\mathrm{A}} \dfrac{Q_{\mathrm{A}} - \dot{x}_{\mathrm{c}} A_{\mathrm{A}}}{(V_{\mathrm{A}} + x_{\mathrm{c}} A_{\mathrm{A}})^2} \\ B A_{\mathrm{B}} \dfrac{-Q_{\mathrm{B}} + \dot{x}_{\mathrm{c}} A_{\mathrm{B}}}{(V_{\mathrm{B}} - x_{\mathrm{c}} A_{\mathrm{B}})^2} \\ 0 \end{bmatrix} \delta x_{\mathrm{c}} + \begin{bmatrix} -B \dfrac{A_{\mathrm{A}}}{(V_{\mathrm{A}} + x_{\mathrm{c}} A_{\mathrm{A}})} \\ B \dfrac{A_{\mathrm{B}}}{(V_{\mathrm{B}} - x_{\mathrm{c}} A_{\mathrm{B}})} \\ H \end{bmatrix} \delta \dot{x}_{\mathrm{c}} \\ &= \mathbf{b}_{\mathrm{cm}} \delta x_{\mathrm{c}} + \mathbf{c}_{\mathrm{cm}} \delta \dot{x}_{\mathrm{c}}. \end{aligned} \tag{37}$$

By using Eqs. (26)–(32) results in

$$\begin{aligned}\delta \mathbf{f}_{\mathrm{cyl}} &= \left(\mathbf{c}_{\mathrm{cm}} \mathbf{B}_{\mathrm{c}} + \frac{1}{L_{\mathrm{n}}} \mathbf{b}_{\mathrm{cm}} \mathbf{x}^{\mathrm{T}} \mathbf{A} \right) \delta \mathbf{x} + \frac{1}{L_{\mathrm{n}}} \mathbf{c}_{\mathrm{cm}} \mathbf{x}^{\mathrm{T}} \mathbf{A} \delta \dot{\mathbf{x}} \\ &= \mathbf{J}_{\mathrm{cm}} \delta \mathbf{x} + \mathbf{C}_{\mathrm{cm}} \delta \dot{\mathbf{x}}. \end{aligned} \tag{38}$$

[2] Note that only F_{c} dependends on the cylinder variables.

Part 2: Variation of the second part in (36) is simply

$$\delta \mathbf{f}_{\text{cyl}} = \begin{bmatrix} 0 \; 0 & 0 \\ 0 \; 0 & 0 \\ 0 \; 0 & -\dfrac{|\dot{x}_{\text{c}}|}{g(\dot{x}_{\text{c}})} \end{bmatrix} \delta \mathbf{z} = \mathbf{J}_{\text{cc}} \delta \mathbf{z}. \tag{39}$$

4.2 Bending Flexible Hydraulic Cylinders

In real life, hydraulic cylinders can bent therefore a bending flexible cylinder (BF) model is developed. In addition, the rod arm slides inside the cylinder lining so a telescopic movement has to be modeled. The starting points in developing this element are to capture the bending flexibility and then modify the coupling between the mechanical variables and the cylinder variables accordingly. Moreover, the sliding between the cylinder members has to be accounted for.

The bending flexibility can be modeled if the members are treated as beam elements [10], and the sliding between the members can be captured by the C^1-continuous slide-spring element introduced in [13]. The beam model of the hydraulic cylinder is shown in Fig. 6, where 4 beams are used to model the cylinder lining and 4 beams to model the cylinder arm. The element mesh can be given separately for each member of the cylinder element. This super element is connected to the mechanical system via the attachment points located at the farthest end-nodes, marked as N1 and N2. In simulations, as in reality, the members are on top of each other, but for the sake of clarity they are shown in Fig. 6 as separated. In Fig. 6, the lower beam column is the lining and the upper one is the cylinder rod.

Unlike in the truss element cylinder, where the fluid volume was computed using the relative piston displacement x_{c}, the bending flexible cylinder element introduces the piston displacement as an absolute quantity. The fluid volume of the A chamber is between nodes 1 and 2, whereas that of the B chamber is between nodes 2 and 3. Thus the node 2 in Fig. 6 is identified as the cylinder piston, node 3 as the rod guide and node 1 is the bottom of the cylinder. The pressure evolution laws are written for the two chambers separately, and the pressure-induced forces are regarded as external

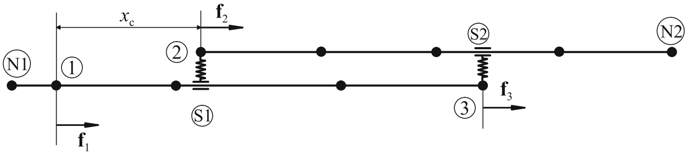


Fig. 6 A bending flexible hydraulic cylinder where the members are modeled as Reissner's geometrically exact beam elements. The element is attached to the mechanical system using nodes *N1* and *N2*

loads to the cylinder element, indicating there is no need to write the equilibrium equation for the cylinder piston.

The sliding between the members is then accounted for using the C^1-continuous slide-spring elements. The hydraulic cylinder has two sliding surfaces and therefore two slide-spring elements are used. Slide 1 corresponds to the cylinder piston sliding inside the lining, and Slide 2 to the arm sliding through the rod guide.

The slide-spring element is utilized in the bending flexible hydraulic cylinder. Since the spring is constrained between two nodes of a beam element, a special method has to be developed when the slide-spring changes from one beam to another. This procedure requires study of the slide degree of freedom, which is restricted to $s \in [0, 1]$. Therefore, whenever this requirement is violated the cylinder element needs to be re-meshed so that the requirement is still met.

Since the cylinder members are modeled as beam elements, the nodal degrees of freedom are translations and rotations. Therefore, it is possible to take the curvature of the cylinder into account when the pressure forces are computed. The pressure forces are always perpendicular to the surface, see [1]. The pressure forces are therefore follower forces, and they are dependent on the current configuration of the cylinder.

4.2.1 Cylinder Variables

First, the frictionless cylinder element is described, and thus only the pressure variables are included. Thus the state equation for the cylinder variables is simply

$$\dot{\mathbf{z}} = \begin{bmatrix} \dot{p}_{\mathrm{A}} \\ \dot{p}_{\mathrm{B}} \end{bmatrix} = \begin{bmatrix} B\dfrac{Q_{\mathrm{A}} - \dot{x}_{\mathrm{c}} A_{\mathrm{A}}}{V_{\mathrm{A0}} + x_{\mathrm{c}} A_{\mathrm{A}}} \\ B\dfrac{-Q_{\mathrm{B}} + \dot{x}_{\mathrm{c}} A_{\mathrm{B}}}{V_{\mathrm{B0}} - x_{\mathrm{c}} A_{\mathrm{B}}} \end{bmatrix}. \tag{40}$$

The cylinder piston position x_{c} is defined as an absolute distance from Node 1 in Fig. 6, unlike in the truss element cylinder. Therefore, the definition of the volumes for the A and B chambers, V_{A0} and V_{B0} respectively, also have to be changed accordingly. The dead volumes are then the actual ones at $x_{\mathrm{c}} = 0$ and $x_{\mathrm{c}} = x_{\mathrm{c\_max}}$, respectively. Otherwise the $\dot{\mathbf{z}}$ is similar to the one for the truss element cylinder.

