Upscaling, analysis, and iterative numerical solution schemes for thermo-poroelasticity

Mats Kirkesæther Brun

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2019

UNIVERSITY OF BERGEN

Upscaling, analysis, and iterative numerical solution schemes for thermo-poroelasticity

Mats Kirkesæther Brun

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 28.06.2019

Year: 2019

Title: Upscaling, analysis, and iterative numerical solution schemes for thermo-poroelasticity

Name: Mats Kirkesæther Brun

Print: Skipnes Kommunikasjon / University of Bergen

Preface

This dissertation is submitted as a partial fulfillment of the requirements for the degree of Doctor Philosophy (PhD) at the University of Bergen. The advisory committee has consisted of Florin Adrian Radu (University of Bergen), Jan Martin Nordbotten (University of Bergen, Princeton University) and Inga Berre (University of Bergen).

The PhD project has been financially supported by the Norwegian Research Council Toppforsk project 250223 (The TheMSES project: https://themses.w.uib.no).

Acknowledgements

Doing a PhD is an interesting experience. In the beginning you feel like you are wandering around in a forest at night without any sense of direction, but by working systematically there eventually comes a time when you start to get a sense of what it is you are actually doing. Frustrations and setbacks are plentiful, but it is still a learning experience like no other, and in the end you would not have been without it.

There are many people to whom I owe my sincere gratitude and appreciation. Firstly, I would like to thank my main advisor Florin A. Radu for the continuous support and excellent supervision during my entire PhD study. I would also like to thank my two co-advisors, Inga Berre and Jan Martin Nordbotten, whose positive attitude, inspiring words and immense knowledge has been crucial for the completion of this research and for writing this dissertation. I could not have asked for a better advisory committee.

I also owe a special thanks to Elyes Ahmed, whose entry into the Porous Media Group as a post-doc one year after I started my PhD resulted in a significant improvement in the quality of my research. I am immensely grateful for our collaboration, which has resulted in two papers so far, and hopefully we will continue to collaborate also in the future. To all the rest of my colleagues in the Porous Media Research Group, I am sincerely grateful for the welcoming and friendly environment we have had both at work and also socially. I feel very lucky to have been a part of this research group for the past three years.

I would also like to thank Thomas Wick for hosting me during my six week exchange at the Leibniz University in Hannover, and for the enthusiasm and inspiring attitude with which our collaboration was approached.

Last but not least, I would like to thank my family; my two brothers and my two sisters, and especially my mom and my dad, for always being supportive and always believing in me, throughout the PhD study and also in life in general.

Abstract

The main objectives of this research project is to provide part of the mathematical models and simulation technology required to assess large-scale deployment of *thermo-mechanical subsurface energy storage* in the context of *intermittent renewable energy*. Within this overarching framework, two main topics are discussed in this dissertation: *Thermo-poroelasticity*, i.e., the coupling of *geomechanics*, *flow*, and *heat* within a porous material, and *phase field brittle fracture propagation*, i.e., the temporal and spatial tracking of brittle fracture evolution by creating a diffusive zone around fracture surfaces through an auxiliary variable known as a *phase field*. These subjects are highly relevant for e.g., applications involving injection of heated fluids at high pressures into the subsurface, where the increased pressure may cause a fracturing of the rock matrix, in addition to the induced temperature gradient affecting the poroelastic properties of the surrounding medium.

The part of this dissertation focused on thermo-poroelasticity can again be separated into three parts; (1) *modeling*, (2) analysis and (3) *numerical implementation*. In part (1), formal upscaling techniques (i.e., *homogenization*) are employed in order to derive the constitutive thermo-poroelastic equations from the known equations governing the physical processes at the pore-scale. Homogenization is a well-known and trusted technique, which is applicable in situations where the physical processes in question can be viewed from several scales, and where there is some uniformity or periodicity on the smaller scale. Within the context of porous media, the two relevant scales are the *pore-scale* and the *macro-scale*. Viewed from the pore scale, a porous medium consists of solid grains (obeying the laws of solid mechanics), and a fluid which is saturating the space in between the grains (obeying the laws of fluid mechanics), where the two processes are coupled at the mutual interface (i.e., at the grain surfaces). At the macro-scale, however, the porous medium is considered as a homogenous material obeying its own set of physical laws, but which is still somehow implicitly dependent on (or rather, the culmination of) the processes which are taking place at the pore-scale. Assuming some uniformity in the distribution of grains (formally, *periodicity*), the technique of homogenization allows for deriving the macro-scale model equations from the micro-scale model equations.

Part (2) is concerned with analyzing the thermo-poroelastic system derived in part (1). The existence and uniqueness of a weak solution to this model problem is established in the fully mixed formulation, under some natural assumptions on the regularity

of the source and initial data, as well as imposing some constraints on the effective coefficients. Since the upscaled thermo-poroelastic system is a nonlinear one, this analysis is done in two steps: First, a linearized system is analyzed using a standard *Galerkin technique* together with the *weak compactness* properties of the relevant function spaces. Finally, the previously analyzed linearized system is used to design an iterative scheme in order to approximate the original nonlinear system. By employing a contraction argument, the convergence of this iterative scheme is proved, which implies the existence and uniqueness of a weak solution to the full nonlinear problem.

Part (3) covers the numerical analysis and implementation of the thermo-poroelastic system derived in part (1). Here, six different iterative algorithms are proposed, all based on the linearization technique employed in part (2) when analyzing the full nonlinear system, as well as the stabilized splitting scheme known informally as the 'L-scheme'. These six algorithms involve different combinations of coupling / decoupling of the three subproblems involved (flow, mechanics and heat), i.e., at each iteration either a linearized system is solved monolithically, two subproblems are solved together decoupled from the third, or all three subproblems are decoupled. As such, these six algorithms exhaust all possibilities of coupling / decoupling of the three subproblems. The convergence of all six algorithms are proved using a contraction argument, similar to the one employed in part (2), except that only the fully discrete formulation is considered. Several numerical tests validate the robustness and efficiency of the algorithms. Furthermore, the performance of all six algorithms are compared with respect to a wide range of different physical regimes (i.e., with respect to various coupling strengths between the three subproblems).

Finally, the last part of this dissertation concerns brittle fracture propagation in a quasi-static elastic medium, where the fracture evolution is tracked by a phase field variable. In particular, the numerical approximation of such models. Phase field brittle fracture problems are notoriously difficult to solve, and currently no universally accepted method exists. Hence, this work involves designing a novel iterative algorithm for brittle fracture phase field models, analyzing its convergence, and testing it in detail with several numerical benchmark problems. The proposed algorithm is based on a linearization as well as stabilization of the model, where the two subproblems (*phase field* and *mechanics*) are solved separately at each iteration, while sharing updated solution information. Under the natural conditions that the mechanical elastic energy remains bounded, and that the diffusive zone around crack surfaces must be sufficiently thick, monotonic convergence of the proposed scheme is proved. These properties are also confirmed by the extensive numerical tests.

Outline

This dissertation consists of two parts. Part I gives an overview of the scientific theory and mathematical methods that are relevant to this research project. Part II consists of the papers that constitute the scientific results.

Part I is organized as follows: In Chapter 1, an introduction to the thesis is given, where the different research topics are discussed, and placed within the overarching framework of thermo-mechanical energy storage. In Chapter 2, the mathematical models employed in this thesis are discussed: First, an introduction to poroelasticity is given, with an emphasis on the linear system of equations known as Biot's quasi-static consolidation model. The extension of this system to the non-isothermal case is also discussed. Next, the theory of phase fields is presented. A brief overview of the general theory is given, followed by a more detailed discussion about application of phase fields for brittle fracture propagation. Next follows two chapters devoted to two specific mathematical techniques which are central to research done in this dissertation: First, in Chapter 3, the theory of homogenization is presented, specifically the method known as 'two-scale asymptotic expansions'. A model homogenization problem is then gone through is some detail. Next, in Chapter 4 an overview of iterative numerical methods is presented. Two specific such methods are then discussed in more detail; Newton's method and the Fixed Stress Splitting / L-scheme methods. Here, there is also provided a detailed example where the L-scheme is used to design an algorithm for a nonlinear coupled model problem. A convergence proof of this algorithm is also given, which demonstrates how convergence rates usually are derived in the context of Fixed Stress Splitting / L-scheme type methods. Chapter 5 contains the introductions to the papers A–D. Finally, in Chapter 6 a summary of the dissertation is given, in addition to some conclusions and also a discussion regarding future outlook.