The cylinder piston position x_{c} is now computed from the distance between Nodes 1 and 2

$$x_{\mathrm{c}} = \sqrt{\mathbf{x}_{\mathrm{c}}^{\mathrm{T}} \mathbf{A} \mathbf{x}_{\mathrm{c}}}, \tag{41}$$

where the vector $\mathbf{x}_{\mathrm{c}}$ collects the current nodal coordinates of Nodes 1 and 2, see Fig. 6. This length is then used to define the fluid volume in the cylinder chambers. It should be noted that rotations are not included in the vector $\mathbf{x}_{\mathrm{c}}$. Defining the displacement vector $\mathbf{u}$ as

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \ \boldsymbol{\Psi}_1 \ \mathbf{u}_2 \ \boldsymbol{\Psi}_2 \end{bmatrix}^{\mathrm{T}}, \tag{42}$$

where the subscripts still refer to Fig. 6. Since these two nodes belong to beam elements, it is possible to find the translational freedoms separately from the rotational freedoms. Since vector $\mathbf{x}_\mathrm{c}$ contains only translations, the mapping between this vector and the vector $\mathbf{u}$ is defined as

$$\mathbf{x}_\mathrm{c} = \mathbf{L}\mathbf{u} \tag{43}$$

where the linear mapping matrix $\mathbf{L}$ has the form

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{44}$$

Using this mapping the variation for the displacement vector $\mathbf{x}_\mathrm{c}$ as $\delta\mathbf{x}_\mathrm{c} = \mathbf{L}\delta\mathbf{u}$ can be found.

In (41) it is assumed that the fluid volume remains straight although this is not actually the case. However, the effect is negligible, as it is shown in [19].

4.2.2 Dynamic Simulation

Derivation of the bending flexible cylinder element is based on the geometrically exact beam model described in detail in [10]. Therefore, the appropriate equations for the fluid fields in the cylinder chambers are derived here. To shorten the presentation, derivations of the Jacobian matrices for the A chamber are given, since the pressure forces can be treated separately for both chambers. Similar derivations can then be performed for the B chamber.

To start with, the pressure force vector acting on Nodes 1 and 2 in Fig. 6 is written as

$$\mathbf{f}_\mathrm{int} = \left[\mathbf{f}_1\ \mathbf{0}\ \mathbf{f}_2\ \mathbf{0}\right]^\mathrm{T}, \tag{45}$$

where the force vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ are

$$\mathbf{f}_1 = -p_\mathrm{A} A_\mathrm{A} \mathbf{R}_1 \mathbf{E}_\mathrm{c} \tag{46}$$

$$\mathbf{f}_2 = p_\mathrm{A} A_\mathrm{A} \mathbf{R}_2 \mathbf{E}_\mathrm{c}. \tag{47}$$

In the above equation the cylinder unit vector in the initial state is denoted with $\mathbf{E}_\mathrm{c}$, and the rotation matrices corresponding to the nodal rotation vectors with $\mathbf{R}_1$ and $\mathbf{R}_2$ for Nodes 1 and 2, respectively. The rotation matrix is computed from the rotation vector using the well known Rodrigues' formula. The zero terms in the internal force vector correspond to the rotations but since no moments arise from the fluid pressure, the terms are set to zero. To obtain the tangent operators, the variations with respect to the nodal translations and rotations according to (42) are taken. Then, the Jacobian matrices arising from the derivation are added to the corresponding degrees of freedom.

Variations of the Internal Force

The stiffness matrix and the coupling terms are now derived from the internal force vector. Writing the implicit variation with respect to the beam element nodal displacement vector as

$$\delta \mathbf{f}_{\text{int}} = \underbrace{\frac{\partial \mathbf{f}_{\text{int}}}{\partial \mathbf{u}}\delta \mathbf{u} + \frac{\partial \mathbf{f}_{\text{int}}}{\partial \dot{\mathbf{u}}}\delta \dot{\mathbf{u}}}_{\text{Part 1}} + \underbrace{\frac{\partial \mathbf{f}_{\text{int}}}{\partial \mathbf{z}}\delta \mathbf{z}}_{\text{Part 2}}, \tag{48}$$

from which can be noted that Part 1 yields to a stiffness and damping matrix, and Part 2 to a coupling term. Since the internal force vector is not dependent on the velocities, there is no damping matrix.

Part 1: The first part of the variation leads to the stiffness matrix, and it is simply computed using the variations presented in [10], thus giving,

$$\delta \mathbf{f}_{\text{int}} = \begin{bmatrix} p_{\text{A}} A_{\text{A}} \mathbf{R}_1 \widetilde{\mathbf{E}}_{\text{c}} \mathbf{T}_1 \delta \boldsymbol{\Psi}_1 \\ \mathbf{0} \\ -p_{\text{A}} A_{\text{A}} \mathbf{R}_2 \widetilde{\mathbf{E}}_{\text{c}} \mathbf{T}_2 \delta \boldsymbol{\Psi}_2 \\ \mathbf{0} \end{bmatrix} = \mathbf{K}_{\text{mm}} \delta \mathbf{u}. \tag{49}$$

This stiffness is added to the translation freedoms of Nodes 1 and 2. A similar equation for the B chamber can also be written, where different rotation vectors are used, as well as the pressure of the B chamber and the corresponding area. The stiffness matrix in the component form is rather peculiar because no diagonal terms arise. The stiffness matrix is

$$\mathbf{K}_{\text{mm}} = p_{\text{A}} A_{\text{A}} \begin{bmatrix} \mathbf{0} & \mathbf{R}_1 \widetilde{\mathbf{E}}_{\text{c}} \mathbf{T}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{R}_2 \widetilde{\mathbf{E}}_{\text{c}} \mathbf{T}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{50}$$

Part 2: The second part of the variation, where the internal force vector is varied with respect to the chamber pressures is simply

$$\delta \mathbf{f}_{\text{int}} = \delta \mathbf{f}_{\text{int}} = A_{\text{A}} \left[-\mathbf{R}_1 \mathbf{E}_{\text{c}} \ \mathbf{0} \ \mathbf{R}_2 \mathbf{E}_{\text{c}} \ \mathbf{0} \right]^{\text{T}} \delta p_{\text{A}} = \mathbf{J}_{\text{mc}} \delta p_{\text{A}}. \tag{51}$$

Similar derivations are then performed for the B chamber to obtain the corresponding Jacobian matrices.

Variations of the Cylinder State Equation

To derive the cylinder Jacobian and the other coupling matrix, the variation of $\mathbf{f}_{\text{cyl}}$ has the form

$$\delta\mathbf{f}_{\mathrm{cyl}} = \underbrace{\frac{\partial\mathbf{f}_{\mathrm{cyl}}}{\partial\mathbf{u}}\delta\mathbf{u} + \frac{\partial\mathbf{f}_{\mathrm{cyl}}}{\partial\dot{\mathbf{u}}}\delta\dot{\mathbf{u}}}_{\text{Part 1}} + \underbrace{\frac{\partial\mathbf{f}_{\mathrm{cyl}}}{\partial\mathbf{z}}\delta\mathbf{z}}_{\text{Part 2}}, \tag{52}$$

from which the first part yields a coupling matrix, whereas the second part is the cylinder Jacobian $\mathbf{J}_{\mathrm{cc}}$.

Part 1: The variation of the first part yields two coupling terms because the cylinder state is dependent on both the displacement x_{c} and the velocity $\dot{x}_{\mathrm{c}}$. Therefore it is required to derive the variations for both. The variation for x_{c} is taken from (41). This variation has been given in (26). The variation for $\dot{x}_{\mathrm{c}}$ is given in (32) by substituting L_{n} with x_{c}.

$$\begin{aligned}
\delta\mathbf{f}_{\mathrm{cyl}} &= \begin{bmatrix} -BA_{\mathrm{A}}\dfrac{(Q_{\mathrm{A}} - A_{\mathrm{A}}\dot{x}_{\mathrm{c}})}{(V_{\mathrm{A0}} + A_{\mathrm{A}}x_{\mathrm{c}})^2} \\ BA_{\mathrm{B}}\dfrac{(-Q_{\mathrm{B}} + A_{\mathrm{B}}\dot{x}_{\mathrm{c}})}{(V_{\mathrm{B0}} + A_{\mathrm{B}}(L_{\mathrm{p}} - L_{\mathrm{ap}} - x_{\mathrm{c}}))^2} \end{bmatrix}\delta x_{\mathrm{c}} \\
&\quad + \begin{bmatrix} -\dfrac{A_{\mathrm{A}}B}{V_{\mathrm{A0}} + A_{\mathrm{A}}x_{\mathrm{c}}} \\ \dfrac{A_{\mathrm{B}}B}{V_{\mathrm{B0}} + A_{\mathrm{B}}(L_{\mathrm{p}} - L_{\mathrm{ap}} - x_{\mathrm{c}})} \end{bmatrix}\delta\dot{x}_{\mathrm{c}} \\
&= \mathbf{b}_{\mathrm{cm}}\delta x_{\mathrm{c}} + \mathbf{c}_{\mathrm{cm}}\delta\dot{x}_{\mathrm{c}}.
\end{aligned} \tag{53}$$

By substituting the variations of x_{c} and $\dot{x}_{\mathrm{c}}$ the coupling matrices are obtained

$$\begin{aligned}
\delta\mathbf{f}_{\mathrm{cyl}} &= \left(\frac{1}{x_{\mathrm{c}}}\mathbf{b}_{\mathrm{cm}}\mathbf{x}_{\mathrm{c}}^{\mathrm{T}}\mathbf{A} + \mathbf{c}_{\mathrm{cm}}\mathbf{B}_{\mathrm{c}}\right)\delta\mathbf{x}_{\mathrm{c}} + \frac{1}{x_{\mathrm{c}}}\mathbf{c}_{\mathrm{cm}}\mathbf{x}_{\mathrm{c}}^{\mathrm{T}}\mathbf{A}\delta\dot{\mathbf{x}}_{\mathrm{c}} \\
&= \left(\frac{1}{x_{\mathrm{c}}}\mathbf{b}_{\mathrm{cm}}\mathbf{x}_{\mathrm{c}}^{\mathrm{T}}\mathbf{A} + \mathbf{c}_{\mathrm{cm}}\mathbf{B}_{\mathrm{c}}\right)\mathbf{L}\delta\mathbf{u} + \frac{1}{x_{\mathrm{c}}}\mathbf{c}_{\mathrm{cm}}\mathbf{x}_{\mathrm{c}}^{\mathrm{T}}\mathbf{A}\mathbf{L}\delta\dot{\mathbf{u}} \\
&= \mathbf{J}_{\mathrm{cm}}\delta\mathbf{u} + \mathbf{C}_{\mathrm{cm}}\delta\dot{\mathbf{u}},
\end{aligned} \tag{54}$$

where the relation between $\mathbf{x}_{\mathrm{c}}$ and $\mathbf{u}$ has been used.

Part 2: The second part of the differentiation is also zero since the pressure evolutions are not dependent on the chamber pressures themselves, thus $\mathbf{J}_{\mathrm{cc}} = \mathbf{0}$.

The Jacobian matrices related to displacements and velocities have been presented above. Expressions for the matrices are simpler than they were with the truss element cylinder. This is because there is no need to write the equilibrium equation of the cylinder piston. In addition the equations can be derived separately for the A and B chambers.

4.3 Bending Flexible Hydraulic Cylinder with Friction

The bending flexible hydraulic cylinder (BF) model derived above did not have the friction model used with the truss element cylinder. Consequently, the LuGre model was also implemented into the C^1-continuous slide-spring element, so that it is possible to model frictional slides. The friction model is implemented into the slide-spring element, which is presented in its original form in [13]. The frictional C^1-continuous slide spring element is then applied to modeling the hydraulic cylinder, as presented in the previous section.

Here it is emphasized that, apart from the friction model, the formulation for this frictional bending flexible cylinder element is the same as for the frictionless one introduced in the previous section. The difference is in the cylinder state equation, where the bristle deflection is included as a separate variable. The state equations concerning the cylinder chamber pressures need not to be modified in any way. Therefore it is only required to derive the Jacobian matrices of the friction model and use them alongside the ones from the frictionless formulation.