Part II contains the scientific results, consisting of the following four papers:

Paper A Brun, Mats K and Berre, Inga and Nordbotten, Jan M and Radu, Florin A (2018). Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium. In *Transport in Porous Media*, 124(1), p. 137–158, Springer.

Paper B Brun, Mats Kirkesæther and Ahmed, Elyes and Nordbotten, Jan Martin and Radu, Florin Adrian (2019). Well-posedness of the fully

coupled quasi-static thermo-poroelastic equations with nonlinear convective transport. In *Journal of Mathematical Analysis and Applications*, 471(1–2), p. 239–266, Elsevier.

Paper C Brun, Mats Kirkesæther and Ahmed, Elyes and Berre, Inga and Nordbotten, Jan Martin and Radu, Florin Adrian (2019). Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. *In review (2019)*.

Paper D Brun, Mats Kirkesæther and Wick, Thomas and Berre, Inga and Nordbotten, Jan Martin and Radu, Florin Adrian (2019). An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters. *In review (2019)*.

Contents

Part I

Scientific Background

Chapter 1

Introduction

The purpose of this chapter is to give an introduction to the different topics discussed in this dissertation, and furthermore, to place these within the overarching framework of the research project as a whole, which is *geothermal energy storage*, and in particular *thermo-mechanical subsurface energy storage*.

In general, the subsurface consists of more or less permeable rock or sediment types, and if the permeability is large enough, fluid can flow freely through these structures. In the simplest description, (single phase) flow in the subsurface is modeled by a single conservation law, which is the conservation of fluid mass flux (Darcy flow). However, for geothermal applications, this model is too simple, and heat transfer must also be included into the aforementioned model. Specifically, natural heat transfer in the subsurface occurs primarily due to diffusion and convection [9]. Thus, to account for this, an energy conservation equation is introduced which is coupled to the mass conservation equation through a convective transport term (i.e., the transfer of heat with the movement of fluid particles) [19]. The energy conservation equation may also be coupled back to the mass conservation equation by including fluid density variations with temperature. A common simplification in this regard is the *Boussinesq approximation* (see e.g., [56] for further details). If the fluid movement is only due to density differences of the fluid in different spatial regions, then the flow is said to be a *natural convection current*. These density differences in the fluid are due the fact that when heated, the fluid becomes less dense. This spatial temperature variation is called the *geothermal gradient*, and is the driving force of the natural convection currents. Moreover, when the natural convection currents form closed curves, they are called *natural convection cells*. Such natural convection cells are very important for applications of geothermal energy storage in the subsurface, since these currents can be taken advantage of, and, given favorable conditions, can keep injected fluids heated for long periods of time.

However, in these descriptions of the subsurface, the permeable solid skeleton is considered as fixed in space and time. A natural extension is therefore to include the elastic response of the solid matrix into the modeling. This becomes very important when considering applications involving injection/extraction of heated fluids at high pressures,

such as e.g., thermo-mechanical subsurface energy storage. In particular, when simulating subsurface-processes in the proximity of injection/extraction wells, where the pressure differences are very large, a non-rigid solid skeleton may be necessary to include in the mathematical model for an accurate simulation. Moreover, if the pressure differences are large enough, an elastic solid may even be too simple a model, since fracturing of the matrix can occur. This then necessitates further complexities in the modeling, such as e.g., phase fields.

1.1 Geothermal energy storage

Thermo-mechanical subsurface energy storage is the process of injecting hot fluids at high pressures into a subsurface reservoir, where upon extraction at a later time, both thermal and mechanical energy may be recovered. This strategy is especially attractive in areas where there is great variability in the energy output coming from intermittent renewable energy: At times when there is a surplus in the energy output, effective means for long time storage of this energy is available, and at a later time when there is a shortage in the energy output, this stored energy may be recovered and utilized as electrical power. Furthermore, areas where there additionally are favorable geological conditions (such as permeable layers in the subsurface, depleted oil and gas reservoirs, or saline aquifers) are especially well suited.

Upon extraction of the injected fluids, there will necessarily be a significant loss during the conversion from thermo-mechanical energy to electric power. However, this loss can be mitigated by employing thermo-mechanical energy storage in areas where a large fraction of the power demand covers heating (hence, no such conversion is necessary), or in areas where there is a lot of waste heat generated from e.g., industrial processes (thus avoiding the conversion from electrical to thermo-mechanical energy before injection into the subsurface). However, in order to optimize the effectiveness of thermo-mechanical energy storage, there is also needed accurate and reliable models of how the induced temperature gradients and increased pressures, which occur when injecting heated fluids at high pressures, affect the natural processes in the subsurface. In fact, even more fundamental is a good understanding of the natural heat transfer mechanisms which are taking place in the subsurface from before.

The key processes which must be accounted for in any accurate mathematical model in this context is fluid flow, heat transfer and mechanical deformation (both elastic and plastic, i.e., fracturing). These three processes are fully coupled, i.e., each one is depending on the other two in some way. This leads to challenging problems, both from the modeling and analysis point of view, but also from the point of view of numerical simulations. This dissertation concerns all of these topics, and hence provides new insight into the fundamental science which is necessary for large-scale deployment of thermo-mechanical subsurface energy storage.

1.2 Relevant mathematical models

The simplest model for the coupling of flow and elastic deformation in a porous medium, i.e., *poroelasticity*, is the famous Biot's quasi-static consolidation model, which will be discussed in some detail in the next chaper. In short, this is a linear model which accounts for the fact that an increase in fluid pressure induces a dilation of the pores, which again results in an elastic response of the solid matrix. Compared to the Darcy flow model discussed above, there is now also a conservation of momentum equation, and the flow (mass conservation) equation is modified to account for the fact that fluid mass flux is balanced by a change in porosity. Biot's model is also readily extended to include thermal effects, thus resulting in a *thermo-poroelastic* system, i.e., the coupling of flow, elastic deformation and heat transfer within a porous medium. Even with constant fluid density, this system exhibits a fully coupled structure between all three processes involved. A significant part of this dissertation concerns different aspects of thermo-poroelasticity. In particular, the derivation of a thermo-poroelastic system using formal upscaling techniques, the mathematical analysis of this system, and finally its numerical implementation.

The subsurface is not a homogenous environment, but rather a combination of different rock and sediment types, each with different material characteristics, resulting in discontinuities in both variables and coefficients for any large-scale simulation. Fractures in the matrix are special in this regard since, due to their small aperture, they can be regarded as two-dimensional surfaces within a three-dimensional domain. The modeling of a fractured reservoir thus entails dealing with surfaces of discontinuity in the computational mesh. A common way to do this is to build the defect (fracture) into the computational mesh from the beginning. However, injecting fluids at high pressures into the subsurface may cause a fracturing of the matrix, or further extend the existing fractures. The modeling of such phenomena then means simulating the spatial and temporal evolution of a lower dimensional surface, over which the variables in question are discontinuous. The theory of *phase fields* is commonly employed in the modeling of fracture evolution, since the phase field creates a diffusive transition zone around fracture surfaces, thus avoiding the problem of an evolving lower dimensional surface. Phase fields, especially in relation to brittle fracture propagation, will also be discussed in the following sections. However, the evolution of fractures are herein only considered within the context of a quasi-static elastic material, and not a (thermo-)poroelastic material.

Chapter 2

Mathematical Models

In this chapter, and overview of the mathematical models employed in this research project are presented. First, *poroelasticity* is discussed. In particular, the two-field coupled linear model which is known as *Biot's quasi-static consolidation model*. This model has served as the backdrop during this research project for the extension to *thermo-poroelasticity* (i.e., to a non-isothermal Biot-type model), which is discussed next. A derivation of the Biot model is presented briefly in the context of mass and momentum conservation. However, in this derivation, some linearized constitutive laws are considered as given and not explained in full detail (for a justification of these, see e.g. [23, 34, 36, 47, 86]). The non-isothermal extension of the Biot system is also discussed, but not in the context of conservation laws. A more involved derivation using upscaling techniques is needed, which is in fact the strategy used in this research project. Finally, an overview of the theory of *phase fields* is given, with emphasis on the use of phase fields for *brittle fracture propagation within a quasi-static elastic material*. Finally, the advantages as well as the limitations of this theory are also discussed.