4.3.1 Cylinder Variables

First, the cylinder variables for the frictional cylinder are defined. This task is merely combination of (23)–(40), with two friction variables. Thus, the cylinder variables now consist of the rates of chamber pressures and of the time rates of the bristle deflections as

$$\dot{\mathbf{z}} = \begin{bmatrix} \dot{p}_{\mathrm{A}} \\ \dot{p}_{\mathrm{B}} \\ \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} B\dfrac{Q_{\mathrm{A}} - \dot{x}_{\mathrm{c}} A_{\mathrm{A}}}{V_{\mathrm{A0}} + x_{\mathrm{c}} A_{\mathrm{A}}} \\ B\dfrac{-Q_{\mathrm{B}} + \dot{x}_{\mathrm{c}} A_{\mathrm{B}}}{V_{\mathrm{B0}} - x_{\mathrm{c}} A_{\mathrm{B}}} \\ v_1 - \dfrac{|v_1|}{g(v_1)} z \\ v_2 - \dfrac{|v_2|}{g(v_2)} z \end{bmatrix}, \tag{55}$$

where the actual slide velocities for the friction forces are denoted with v_i and the subscript i indicates the slide number; see also Fig. 6. The actual sliding velocity is computed as

$$v_i = L_{\mathrm{ele}} \dot{s}_i, \tag{56}$$

where $\dot{s}_i$ is the material time derivative of the sliding degree of freedom and L_{ele} is the current length of element.

4.3.2 Cylinder Formulation with LuGre Friction

The degrees of freedom for the C^1-continuous slide-spring element are $\mathbf{u}_{c1} = [\mathbf{u}_1\ \boldsymbol{\Psi}_1\ \mathbf{u}_2\ \boldsymbol{\Psi}_2\ \mathbf{u}_3\ \boldsymbol{\Psi}_3\ s]^{\mathrm{T}}$. Here, only interest is only given to the last term, which is the slide's position on the beam. The values of s are restricted between $s \in [0, 1]$.

In Sect. 4.2 equations for the frictionless bending flexible cylinder are derived. As it can be seen from (55), the friction variables are added to the cylinder state equation as separate variables. Therefore, only the modifications required in order to utilize the friction model are presented. The hydraulic coupling is precisely the same as in the previous cylinder model. The Jacobians of the friction law and the model for a general slide are derived. These equations can then be utilized in the systems with n slides.

Variations of the Friction Force

The scalar friction force is treated as an external force into the slide and for the LuGre model has the expression

$$F_{\mu_i} = (k_0 z_i + k_1 \dot{z}_i + F_{\mathrm{v}} v_i)\,, \tag{57}$$

where the same friction model coefficients for both slides are assumed. The index i denotes the slide number. In sequel, the subscript is dropped and the equations are derived only for the first slide. The Jacobian matrix for the friction force follows from the first variation

$$\delta F_\mu = (F_{\mathrm{v}} + k_1 H)\,\delta v + \left(k_0 - k_1 \frac{|v|}{g(v)}\right)\delta z = c_\mu \delta v + c_{\mathrm{z}} \delta z. \tag{58}$$

Now, the Jacobian matrices are only scalars. However, for notational reasons it is denoted: $\mathbf{C}_{\mathrm{mm}} = c_\mu$ and $\mathbf{J}_{\mathrm{mc}} = c_{\mathrm{z}}$.

Variations of the Cylinder State Equation

Since there is no coupling between the pressures and the friction variables in (55), the variation is simply

$$\delta \dot{z} = \delta v - \frac{|v|}{g(v)}\delta z - z\delta\left(\frac{|v|}{g(v)}\right) = H(z, v)\delta v - \frac{|v|}{g(v)}\delta z, \tag{59}$$

where the function $H(z, v)$ is taken from (29) and is the derivative of the function g with respect to the velocity. The Jacobian terms are $\mathbf{C}_{\mathrm{cm}} = H(z, v)$ and $\mathbf{J}_{\mathrm{cc}} = -|v|/g(v)$.

The initial state of this frictional bending flexible cylinder model can be computed similarly as for the truss element cylinder, see [19, Sect. 4.2.3]. The only difference is that the actual displacement of the slide is used as the argument for the friction function. The pressure scaling factor has also been included in the static friction model.

5 Integration of the Coupled Two-Field Problem

Time integration of a coupled mechanical and hydraulic system has been discussed by [3], where a multi-rate integration scheme is developed. In the multi-rate scheme the hydraulic state equation is integrated with a different time step and a different integrator than the mechanical system. The idea of multi-rate integration has been discussed in a conceptual sense in [1]. In the multi-rate integration, shorter time steps are used for the hydraulic state equation than used for the mechanical system. For a different strategy of the multi-rate integration see [14].

The coupled mechanical-hydraulic system is stiff and highly damped, therefore the standard Crank-Nicolson rule is not an ideal integrator, see [8, 9]. Therefore the L-stable Rosenbrock method is chosen in the present study and it is described briefly in the next section.

5.1 *The Rosenbrock Method*

Semi-implicit methods have been shown to work well for stiff differential equations [6]. The Rosenbrock method is based on an implicit form, however, instead of using an iterative solution for the linearized equations, only one iteration is performed with the Runge-Kutta type time-stepping, see [15].

In order to utilize the Rosenbrock method, the equations of motion is written as a system of two first order differential equation systems

$$\mathbf{H}\begin{bmatrix}\dot{\mathbf{z}}\\ \dot{\mathbf{q}}\\ \dot{\mathbf{v}}\end{bmatrix} = \begin{bmatrix}\mathbf{f}_{\text{cyl}}(\mathbf{z},\mathbf{q},\dot{\mathbf{q}},t)\\ \mathbf{v}\\ \mathbf{g}(\mathbf{q},\dot{\mathbf{q}},\mathbf{z},t)\end{bmatrix}, \tag{60}$$

where $\mathbf{v} = \dot{\mathbf{q}}$ and the matrix $\mathbf{H}$ is $\mathbf{H} = \operatorname{diag}(\mathbf{I}, \mathbf{I}, \mathbf{M})$. The system (60) is briefly written as

$$\mathbf{H}\dot{\mathbf{x}} = \mathbf{r}(t, \mathbf{x}). \tag{61}$$

The Rosenbrock scheme is a diagonally implicit Runge-Kutta method (DIRK), where advancing in time for an autonomous problem is performed as

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \sum_{i=1}^{s} b_i \mathbf{k}_i. \tag{62}$$

where

$$\mathbf{k}_i = \Delta t\ \mathbf{r}\left(\mathbf{x}_n + \sum_{j=1}^{i-1} \alpha_{ij}\mathbf{k}_j + \alpha_{ii}\mathbf{k}_i\right) \qquad i = 1, \ldots, s, \tag{63}$$

where s is the number of stages. However, this form is implicit. For the Rosenbrock method this equation is linearized and only one iteration is performed for solving the vectors $\mathbf{k}_i$, for details see [6, 15]. In contrast to purely explicit methods, the Jacobian matrix need to be computed only at the beginning of the time step x = x_n. Therefore only one solution of the system (61) is necessary within a time step.