2.1 Poroelasticity

The field of *poroelasticity* concerns the interaction between elastic mechanical deformation and viscous fluid flow within a porous material. Poroelasticity can therefore be seen as a combination of porous media flow (diffusion) and classical linear elasticity, i.e., the fluid flow (or more accurately, the fluid pressure) influences the elastic mechanical deformation, and vice versa. This results in a coupled system of equations, which is quite similar to the classical thermoelastic system (this comparison will be outlined more carefully in the next sections). A number of comprehensive textbooks related to the field of poroelasticity exists, see e.g. [36, 38, 94].

Fully detailed (or resolved) descriptions of saturated porous media is in general quite complicated, i.e., the fluid and solid are considered as separate physical domains wherein separate physical processes occur, but coupled through boundary conditions at a mutual interface with a complicated geometry. For this reason, *effective descrip-*

tions are most frequently employed. This means that the fluid saturated porous material is considered as a homogenous medium, even though it is in fact microscopically heterogenous. The quantities of interest, such as elastic displacement and fluid pressure, must therefore be interpreted as averaged or *effective* quantities, and single points in the model domain must be interpreted as *representative elementary volumes* (*REVs*). Poroelasticity has many similarities with solid mechanics, although a notable difference (even in the simplest description) is the additional variable needed which is the *fluid pressure*. Thus, in addition to the elastic momentum conservation equation there is also needed a mass conservation equation. Further complications are also frequently considered, such as two- (or multi-) phase flow [37, 83, 84], unsaturated flow [63, 81, 81], reactive flow [19, 20, 57], or thermal flow [18, 21, 59, 93] (or any combination of the above).

A common assumption in poroelasticity is that of *quasi-static deformation*. This is essentially the same as assuming inertia effects are negligible. When considering consolidation of a linearly elastic porous medium, which is saturated by a slightly compressible viscous fluid, the time scale of the consolidation process is much longer than that of the fluid flow, and thus the quasi-static assumption arises naturally for a wide range of real-world scenarios. The classical mathematical model for linear poroelasticity is known as *Biot's model for quasi-static deformation*, which will be discussed in the next section. An extension of this model which will also be discussed is that of *thermo-poroelasticity*. In particular, a thermo-elastic solid is saturated by a slightly compressible viscous fluid where heat is transported both by convection and diffusion. In this sense, thermo-poroelasticity can be seen as both an extension of poroelasticity to the non-isothermal case, and also as an extension of classical thermoelasticity to a fully saturated porous material. (Thermo-)poroelasticity is an important subject with several applications, e.g., geothermal energy storage, enhanced oil recovery, nuclear waste disposal, and biomedical applications, to name a few.

2.1.1 Biot's consolidation model

In this section the conservation equations and constitutive laws which model fluid diffusion in a linearly elastic medium with negligible inertia effects, i.e. Biot's model for quasi-static consolidation, is presented. To this end, we denote the fluid pressure by $p(x,t)$, and the displacement vector of the solid structure by $\mathbf{u}(x,t)$. The model is derived as a set of conservation laws, combined with some constitutive relationships specific to poroelasticity. The following presentation is based on [36, 80, 88].

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be an elastic permeable body with density ρ (assumed to be constant), saturated by a slightly compressible viscous fluid. For an arbitrary control volume $V \subset \Omega$, we have that the momentum of the corresponding part of the solid matrix is given by

$$\rho \int_V \frac{\partial \mathbf{u}}{\partial t}(x,t)\mathrm{d}x. \tag{2.1}$$

On the boundary of the control volume V there is traction forces applied by the remaining

Figure 2.1: Belgian-American applied physicist M. A. Biot (1905-1985). Biot made important contributions in geophysics, thermodynamics and engineering. Most notably, he wrote several founding works on poroelasticity. Picture from [12].

part of the body, i.e. by $\Omega \setminus V$, given by

$$\int_{\partial V} \sigma \nu \mathrm{d}s, \tag{2.2}$$

where σ is the symmetric tensor valued poroelastic stress field, the components of which represents the forces on arbitrary surfaces within Ω, and where ν is the outward unit normal vector field of ∂V (the product $[\sigma(x,t)\nu(x)]_i = \sigma_{ij}(x,t)\nu_j(x)$ thus represents the traction at the point $x \in \partial V$ at time t). If $\mathbf{f}(x,t)$ is a vector representing the external body forces acting on Ω, then the principle of conservation of momentum says that the change in momentum of some arbitrary control volume must be balanced by the traction forces acting on the same control volume, in addition to any external body forces. We can thus write conservation of momentum for V as

$$\rho \frac{\partial}{\partial t} \int_V \frac{\partial \mathbf{u}}{\partial t}(x,t)\mathrm{d}x = \int_{\partial V} \sigma \nu \mathrm{d}s + \int_V \mathbf{f}(x,t)\mathrm{d}x. \tag{2.3}$$

Since the control volume V is arbitrary within Ω, and by the Divergence Theorem, we can write the above in differential form as

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} - \nabla \cdot \sigma = \mathbf{f}, \quad \text{for all} \quad x \in \Omega, t > 0, \tag{2.4}$$

Let now $\eta(x,t)$ denote the fraction of the point (i.e., the *REV*) $x \in \Omega$ at time t which is occupied by fluid, i.e., the *fluid content*, or equivalently, the *porosity*. The quantity of fluid within the control volume V is then obtained by the following integral

$$\int_V \eta(x,t)\mathrm{d}x. \tag{2.5}$$

The instantaneous flow rate of fluid relative to the solid matrix (i.e., *the rate of fluid mass per area*) is the fluid flux, denoted by the vector $\mathbf{q}(x,t)$. The rate at which fluid accumulates (or leaves) the control volume V is then given by integrating the normal component of $\mathbf{q}$ over ∂V, i.e.

$$\int_{\partial V} \mathbf{q} \cdot \nu \mathrm{d}s. \tag{2.6}$$

If $h(x,t)$ is a source density of fluid, then the principle of conservation of mass says that the change in fluid mass in some arbitrary control volume must be balanced by the amount of fluid flowing into (or out of) the same control volume, in addition to any fluid source or sink within the same control volume. We can thus write conservation of mass for V as

$$\frac{\partial}{\partial t}\int_V \eta(x,t)\mathrm{d}x + \int_{\partial V} \mathbf{q}\cdot\nu \mathrm{d}s = \int_V h(x,t)\mathrm{d}x. \tag{2.7}$$

Again, since the control volume V is arbitrary, and by the Divergence Theorem, we can write conservation of mass in differential form as

$$\frac{\partial \eta}{\partial t} + \nabla\cdot\mathbf{q} = h, \quad \text{for all} \quad x\in\Omega, t>0. \tag{2.8}$$

What remains now is some constitutive laws for the total poroelastic stress $\sigma(x,t)$, fluid content $\eta(x,t)$, and fluid mass flux $\mathbf{q}(x,t)$. These should be such that in the end, we have two equations where the only variables are the elastic displacement $\mathbf{u}$ and the fluid pressure p. These constitutive laws should also necessarily introduce a coupling between the momentum and mass conservation equations. Specifically, an increase in fluid pressure should induce added stress in the matrix, which again should result in a dilation of the porous structure, and vice versa; a compression of the medium should induce increased pore pressure (if the compression is fast enough relative to the fluid flow rate). Moreover, the driving force of the fluid mass flux is the pressure difference, i.e., the *pressure gradient*. We can therefore write the constitutive laws for the poroelastic stress, fluid content, and fluid mass flux tentatively as $\sigma = \sigma(p,\mathbf{u})$, $\eta = \eta(p,\mathbf{u})$ and $\mathbf{q} = \mathbf{q}(p)$, respectively.