For non-autonomous systems, an additional term is added to the basic form in (63) to account for the time dependency, see [6]. After linearization, the form for solving vectors $\mathbf{k}_i$ is written

$$\begin{aligned}(\mathbf{H} - \gamma\Delta t\mathbf{J})\,\mathbf{k}_i = \Delta t\mathbf{r}\left(t_n + \alpha_i\Delta t,\ \mathbf{x}_n + \Delta t\sum_{j=1}^{i-1}\alpha_{ij}\mathbf{k}_j\right)& \\ + \gamma_i\Delta t^2\frac{\partial\mathbf{r}(t_n, \mathbf{x}_n)}{\partial t} + \Delta t\mathbf{J}\sum_{j=1}^{i-1}\gamma_{ij}\mathbf{k}_j,&\end{aligned} \tag{64}$$

where $\mathbf{J}$ is the full Jacobian having the following block structure

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{\mathrm{cc}} & \mathbf{J}_{\mathrm{cm}} & \mathbf{C}_{\mathrm{cm}} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ -\mathbf{J}_{\mathrm{mc}} & -\mathbf{K}_{\mathrm{mm}} & -\mathbf{C}_{\mathrm{mm}} \end{bmatrix}. \tag{65}$$

It is possible to utilize the symmetry of the matrices by using the block factorization scheme, see [12]. For a two-stage Rosenbrock scheme the coefficients α_{ij} and γ_{ij} are

$$\gamma_{ij} = \begin{bmatrix} \gamma & 0 \\ -\gamma & \gamma \end{bmatrix} \qquad \alpha_{ij} = \begin{bmatrix} 0 & 0 \\ 1/2 & 0 \end{bmatrix} \tag{66}$$

with $\gamma = 1 - 1/\sqrt{2}$, see [16]. In addition the nonzero coefficients of $b_i\ \alpha_i$ are $\alpha_2 = 1/2,\ b_1 = 1/4,\ b_2 = 3/4$.

Since the Rosenbrock method does not require corrector iterations, it is computationally efficient. However, the L-stability is only guaranteed if the Jacobian matrices are correct which is important in eliminating the erroneous high frequency vibrations from the system [6, 15].

6 Numerical Examples

In this section behaviour of the two formulations; the truss element and the bending flexible cylinder are compared. Two test cases for a boom movement are given and the initial values used for both examples are presented in Table 1.

6.1 *Influence of the Stribeck Effect*

The boom is initially in a horizontal position. A mass of 400 kg is then placed at the boom's tip. The hydraulic cylinder element is positioned between nodes D and B, see Fig. 7. After the initial state is solved by the Newton's scheme, the dynamic simulation is started with the Rosenbrock scheme using a constant time-step of 0.001 s. Procedure for the initial state computation is described in [20].

The flow rate for the hydraulic cylinder is defined under the premise that the fluid is incompressible giving the inbound flow rate of $Q_{\mathrm{in}} = v_{\mathrm{e}} A_{\mathrm{A}} = 15.6\ \mathrm{l/min}$. The outbound flow rate is computed according to the model presented in [4].

In the following example the influence of the Stribeck effect to the behaviour of the system response is investigated in a lifting motion of a boom. The Stribeck velocity is given in Table 2 and the Coulomb friction is now set to 70 % of the static friction.

In Fig. 8 comparison of the boom stresses for both the truss-element cylinder, TC, and the bending flexible cylinder, BF, are shown. The "S" denotes the model where the Stribeck effect is included. At first glance, the peak stress in the boom does not appear to be affected by the Stribeck effect. However, when comparing the frictional cylinder models with the ones where the Stribeck effect is included, more oscillation can be noticed. Although the peak stress value remains the same, the increased oscillation could be a factor in any fatigue assessment of the structure, so it is useful to capture the effects of the changing friction force.

The friction forces and the friction parameters in Fig. 9 are plotted separately for both cylinder models, once again, both with and without friction and the Stribeck effect. The oscillation with a low frequency is due to the mass displacement, whereas the higher frequency oscillation is due to the stresses in the boom system. Because a corresponding high amplitude oscillation in the friction forces cannot be detected, a similar high-amplitude oscillation in the boom stresses is naturally absent.

Table 1 The initial values for the lifting boom example

Dimension	Symbol	Value	Unit
Steel density	ρ_s	7850	kg/m^3
Steel Young's modulus	E_s	200	GPa
Beam height	h_b	0.12	m
Beam width	w_b	0.08	m
Wall thickness	t_b	0.005	m
Beam linear density	ρ_b	14.92	kg/m^3
Point mass a the boom end	m	400	kg
Lining length	L_p	0.85	m
Rod length	L_r	0.85	m
Attachment element length	L_ap	0.1	m
Cylinder attachment point	h_0	1.0	m
Outer diameter of lining	D_u	0.08	m
Inner diameter of lining	D_s	0.07	m
Dead volume of A chamber	V_A	0.0029	m^3
Bulk modulus	B_oil	2000	MPa
Rod diameter	D_r	0.042	m
Spring stiffness coefficient	k_x	1×10^8	N/m
Spring stiffness coefficient	$k_\mathrm{y}, k_\mathrm{z}$	9×10^7	N/m
Spring stiffness coefficient	$k_\mathrm{rx}, k_\mathrm{ry}, k_\mathrm{rz}$	0	N/m

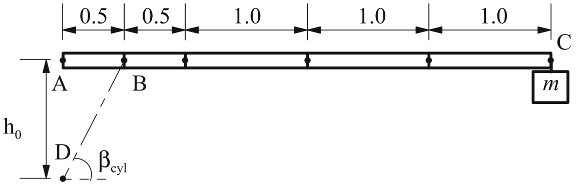


Fig. 7 Boom system with an external load at the tip of the boom

6.2 Discussion

In comparing the truss-element cylinder (TC) and the bending flexible cylinder (BF), in this example the bending flexibility does not seem to to have significant effect to the results. The axial deformations of the bending flexible cylinder model create

Table 2 Parameters for the friction model

Quantity	Symbol	Value	Unit
Static friction	F_{st}	3	kN
Coulomb friction	F_{C}	2.1	kN
Viscous friction	F_{v}	0	Ns/m
Stiffness coefficient	k_0	746.5	kN/m
Damping coefficient	k_1	864	Ns/m
Stribeck velocity	v_{Str}	0.1	m/s

Values are the same as given in [18]

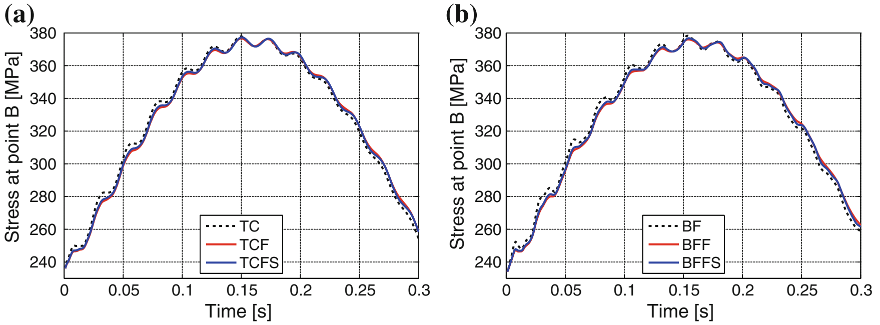


Fig. 8 Comparison of boom stresses with the truss element cylinder (TC) and the bending flexible cylinder (BF). Inclusion of friction and the Stribeck effect is labeled as F and S, respectively

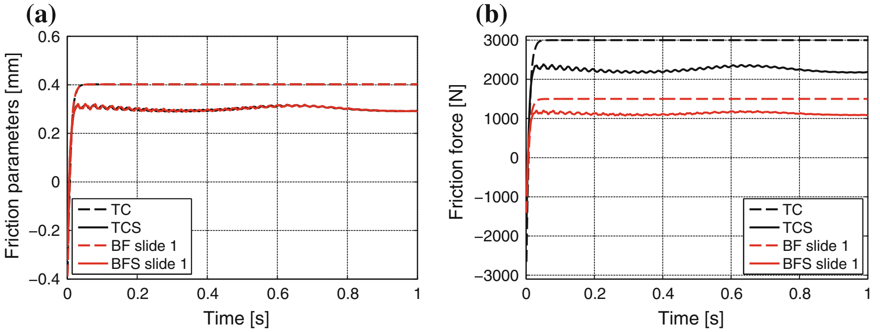


Fig. 9 **a** Comparison of friction parameters and **b** forces with the TC and BF cylinders with and without friction and with the Stribeck effect included

slightly higher initial deformations and the lowest natural frequency is slightly lower due to the increased flexibility. However, these factors can be compensated for by reducing the bulk modulus of the hydraulic oil in the truss-element cylinder.

The Stribeck effect was not as evident in this numerical example. The slide velocity is not as easily defined, and the variations are smaller but faster, thus leading to smaller variations in the friction force. The effects are more local, and related to the slides rather than to the response of the complete boom system.

For this first example, it can be concluded by saying that when the bending flexibility of the cylinder is minor, it is more advantageous to use the truss-element cylinder. The truss-element cylinder has 6 degrees of freedom for the mechanical system, and it has 3 cylinder variables if friction is taken into account. In our numerical example, 4 beam elements were used for both the lining and rod in the meshing of the bending flexible hydraulic cylinder. This resulted in 62 mechanical degrees of freedom, and when friction is included, 4 cylinder variables. Therefore, the truss-element cylinder is computationally more efficient.

6.3 Sudden Stop of a Boom

For the second example, the same boom is used, but instead of analyzing the lifting motion the boom's initial angle is set to 70° and an accident situation in which the boom is allowed to free-fall for 0.5 s is analysed. The point D in Fig. 7 is moved so that the cylinder angle is $\beta_{\text{cyl}} = 83^\circ$ when the boom angle is 70°. The initial length for the cylinder is 2 m, and the cylinder lining and cylinder rod is elongated to 1.2 m. In addition, the mass at the end of the boom is increased to 600 kg.

After defining the initial state the boom is put into a free fall. The flow rate out of chamber A is computed using the modified orifice model, see [4, 19], and the pressure outside the cylinder chamber is 1 bar. For the B chamber, t the pressure is set to zero whenever the pressures fall below 0 bars. The free fall lasts for 0.5 s. Then, all flow rates are stopped in and out of both cylinder chambers and scrutinize the sudden stop. The truss-element cylinder and the bending flexible cylinder are compared through the boom stresses, the displacements at point B, and the mass point.

6.3.1 System Responses with minuscule: Without Friction

The mass and point B displacements are shown in Fig. 10. In the initial state at time $t = 0$ s, both the mass displacement and the point B displacement are higher with the bending flexible cylinder. This is due to the increased flexibility of this cylinder formulation, which was also noted in the previous example.

The displacements for point B and the mass are similar in both the cylinder formulations when the boom is in free fall, but the differences start to become apparent when the flow rates in and out of the cylinder are set to zero. Then, it is observed that with the bending flexible cylinder, point B deflects more in a vertical direction. The truss-element cylinder only transmits energy to the pressurized fluid, since only the fluid is considered incompressible and only axial effects are included in this element. The bending flexible cylinder element, however, is allowed to bend, thus

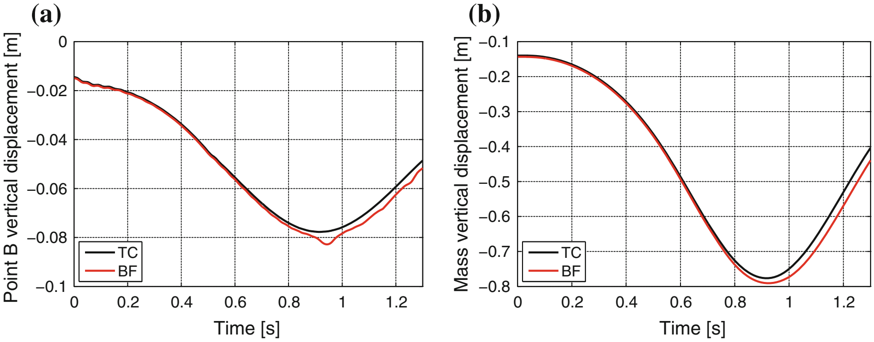
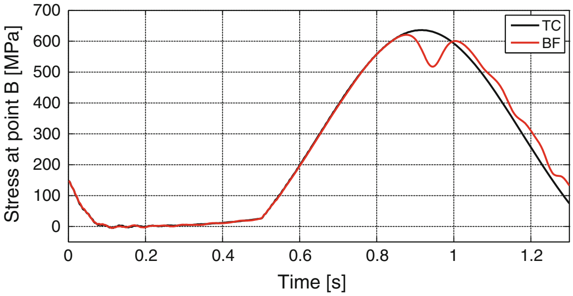


Fig. 10 Vertical displacements of point B and the mass when the sudden stop occurs. The boom is in free fall for 0.5 s and after that all flow rates are set to zero

Fig. 11 Boom stresses at point B when sudden stop is observed with the truss element cylinder and bending flexible cylinder without friction



allowing point B to travel more vertically than it does in the truss-element cylinder. This bending effect is clearly visible when $t = 0.95$ s.