The constitutive laws for the poroelastic stress and the fluid content were first formulated by K. Terzaghi [90] and M. A. Biot [11, 13], but we begin with the constitutive law for the fluid mass flux, which is the famous Darcy's law (somewhat older than the works of Terzaghi and Biot), given by

$$\mathbf{q}(p) = -\frac{1}{\mu_f}\mathbf{K}\nabla p, \tag{2.9}$$

where μ_f is the fluid viscosity and $\mathbf{K}$ is the permeability tensor of the solid matrix. Darcy's law essentially states that fluid will flow from regions of low pressure to regions of high pressure, i.e., the fluid is diffusing through the porous medium. The heterogeneity of the porous medium is thus encoded in the permeability matrix, which in the case of a perfectly isotropic porous medium reduces to a scalar. Introducing the standard Cauchy stress tensor from linear elasticity $\tilde{\sigma}(\mathbf{u})$, the constitutive law for the total poroelastic stress tensor is

$$\sigma(p,\mathbf{u}) = \tilde{\sigma}(\mathbf{u}) - \alpha p\mathbf{I}, \tag{2.10}$$

where $\mathbf{I}$ is the $d\times d$ identity tensor, and $\alpha > 0$ is the *Biot-Willis constant*, which accounts for the coupling between the fluid pressure and the solid matrix deformation. The Cauchy stress is usually given by Hooke's law, which in the case of an isotropic solid material takes the form $\tilde{\sigma}(\mathbf{u}) = 2\mu\mathbf{e}(\mathbf{u}) + \lambda\nabla\cdot\mathbf{u}\mathbf{I}$, where $\mathbf{e}(\cdot) = (\nabla(\cdot) + \nabla(\cdot)^\top)/2$ is the symmetric

gradient, and μ, λ are the material specific constants known as the Lamé parameters. Finally, the constitutive law for the fluid content is

$$\eta(p, \mathbf{u}) = c_0 p + \alpha \nabla \cdot \mathbf{u}, \tag{2.11}$$

where the constant $c_0 \geq 0$ represents the combined compressibility of the fluid and porosity of the solid matrix. An notable fact about the constitutive laws for the total stress and fluid content is the appearance of the coefficient α in both of these. This makes the coupling between the momentum and mass conservation equations into a symmetric one, and the full quasi-static Biot consolidation model (with suitable initial and boundary data) into a *saddle point problem* [17]. These coupling terms should be interpreted in the following way: The term $\alpha p \mathbf{I}$ results from the additional stress of the fluid pressure within the structure, and $\alpha \nabla \cdot \mathbf{u}$ represents the additional fluid content due to the local volume change.

Before writing out the full Biot system, we mention that the inertia term in the momentum equation (2.1) (i.e., the acceleration term) is usually ignored on the basis of a scaling argument, i.e., that the consolidation of the medium is happening slowly enough that the system remains in internal equilibrium throughout. Thus, the characteristic timescale in (2.1) can be considered as very large, and the acceleration term therefore becomes negligible. This results in Biot's quasi-static consolidation model taking the form of the following mixed elliptic-parabolic system of equations

$$-(\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}(x,t)) - \mu \Delta \mathbf{u}(x,t) + \alpha \nabla p(x,t) = \mathbf{f}(x,t), \tag{2.12a}$$

$$\partial_t (c_0 p(x,t) + \alpha \nabla \cdot \mathbf{u}(x,t)) - \nabla \cdot \frac{1}{\mu_f} \mathbf{K} \nabla p(x,t) = h(x,t), \tag{2.12b}$$

for all $x \in \Omega, t > 0$. The above system was first analyzed by R. Showalter [88], where the regularity of the solutions was shown to satisfy

$$\|p(t)\|_{H^2(\Omega)} + \|\mathbf{u}(t)\|_{(H^2(\Omega))^d} \leq \frac{C}{t}, \tag{2.13}$$

for some generic constant $C > 0$. Mixed formulations of the quasi-static Biot system has also been analyzed (taking the Darcy flux $\mathbf{q}$ and/or the total poroelastic stress σ as additional variables), and can be found in e.g. [2, 80, 97]. The quasi-static Biot system can also be derived using upscaling techniques. This is done by considering a 'resolved' porous medium with (Navier-)Stokes flow in the pore space and linear elasticity in the solid grains, coupled at the mutual interface by demanding no-slip and balance of normal forces, and then letting the size of the pores tend to zero (in some appropriate sense), thus producing the above system of equations for the upscaled (homogenized) medium. See e.g. [23, 34, 47, 66, 86] for more details.

Finally, we mention also the existence of the model known as the *Biot-Allard model* [68]. This extends the above quasi-static model to the dynamic situation, i.e., acceleration terms are retained, in addition to introduction of memory effects which are represented by integral terms.

2.1.2 Thermo-poroelasticity

Biot's model for quasi-static consolidation discussed in the previous section is formally equivalent to the classical thermoelasticity system, which describes heat flow through a linearly elastic solid. In this case, the variable p is the temperature distribution of the medium, the constant c_0 is the specific heat capacity, and the tensor $\mathbf{K}/\mu_f$ is the thermal conductivity. The term $\alpha p\mathbf{I}$ is then interpreted as the thermal stress induced by the temperature gradient, and $\alpha\nabla\cdot\mathbf{u}$ as the internal heating due to the dilation rate. The constant α is therefore in this context the thermal stress coefficient. Furthermore, the diffusive conservation equation no longer gives conservation of mass, but rather conservation of energy.

Linear thermoelasticity is also readily extended to the case of a porous medium, just as Biot's consolidation model can be seen as an extension of linear elasticity to a porous medium. This will then necessarily lead to a coupled system three equations, since the temperature is introduced as an additional third variable. For this three-field system to be fully coupled, the constitutive laws for the total stress and fluid content should now depend also on the temperature distribution of the medium, denoted by $T(x,t)$, i.e., $\sigma = \sigma(p,\mathbf{u},T)$ and $\eta = \eta(p,\mathbf{u},T)$. Moreover, in accordance with linear thermoelasticity, and in light of the discussion above, the energy conservation equation should be similar to the mass conservation equation.

Introducing the constant $b_0 \geq 0$ which is the thermal dilation coefficient, and $\beta > 0$ which is the thermal stress coefficient, the modified constitutive laws for the total stress and fluid content reads as

$$\sigma(p,\mathbf{u},T) := \tilde{\sigma}(\mathbf{u}) - \alpha p\mathbf{I} - \beta T\mathbf{I}, \tag{2.14a}$$

$$\text{and} \quad \eta(p,\mathbf{u},T) := c_0 p - b_0 T + \alpha\nabla\cdot\mathbf{u}, \tag{2.14b}$$

respectively (see e.g. [21, 36, 59, 93]). However, the derivation of the energy equation for the thermo-poroelastic system is not so easily done using conservation principles, as was the case in the previous section for the isothermal system. We will instead argue in the following way: If we expect the temperature couplings in the above constitutive laws to have symmetric counterparts in the energy equation, then this should contain the terms $-b_0\partial_t p$ and $\beta\nabla\cdot\partial_t\mathbf{u}$. Thus, if we denote by $a_0 > 0$ the combined thermal capacity of the fluid and solid skeleton, and by $\mathbf{y}(x,t)$ the thermal flux, we can write tentatively

$$\frac{\partial}{\partial t}\int_V \psi(x,t)\mathrm{d}x + \int_{\partial V}\mathbf{y}\cdot\nu\mathrm{d}s = \int_V z(x,t)\mathrm{d}x, \tag{2.15}$$

for some volumetric heat source density $z(x,t)$, and where we defined the *heat content* by $\psi(p,\mathbf{u},T) := a_0 T - b_0 p + \beta\nabla\cdot\mathbf{u}$. Assuming that the heat is diffusing through the porous medium according to Fourier's law of heat conduction, we have the following constitutive law for the heat flux

$$\mathbf{y}(x,t) = -\Theta\nabla T, \tag{2.16}$$

where Θ is the effective thermal conductivity. However, heat is not only transported through the porous medium by diffusion. It is also transported by the movement of the

flow, i.e. by convection. Thus, writing the above tentative energy equation in differential form, and also including an additional term which represents thermal convection yields

$$\begin{aligned}&\partial_t(a_0T(x,t) - b_0p(x,t) + \beta\nabla\cdot\mathbf{u}(x,t))\\&\quad+\frac{c_f}{\mu_f}\mathbf{K}\nabla p(x,t)\cdot\nabla T(x,t) - \nabla\cdot\Theta\nabla T(x,t) = z(x,t),\end{aligned}\tag{2.17a}$$

for all $x \in \Omega, t > 0$, and where $c_f > 0$ is the volumetric specific heat capacity of the fluid. A derivation of this equation using upscaling techniques can be found in e.g. [21, 59, 93]. The momentum and mass conservation equations (2.12a)-(2.12b) corresponding to the constitutive laws (2.14a)-(2.14b) are then given by

$$-(\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}(x,t)) - \mu\Delta\mathbf{u}(x,t) + \alpha\nabla p(x,t) + \beta\nabla T(x,t) = \mathbf{f}(x,t),\tag{2.17b}$$

$$\partial_t(c_0p(x,t) - b_0T(x,t) + \alpha\nabla\cdot\mathbf{u}(x,t)) - \nabla\cdot\frac{1}{\mu_f}\mathbf{K}\nabla p(x,t) = h(x,t),\tag{2.17c}$$

for all $x \in \Omega, t > 0$. Thus, the full quasi-static thermo-poroelastic system is given by equations (2.17a)-(2.17c). This system was analyzed in [22], where existence and uniqueness of a weak solution was established in the fully mixed formulation. It is also possible to consider a temperature dependent density in the above thermo-poroelastic system. This was done in [93], where the previously mentioned *Boussinesq approximation* was employed. In this case, the resulting model exhibits an additional (linear) temperature coupling in the mass conservation equation. Finally, we mention that the above thermo-poroelastic model is only valid for small temperature variations. This is due to the linearizations employed in obtaining the constitutive laws (2.14a)-(2.14b). Systems involving high temperature differences and/or phase transitions necessitates a more advanced model than the one described above.

2.2 Phase fields

The mathematical theory of *phase fields* is developed to be able to model and simulate the development of some *microstructure*, without the need for explicitly tracking the spatial and temporal evolution of individual interfaces between the microstructure and the surrounding medium. The term 'microstructure' used here should be interpreted in a broad sense, i.e., as the spatial arrangement of the defects and/or the spatial arrangement of the phases that have a different structural and/or compositional character than the surrounding matrix. As an example, consider some precipitate which is growing (or shrinking). The precipitate is then separated from the surrounding matrix by an interface which is evolving in space and in time. Thus, the mathematical modeling of such a process involves keeping track of three distinct entities; the *precipitate*, the *matrix*, and the *interface*. The classical approach in order to deal with this type of problem is to consider the precipitate and the matrix as governed by distinct sets of equations, but coupled at the mutual interface through physical boundary conditions (e.g., balance of

normal forces, continuity of fluxes, etc.). This interface is then an evolving surface, the driving force of which is the interaction between the precipitate and the matrix. Within this description, the interface is a $(d-1)$-dimensional surface within a d-dimensional domain, and is thus characterized as a *sharp interface*. This type of modeling is very challenging, especially if the evolution path of the sharp interface is not known *a priori*, which generally is the case. The theory of phase field offers a way to simplify such interfacial evolution problems, while still allowing the original sharp interface problem to be approximated in an asymptotical fashion (at the cost of increasing computational difficulty).

The *phase field* (usually denoted by the Greek letter φ) is an *order parameter* which is introduced into the relevant mathematical model as an additional variable (technically, it is the set of values of the order parameter over the entire domain that is the *phase field*, but the term is also commonly used to refer to the function itself). The boundary conditions at the (problematic) interfaces can then be substituted with an evolution equation for the phase field variable. Thus, the state of the microstructure as a whole is represented continuously by a single variable. However, this comes at the cost of the lower dimensional sharp interface being 'smeared out' in space by the phase field function, i.e., the phase field is designed to take the value zero at the interface, one in the matrix, and varying smoothly between zero and one in a transition zone of some prescribed non-zero (half-)thickness $\varepsilon > 0$ around the interface. The phase field is therefore continuous across the interfacial regions, which is in contrast to the sharp interface description, where the interfaces act as surfaces of discontinuity. The thickness of the artificial transition zone around the interface thus becomes an important model parameter in the new phase field evolution problem; it determines the degree to which the original sharp interface problem is approximated (formally, the original problem is recovered in the limit $\varepsilon \to 0$). Although phase field models in theory are valid for arbitrarily small ε, the computational time t scales with interface thickness as $t/t_0 \sim (\varepsilon/\varepsilon_0)^{-d}$ [82], thus forming a computational bottleneck in regards to the degree to which the original sharp interface problem can be approximated. The figure 2.2 below shows a schematic illustration of a snapshot in time of some evolving microstructure, which is represented by a phase field function.

Early developments of phase field theory was in relation to the solidification dynamics of pure and binary materials [35, 41, 58]. Further developments investigated the effects of anisotropy [24, 25], and the convergence of phase field models to the 'sharp interface'-limit [26]. Furthermore, in [78, 79] there is proposed a general framework for deriving phase field models in a thermodynamically consistent way, thus providing some universality between different phase field models. In many applications it suffices to have only one phase field variable, e.g., when only two distinct phases are present (i.e., a *binary model*), or when tracking the evolution of some structural defects in a 'pure' material. However, it is also possible to consider any finite number of distinct phase field variables for more complicated multi-phase situations [31, 54, 74]. Phase field theory continues to be an active area of research, and is highly relevant for applications in e.g., solidification dynamics, brittle fracture propagation and viscous fingering, to name

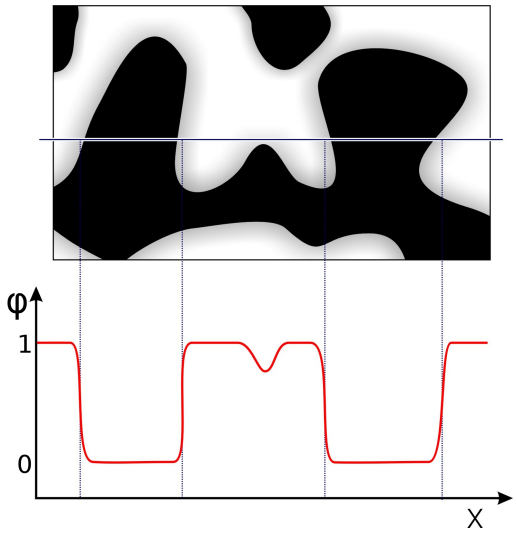


Figure 2.2: Profile of phase field function φ across the horizontal line in the domain. The sharp interfaces of the microstructure is replaced by a diffusive transition zone which is represented by $\{x \in \Omega : 0 < \varphi(x,t) < 1\}$. Source: `wikipedia.org/wiki/Phase_field_models`.

a few. For more details regarding different applications of phase fields, see e.g. [10, 82]

There is a wide variety of phase field models, but common to all is the diffuse interface description of some originally sharp interface problem. The focus here will be on phase field descriptions of brittle fracture propagation in a quasi-static elastic material. But first, another aspect of the general theory of phase fields is addressed, namely the free energy functional.

2.2.1 Energy functional

When formulating a phase field model, a key ingredient is to define a free energy functional of the system in terms of the phase field variable (and its derivatives). Common choices are the *Gibbs free energy* (for an isothermal system at constant pressure) or the *Helmhotz free energy* (for a system at constant temperature and volume). Note that for an isolated system which is not isothermal, an *entropy functional* may turn out to be the most appropriate choice. The necessary theory in this regard was first developed by J. W. Cahn and J. E. Hilliard, by realizing that the free energy of an arbitrary control volume within a heterogenous system cannot depend only on the phase composition within the control volume, but also on the phase composition of the surrounding environment [27, 28]. This is due to the fact that control volumes with equal volume fractions of phase compositions need not be energetically equivalent. Thus, the total free energy of a heterogenous system depends not only on the phase field variable, but also on its derivatives.