It can be seen from Fig. 10b that the vertical displacement of the mass point is also higher when the simulation is carried out using the bending flexible cylinder, for the same reasons. The bending flexibility allows the mass to slow down over a longer distance, and this should also have an effect on the boom stresses.

The bending stresses of the top edge of point B for both cylinders are shown in Fig. 11. The stresses in the initial state are almost identical, although they are slightly higher with the bending flexible cylinder, due to its increased flexibility (Fig. 13).

During the free fall, both cylinder models showed similar stresses at point B, although the stress oscillated slightly more with the bending flexible model. As far as stress is concerned, the responses were almost identical until 0.85 s had elapsed, after which the bending flexible cylinder results in lower boom stresses. As already mentioned, point B is allowed to translate in a more vertical direction with the bending flexible cylinder. This gives the boom more room in which to stop, which is reflected in the reduction in the boom stresses.

The maximum cylinder chamber pressures for the two cylinders are given as 520 bars for the truss-element cylinder and 500 bars for the bending flexible cylinder. Because the cylinder bends, instead of the piston merely sliding within the lining,

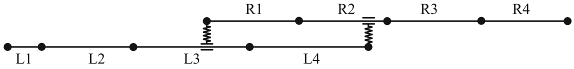


Fig. 12 Element numbering of the bending flexible hydraulic cylinder. The slides are positioned such that they appear during the highest stresses

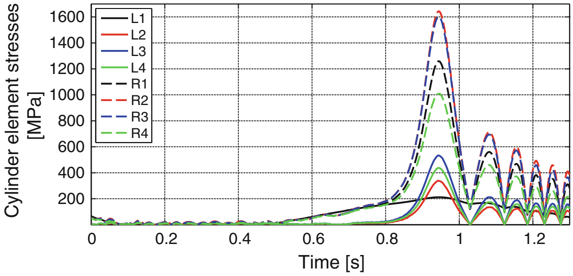


Fig. 13 Stresses of the beam elements along the hydraulic cylinder. Here L denotes elements of the lining and R elements of the rod according to Fig. 12. Stresses are maximum tensile stresses

the pressures are reduced. The A chamber pressures follow the curve given for the stresses in Fig. 11.

Figure 12 shows the element numbering and Fig. 13 the results of all the elements along the cylinder lining and the rod. From this figure it can be seen that the bending flexible cylinder has stresses all along the cylinder, with the maximum stress occurring at the cylinder rod. The bending of the cylinder element can be clearly seen at $t = 0.8$ s, where there is a sudden increase in the stresses. This is exactly at the point where it can be seen a drop in the boom stresses in Fig. 11. The highest stresses occur in the first two elements of the cylinder rod, and as they happen to be the slide-spring element and the following element, this implies that they experience bending. After the first stress peak, however, lower stresses along the beams can be noticed of the cylinder element because the boom stresses also start to oscillate.

6.4 System Responses with Friction

In the previous section the sudden stop of a boom without friction is analyzed. In this section friction is included in the simulations. The friction parameters are as given in Table 2 therefore the Stribeck effect is not included in this simulation.

When the friction force is included, it can be assumed that the boom stresses will be lower during the stop since the friction force lowers the acceleration of the mass during a free fall. Studying Fig. 14 shows the displacements at point B for both cylinder models separately, both with and without friction. With friction included, the initial state displacement is lower for both cylinder models, as was also seen in the previous examples. Notice also that the velocity of point B is indeed

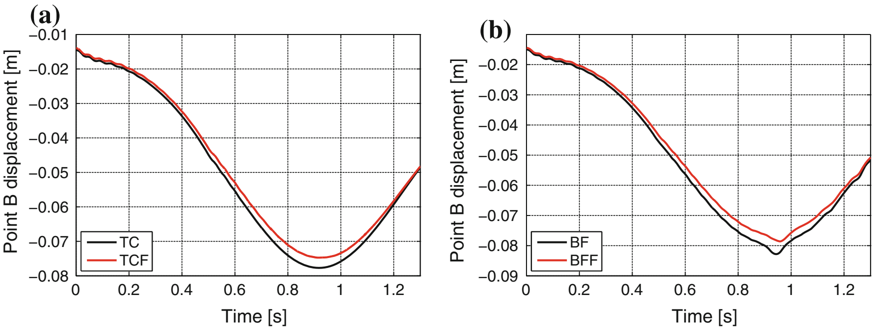


Fig. 14 Point B displacements: **a** for the truss-element cylinder (TC) and **b** for the bending flexible cylinder (BF) with (xxF) and without friction

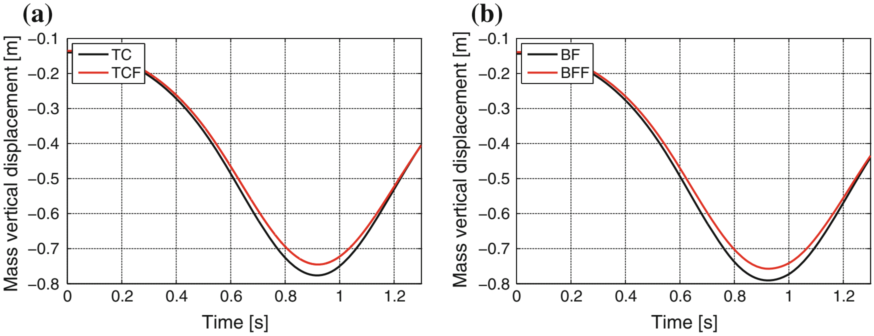


Fig. 15 Displacements of the mass with and without friction: **a** for the TC cylinder and **b** for the BF cylinder

lower in the frictional analysis for both cylinder models, since the gap between the curves representing the frictionless and frictional analysis increases as the simulation progresses.

When the cylinder orifices are closed, and the boom comes to a sudden stop after $t = 0.5$ s, similar results as with the frictionless cylinder are obtained. However, the displacements are now reduced since the friction force reduces the velocity of the mass, and thus the kinetic energy of the whole system. The bending flexible cylinder also bends, thus allowing point B to travel a greater distance than it did with the TC model. In Fig. 15, the mass displacement, shows a similar response.