In what follows, a simplified derivation of a free energy functional is presented: If $f_0(\varphi)$ is the free energy per unit volume of a homogenous system, then the heterogenous energy density $f(\varphi, \nabla\varphi, \nabla^2\varphi, \cdots)$ can be approximated by performing the following Taylor expansion

$$\begin{aligned} f = f_0 &+ \frac{\partial f_0}{\partial \nabla\varphi}\nabla\varphi + \frac{1}{2}\frac{\partial^2 f_0}{\partial(\nabla\varphi)^2}(\nabla\varphi)^2 + \cdots \\ &+ \frac{\partial f_0}{\partial \nabla^2\varphi}\nabla^2\varphi + \frac{1}{2}\frac{\partial^2 f_0}{\partial(\nabla^2\varphi)^2}(\nabla^2\varphi)^2 + \cdots \\ &\vdots \qquad\qquad \ddots \end{aligned} \tag{2.18}$$

Since the heterogenous free energy density must be invariant to a change of sign in the coordinates, the coefficients of the terms involving odd orders of differentiation of the homogenous energy density must be equal to zero. Moreover, further simplification of the above expression is achieved using the following formula, which is obtained by integration by parts

$$\int_V \frac{\partial f_0}{\partial \nabla^2\varphi}\mathrm{d}x = \frac{\partial f_0}{\partial \nabla^2\varphi}\nabla\varphi \cdot \nu - \int_V \frac{\partial}{\partial\varphi}\frac{\partial f_0}{\partial \nabla^2\varphi}(\nabla\varphi)^2\mathrm{d}x, \tag{2.19}$$

where V is some arbitrary control volume, and ν its outward unit normal field. Since the first term on the right hand side of (2.19) involves an odd order of differentiation of f_0, it must be equal to zero by the same argument as above. Since the control volume V was arbitrary, the Taylor expansion of the free energy density (2.18) thus reduces to

$$f = f_0(\varphi) + \frac{1}{2}\left(\frac{\partial^2 f_0}{\partial(\nabla\varphi)^2} - 2\frac{\partial}{\partial\varphi}\frac{\partial f_0}{\partial \nabla^2\varphi}\right)(\nabla\varphi)^2 + \cdots . \tag{2.20}$$

Truncating this expansion after the first and second order terms yields the free energy functional for an arbitrary control volume V within a heterogenous system as

$$E(\varphi) = \int_V f_0(\varphi) + \frac{\lambda^2}{2}(\nabla\varphi)^2\mathrm{d}x, \tag{2.21}$$

where

$$\lambda^2 = \frac{\partial^2 f_0}{\partial(\nabla\varphi)^2} - 2\frac{\partial}{\partial\varphi}\frac{\partial f_0}{\partial \nabla^2\varphi} \tag{2.22}$$

is the *gradient energy coefficient*. In order to have an accurate description of interface properties such as e.g., *anisotropy* and *surface energy density*, an accurate value of this coefficient is needed. In a more general setting than in the above simplified derivation, e.g., if phase transformations occur in some elastic material which then induces displacements, such that this mechanical energy must be accounted for in the energy functional, then, the free energy functional must include a contribution from the induced mechanical elastic energy, which yields a dependence on the displacement vector $\mathbf{u}$, i.e., $E = E(\varphi, \mathbf{u})$.

2.2.2 Brittle fracture propagation

In the context of brittle fracture propagation, the backdrop is some brittle elastic material to which a loading force is applied. If enough loading force is exerted, the brittle material cracks, creating sharp surfaces of discontinuity in the displacement field. However, in a phase field formulation, the phase field variable acts as an indicator function for these surfaces of discontinuity, and creates a diffusive transition zone between the sharp fracture surface and the elastic material. The figure 2.3 below shows a plot of the phase field function for a brittle fracture propagation simulation. The following presentation is based on [16, 42, 65].

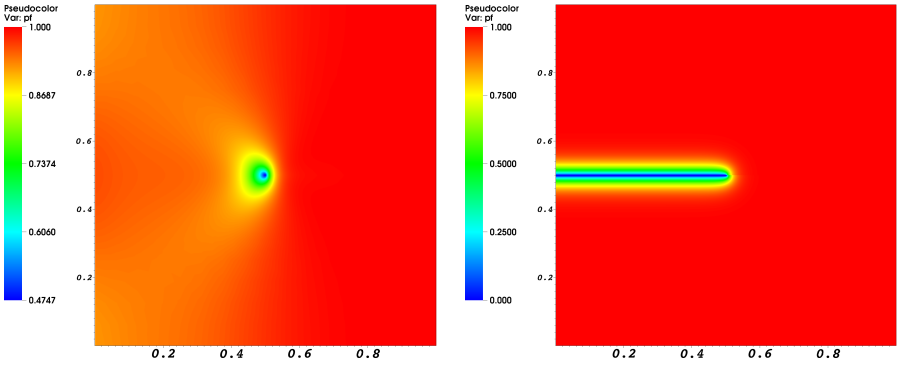


Figure 2.3: Computation of phase field brittle fracture propagation [53]. Left: The crack is beginning to evolve from the middle of the domain. Right: The crack has reached the boundary of the domain. Diffusive zone around the fracture created by the phase field is visible on both figures.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded, open domain, containing some brittle elastic material, wherein $\mathscr{C} \subset \mathbb{R}^{d-1}$ denotes the lower dimensional fracture surface(s), and where the domain boundary $\partial\Omega$ is decomposed into two parts, Γ_D and Γ_N, both of strictly positive measure. If $\mathbf{u}(x, t)$ is the displacement vector, and if $\mathbf{f}(x, t)$ is a force applied on Γ_N, and assuming that fracture surfaces are not reaching the domain boundary $\partial\Omega$, we define the following total energy functional according to Griffith's criterion for brittle fracture [48]

$$E(\mathbf{u}, \mathscr{C}) = \frac{1}{2}(\mathbb{C}\mathbf{e}(\mathbf{u}), \mathbf{e}(\mathbf{u}))_{\Omega\setminus\mathscr{C}} - (\mathbf{f}, \mathbf{u})_{\Gamma_N} + G_c\mathscr{H}^{d-1}(\mathscr{C}), \tag{2.23}$$

where $\mathbb{C} = [C_{ijkl}]_{ijkl}$ is the fourth order tensor containing the elastic material coefficients, and $\mathscr{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff-measure. Furthermore, the constant $G_c > 0$ is the *critical energy restitution rate*, giving the critical value for the elastic energy restitution rate at which fracture propagation occurs. In the above functional, the

first term describes the bulk energy in the intact domain, the second term is the traction boundary forces, and the last term is the surface fracture energy.

We now introduce the phase field function $\varphi(x,t)$, which takes the value 0 in the fracture, 1 in the intact domain, and varies smoothly from 0 to 1 in a transition zone of (half-)thickness $0 < \varepsilon < 1$ around the fracture. Using the phase field function, we can regularize the sharp fracture surface, i.e., we replace the Hausdorff-measure of the crack surface $\mathscr{H}^{d-1}(\mathscr{C})$ by the following regularized fracture functional

$$\Gamma_\varepsilon(\varphi) = \frac{1}{2\varepsilon}\|1-\varphi\|^2 + \frac{\varepsilon}{2}\|\nabla\varphi\|^2. \tag{2.24}$$

Furthermore, we would like to define the elastic energy (and thus the displacement) on Ω rather than on $\Omega\setminus\{x\in\Omega : \varphi(x,t)=0\}$. To this end, we introduce another regularization parameter $0 < \kappa < \varepsilon$, and define the so-called *degradation function* as

$$g(\varphi) = (1-\kappa)\varphi^2 + \kappa. \tag{2.25}$$

With these definitions, we replace the total energy functional (2.23) with the regularized total energy (denoted as E_ε), which takes the following form

$$E_\varepsilon(\mathbf{u},\varphi) = \frac{1}{2}(g(\varphi)\mathbb{C}\mathbf{e}(\mathbf{u}),\mathbf{e}(\mathbf{u})) - (\mathbf{f},\mathbf{u}) + G_c\Gamma_\varepsilon(\varphi). \tag{2.26}$$

With this formulation, we have essentially replaced the fracture by a softer material. Finally, a crack irreversibility condition must be enforced (the crack is not allowed to heal), which takes the form

$$\partial_t\varphi \le 0. \tag{2.27}$$

The problem is then to find the displacement vector and phase field function $\{\mathbf{u}(t),\varphi(t)\}$ that minimizes the regularized energy functional (2.26), while also subject to the irreversibility constraint (2.27). The standard approach when dealing with this type of minimization problem is to restate it in terms of the corresponding Euler-Lagrange equations. Formally speaking, this involves differentiating the expression (2.26) with respect to the arguments, and setting the resulting expressions equal to zero. To this end, we begin by introducing the space of admissible displacements and phase field functions as $V := \{\mathbf{v}\in(H^1(\Omega))^d : \mathbf{v}|_{\Gamma_D} = 0\}$ and $W := H^1(\Omega)\cap L^\infty(\Omega)$, respectively. The Euler-Lagrange equations are then obtained by evaluating the following limits

$$0 = \lim_{s\to 0}\frac{1}{s}(E_\varepsilon(\mathbf{u}+s\mathbf{v},\varphi) - E_\varepsilon(\mathbf{u},\varphi)), \qquad \forall\mathbf{v}\in V, \tag{2.28a}$$

$$0 = \lim_{s\to 0}\frac{1}{s}(E_\varepsilon(\mathbf{u},\varphi+s\psi) - E_\varepsilon(\mathbf{u},\varphi)), \qquad \forall\psi\in W. \tag{2.28b}$$

Incorporating also the irreversibility constraint (2.27), we obtain the following variational inequality problem: Find $(u(t),\varphi(t))\in V\times W$ such that for all $t>0$ there holds

$$(g(\varphi)\mathbb{C}\mathbf{e}(\mathbf{u}),\mathbf{e}(\mathbf{v})) = (\mathbf{f},\mathbf{v}),\ \ \forall\mathbf{v}\in V, \tag{2.29a}$$

$$G_c\varepsilon(\nabla\varphi,\nabla\psi) - \frac{G_c}{\varepsilon}(1-\varphi,\psi) + (1-\kappa)(\varphi|\mathbb{C}\mathbf{e}(\mathbf{u})|^2,\psi) \ge 0. \quad \forall\psi\in W. \tag{2.29b}$$

The time-discrete version of the above system was analyzed in [75], and there it was shown that at least one global minimizer exists, provided sufficient regularity of domain and loading force. We mention also that the analysis of a pressurized phase field brittle fracture model can be found in [72, 73].

2.2.3 Limitations of phase fields

The theory of phase fields is particularly well suited for modeling and simulating the development of microstructure, as introduction of the phase field variable(s) effectively removes the biggest difficulties associated with the classical description of such microstructural (i.e. sharp interface) evolution problems. Many of the advantages of the phase field description is discussed in the introduction of this chapter, but there are also some important limitations associated with phase fields which are not easy to overcome. Specifically, due to the approximating nature of phase fields, some limitations are in fact 'built-in' to the theory, and thus *cannot* be overcome.

The most obvious disadvantages of the theory relates to the thickness of the artificial diffusive zone around the sharp interfaces, i.e., to the universal model parameter ε. This parameter may easily be set to unrealistic values. This may result in an unacceptable degree of loss of detail, and appearance of unphysical phenomena, even if this is not clear from the modeler's point of view. For any given problem, to determine exactly for what values of ε the model becomes unphysical is very difficult, and may even be impossible. On the other hand, if ε is chosen too small, computational issues become apparent, and a lot of the difficulties associated with the original sharp interface model resurface. This is especially true for large domain simulations. Consider e.g., a brittle fracture evolution problem. In reality the fractures within the elastic body have an aperture, but this is small enough that fractures are usually considered as lower dimensional surfaces. Setting ε equal to the average fracture aperture should then produce a realistic model, but for practical computations, such small values of ε necessitate an extremely fine computational grid. Thus, the situation might be that an unphysical value of ε is demanded for a feasible practical computation. The simulation will nevertheless produce a smooth looking phase field function, but if the result is quantitatively comparable to reality is another question entirely. However, phase fields is still a powerful theory which continues to receive a great amount of attention. Some of the limitations discussed in the above may also be mitigated by development of more advanced simulation technology.

Chapter 3

Homogenization

3.1 Introduction

Homogenization is an *asymptotic analysis theory* within the broader field of applied mathematics, which has its origins in the field of *material engineering*, specifically from subfields dealing with *composite materials*. In particular, homogenization theory originates from the attempt to understand how the constitutive equations of a composite material can be derived from the constitutive equations of the components constituting the composite, and from the components' topological and geometric distributions within the composite. In other words, the aim of homogenization theory is to establish the macroscopic behavior of a system which is microscopically heterogenous, by 'smoothing' the microscopic heterogeneities. The development of this theory is due, at least in part, to the realization that in many applications dealing with complex heterogenous media, the relevant properties of the medium may only be the *effective* ones, i.e., the average properties which does not depend explicitly on the microscopic heterogeneities, and which only emerge at a scale much larger than said heterogeneities. Homogenization theory provides the mathematical framework necessary to bridge the gap between the (possibly prohibitively complex) microscopic description, and the (less complex) effective macroscopic description. Hence, given a micro-scale model of some heterogenous material, homogenization can reveal the effective characteristics of this material by replacing the original complex heterogenous material by a new homogenous (fictitious) one, possessing only effective properties.

They key idea in homogenization is that a heterogenous body, as long as the heterogeneities are sufficiently small and sufficiently evenly distributed, appears as homogenous from the macroscopic point of view. Thus, regarded as a homogenous body from a sufficiently 'zoomed out view', macroscopic properties of the (microscopically) heterogenous body emerge, which does not exist at the micro-scale (such as e.g., *permeability*, *porosity*, etc.). The successful characterization of these macroscopic properties by way of the microscopic properties is the remarkable utility of homogenization theory, and the reason it continues to be an active field of research.

Although beginning as a subfield within material engineering, homogenization theory has steadily become more grounded in mathematics, and is now viewed purely as a mathematical theory. Due to the works of Nguetseng [76, 77] and Allaire [3] there is now even rigorous notions of weak compactness specifically tailored for dealing with problems coming from homogenization. Thus, one can obtain the macroscopic model as a limiting case when letting the size of the microscopic heterogeneities tend to zero. However, this (rigorous) theory is beyond the scope of this dissertation; the focus here will be on the formal 'two-scale asymptotic expansion' method.

For a comprehensive overview of homogenization, see the textbooks [33, 49]. For an overview of the rigorous theory, see e.g. [43].

3.1.1 Applicability of homogenization

Application of homogenization theory is most effective for problems involving heterogenous media with some uniformity of the heterogeneities, or more generally, for boundary value problems involving coefficients with *low-amplitude, high-frequency oscillations*. For some heterogenous body occupying a domain Ω, we can summarize the main requirements for homogenization to be applicable in the following bullet points:

- That the heterogeneities are small compared to the size of Ω.
- That the heterogeneities are uniform in size.
- That the heterogeneities are evenly distributed throughout Ω.

Thus, two scales characterize the domain Ω: The scale of the heterogeneities, and the global scale of Ω itself. As long as these two scales are separated by sufficient orders of magnitude (usually $1e6 - 1e12$), they can be considered as almost independent. Additionally, if the microscopic heterogeneities within Ω are sufficiently uniform in size and sufficiently evenly distributed, they may be considered as being periodic, even if this is strictly not the case in reality. As a rule of thumb, the greater the scale separation is, and the greater the uniformity of the heterogeneities, the more reliable the results produced by homogenization. The use of the word 'sufficiently' in the above must therefore be interpreted within the specific context.

The heterogenous body occupying Ω may be a fine mix of several constituents, or it may be a single material with many small perforations. Hence, e.g., *porous structures* are excellent candidates for application of homogenization: In a large-scale porous structure, say, a subsurface reservoir, the two characteristic scales would be the size of an average pore and the size of the porous reservoir itself, leading to an exceptionally great scale separation. As there is also usually some uniformity in the grains making up the porous skeleton, we can consider the three bullet points above to be fulfilled.

Finally, we mention some of the literature on applications of homogenization in the context of porous media: In [23] a formal upscaling leading to the quasi-static Biot model was undertaken, and in the book [86] a rigorous upscaling can be found. In [34, 47] the rigorous derivation of dynamic Biot-type models can be found, corresponding to

different scalings of the micro-scale model. In [40] the case of an inviscid fluid filling the pore space is treated.