Since smaller displacements for both point B and the mass are obtained, it may be assumed that also find lower boom stresses due to the reduced kinetic energy will be obtained. The boom stresses are given for both cylinder formulations, with and without friction in Fig. 16, and these confirm the assumption about reduced boom stresses. Although the reduction is 4 % for the truss element cylinder and 3.3 % for the bending flexible cylinder model, the boom stress responses are similar for both

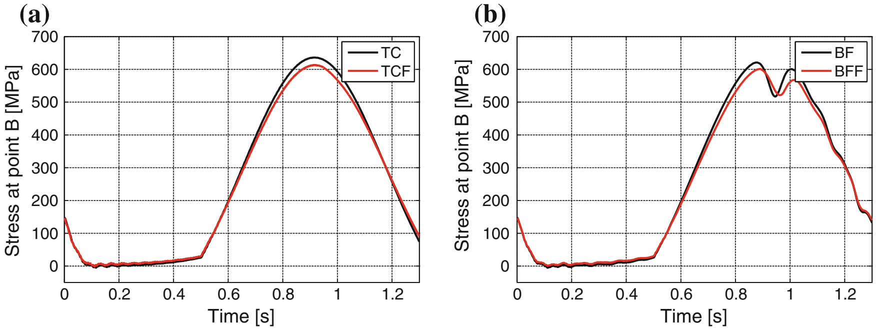
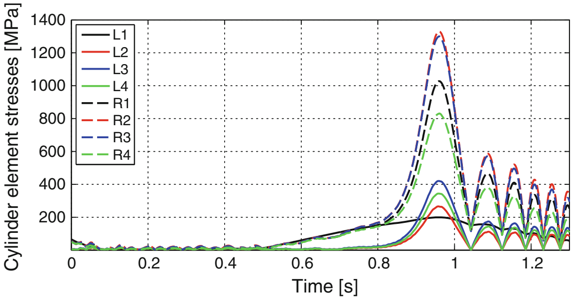


Fig. 16 Boom bending stresses at the top edge of point B with the two cylinder formulations with and without friction

Fig. 17 Stresses of the beam elements along the hydraulic cylinder. Here L denotes elements of the lining and R elements of the rod according to Fig. 12



the truss-element cylinder and the bending flexible cylinder, both with and without friction.

As with the previous example, the cylinder stresses are shown in Fig. 17. The lower kinetic energy is reflected in the stresses along the cylinder element, where there is a reduction of 23 % in the highest stress value. Nevertheless, the element with highest stress is the same. The chamber pressures in the frictional setting read 500 bars for the truss-element cylinder and 490 bars for the bending flexible cylinder.

7 Concluding Remarks

The literature does not cover the modeling of hydraulic cylinders for multibody simulations extensively. In this chapter, the two hydraulic cylinder models developed in [19] are presented. The elements are derived to be compatible with finite elements, and the initial state can be computed using the Newton scheme since the cylinder variables are embedded in the mechanical variables. On the other hand, the cylinder models to be used in dynamic simulations introduce new variables.

When the hydraulic cylinder is modeled using either the TC or BF models, the hydraulic cylinder model works as an interface between the hydraulic control system and the mechanical system. Therefore, the input for the mechanical system from the state of the hydraulic control system can be obtained.

So far any conclusive remarks about a suitable time integration scheme from this study cannot be made. The coupled system is stiff and highly dissipative, which is why an efficient time integration scheme is required for the time-stepping. The concept of multi-rate integration could prove to be a good approach.

The bending flexible cylinder model could be improved by using a mixed formulation approach to also account for the pressure forces on the cylinder lining. In this study, these forces are neglected completely. The pressure inside the cylinder chamber tends to swell the cylinder thus increasing the lining stresses, but it also affects the friction by reducing the pre-stressing of the seals. Accounting for the clearance is also a factor to be included when computing a reliable critical load for buckling. In particular, these factors should be included when cylinders are being designed.

This study shows that by modeling the hydraulic cylinder with a specialized element can have an impact on the system response. The proposed cylinder elements proposed here work within the framework of finite elements, and the results are very promising.

References

1. Bathe KJ (1996) Finite Element Procedures. Prentice Hall
2. Bauchau O, Liu H (2006) On The Modeling of Hydraulic Components in Rotorcraft Systems. Journal of the American Helicopter Society 51(2):175–184
3. Cardona A, Géradin M (1990) Modeling of a hydraulic actuator in flexible machine dynamics simulation. Mechanism and Machine Theory 25(2):193–207
4. Ellman A, Piché R (1999) A two regime orifice flow formula for numerical simulation. Journal of Dynamic Systems, Measurement and Control, Transactions of the ASME 121(4):721–724
5. Géradin M, Cardona A (2001) Flexible Multibody Dynamics: A Finite Element Approach. J. Wiley & Sons
6. Hairer E, Wanner G (1991) Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems. Springer Series in Computational Mathematics 14, Springer
7. Holzapfel GA (2000) Nonlinear Solid Mechanics - A Continuum Approach for Engineering, 1st edn. John Wiley & Sons
8. Ibrahimbegović A, Mamouri S (1999) Nonlinear dynamics of flexible beams in planar motion: Formulation and time-stepping scheme for stiff problems. Computers & Structures 70(1):1–22
9. Ibrahimbegović A, Mamouri S (2002) Energy conserving/decaying implicit time-stepping scheme for nonlinear dynamics of three-dimensional beams undergoing finite rotations. Computer Methods in Applied Mechanics and Engineering 191(37–38):4241–4258
10. Mäkinen J (2007) Total Lagrangian Reissner's geometrically exact beam element without singularities. International journal for numerical methods in engineering 70(9):1009–1048
11. Marjamäki H, Mäkinen J (2003) Modelling telescopic boom - the plane case: Part I. Computers & Structures 81(16):1597–1609
12. Marjamäki H, Mäkinen J (2006) Modelling a telescopic boom - the 3D case: Part II. Computers & Structures 84(29-30):2001–2015
13. Marjamäki H, Mäkinen J (2009) Total Lagrangian beam element with C^1-continuous slide-spring. Computers & Structures 87:534–542

14. Naya M, Cuadrado J, Dopico D, Lugris U (2011) An efficient unified method for the combined simulation of multibody and hydraulic dynamics: Comparison with simplified and co-integration approaches. Archive of Mechanical Engineering 58(2):223–243
15. Piché R (1995) An L-stable Rosenbrock method for step-by-step time integration in structural dynamics. Computer Methods in Applied Mechanics and Engineering 126(3–4):343–354
16. Shampine L, Reichelt M (1997) The MATLAB ode suite. SIAM Journal on Scientific Computing 18(1):1–22
17. Viersma TJ (1980) Analysis, Synthesis, and Design of Hydraulic Servosystems and Pipelines. Elsevier Scientific Publishing Company
18. Canudas de Wit C, Olsson H, Astrom K, Lischinsky P (1995) New model for control of systems with friction. IEEE Transactions on Automatic Control 40(3):419–425
19. Ylinen A (2015) Hydraulic cylinder models for flexible multibody system simulation. PhD thesis, Department of Mechanical Engineering and Industrial Systems, Tampere University of Technoogy
20. Ylinen A, Marjamäki H, Mäkinen J (2014) A hydraulic cylinder model for multibody simulations. Computers & Structures 138:62–72