3.1.2 The two-scale asymptotic expansion method

As mentioned earlier, the focus here will be on the homogenization technique known as the *two-scale asymptotic expansion method* within the *periodic framework* [33, 49]. The periodicity assumption is very common in applications of the two-scale asymptotic expansion method, and in homogenization in general. Although this assumption significantly simplifies the technical procedure, it is not a necessary requirement. In particular, at a certain point during the derivation, one partitions the original heterogenous domain into small cells (or REVs in the context of porous media), which from the macro-scale point of view can be regarded as *infinitesimal points*. The periodic assumption is then essentially the same as assuming the heterogeneities behind each such (macro-scale) point is always arranged in an identical fashion. Therefore, only one local description of the micro-scale heterogeneities, known as a 'reference cell', is necessary in order to complete the homogenization procedure. Without the periodic assumption there is also needed information about how the arrangement of the micro-scale heterogeneities change from (macro-scale) point to (macro-scale) point, i.e., a parametrization of the reference cell is needed. In the upscaled model, however, this is only the difference between constant or spatially dependent effective coefficients. Throughout this chapter we will always assume to be within the periodic framework.

More important than the absolute value of the two scales characterizing the relevant model is their ratio, i.e., if $l > 0$ is the characteristic scale of the microscopic heterogeneities and $L > 0$ is the characteristic scale of the macroscopic domain, the dimensionless parameter $\varepsilon := l/L << 1$, known as the *scale separation parameter* is defined. After a non-dimensionalization of the relevant model (i.e. the macro-scale domain is now of $\mathcal{O}(1)$ while the microscopic heterogeneities are of $\mathcal{O}(\varepsilon)$), the following two-scale asymptotic expansion of the relevant variable(s) is postulated, i.e. if $u : \Omega \to \mathbb{R}$ is the quantity in question, then

$$u(x) = u^0(x, x/\varepsilon) + \varepsilon u^1(x, x/\varepsilon) + \varepsilon^2 u^2(x, x/\varepsilon) + \cdots, \tag{3.1}$$

for some functions u^j, $j = 1, 2, 3, ...$, which are periodic in the second argument. The fact that the u^j depends on x and x/ε needs to be understood in the following sense: Since x is the dimensionless position of a point within the (dimensionless) domain Ω, we have that the needed variations in x when describing macroscopic changes is of $\mathcal{O}(1)$. Similarly, the needed variations in x when describing microscopic changes is of $\mathcal{O}(\varepsilon)$, moreover, these changes are periodic with period ε. The 'blown-up' variable x/ε can therefore be viewed as giving the position within one periodic reference cell, rescaled such that the period is 1. The first term on the right hand side of (3.1), i.e. u^0, thus represents the 'slow' macroscopic changes of $\mathcal{O}(1)$ in u, the next term, u^1, represents the 'faster' changes in u coming from the microscopic heterogeneities of $\mathcal{O}(\varepsilon)$, and similarly for the higher order terms. The figure 3.1.2 below illustrates the idea behind the postulated two-scale

expansions.

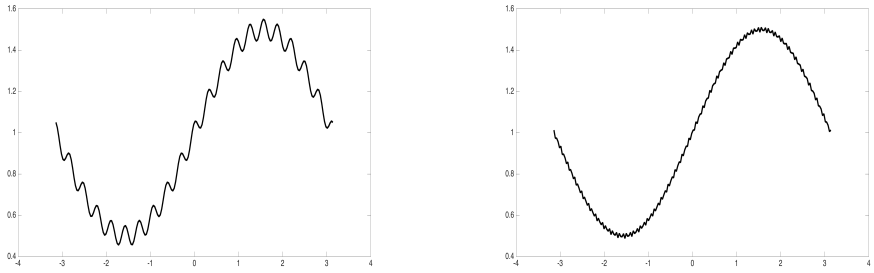


Figure 3.1: Graph of $f^\varepsilon(x) = \sin(x)/2 + 1 + \varepsilon\cos(x/\varepsilon)$ for $\varepsilon = 1/20$ (left) and $\varepsilon = 1/80$ (right). These figures illustrate the idea behind the two-scale expansion: The function f^ε has 'slow' changes of $\mathcal{O}(1)$, and 'fast' changes of $\mathcal{O}(\varepsilon)$, and therefore admits a two-scale expansion (which is also evident from its formula). Example taken from [43].

3.2 A classic homogenization example

The details of the two-scale asymptotic expansion method is best outlined with an example. Hence, this section will be devoted to going through a model homogenization problem is some detail. The following presentation is based on [33].

3.2.1 Problem description

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be an open and bounded set occupied by a heterogenous body consisting of a fine mix of two constituents, which we hereby name constituent A and constituent B. The figure 3.2.1 below shows a schematic of the situation.

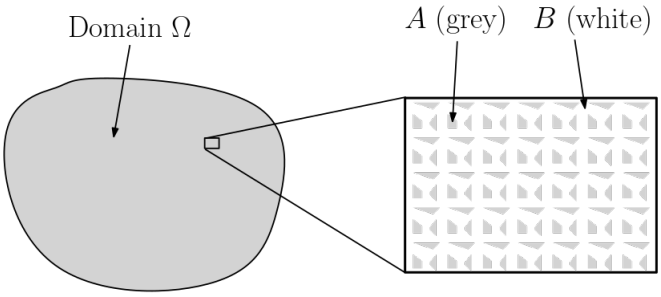


Figure 3.2: Schematic of the composite material. The domain Ω appears homogenous from the macroscopic point of view, but when zooming in on a small area, the microstructure becomes visible, i.e., the material is *microscopically heterogenous*. In this figure, constituent A is colored grey, while constituent B is the white part in between.

We denote by Ω_A and Ω_B the open subdomains of Ω containing all of constituent A and B, respectively, satisfying $\Omega_A \cap \Omega_B = \emptyset$, and such that these subdomains and their mutual interface make up the larger domain, i.e., $\Omega = \Omega_A \cup \Omega_B \cup (\partial\Omega_A \cap \partial\Omega_B)$. Furthermore, we assume that both subdomains are isotropic and have constant thermal conductivities γ_A and γ_B, respectively. Let now $u_{A,B} : \Omega_{A,B} \to \mathbb{R}$ and $\mathbf{q}_{A,B}(x) := \gamma_{A,B}\nabla u_{A,B}(x)$ be respectively the temperature distributions and thermal fluxes in the two subdomains. Using this, we define the global temperature distribution, thermal conductivity, and thermal flux by

$$u(x) := \begin{cases} u_A(x), & x \in \Omega_A, \\ u_B(x), & x \in \Omega_B, \end{cases} \tag{3.2}$$

and

$$\gamma(x) := \begin{cases} \gamma_A, & x \in \Omega_A, \\ \gamma_B, & x \in \Omega_B, \end{cases} \tag{3.3}$$

and

$$\mathbf{q}(x) := \begin{cases} \mathbf{q}_A(x), & x \in \Omega_A, \\ \mathbf{q}_B(x), & x \in \Omega_B, \end{cases} \tag{3.4}$$

respectively. From physical considerations we then have continuity of temperature and thermal flux at the internal interface between composite A and composite B, i.e.

$$u_A = u_B \qquad \text{on } \partial\Omega_A \cap \partial\Omega_B, \tag{3.5a}$$

$$\mathbf{q}_A \cdot \nu_A = \mathbf{q}_B \cdot \nu_B \qquad \text{on } \partial\Omega_A \cap \partial\Omega_B, \tag{3.5b}$$

where we denote the unit normal fields of subdomains Ω_A and Ω_B by ν_A and ν_B, respectively, satisfying $\nu_A = -\nu_B$. If $f : \Omega \to \mathbb{R}$ is a heat source acting on the domain Ω, and if zero temperature is prescribed at the outer boundary $\partial\Omega$, then the global temperature satisfies the following boundary value problem (assuming the system is equilibrium): Find the temperature distribution $u : \Omega \to \mathbb{R}$ such that

$$-\nabla \cdot (\gamma \nabla u) = f, \qquad \text{in } \Omega, \tag{3.6a}$$

$$u = 0, \qquad \text{on } \partial\Omega. \tag{3.6b}$$

Note that it follows from (3.5b) that the temperature gradient is a discontinuous function, which implies that the problem (3.6a)-(3.6b) should be interpreted in the sense of weak derivatives. However, for simplicity, we choose *not* to adapt the variational formulation of this problem in the following presentation.

Due to the fine mixture of the two constituents, the global thermal conductivity is now oscillating very rapidly between the two values, γ_A and γ_B. This makes the above model problem very difficult to treat, especially from the numerical point of view. If the mixture is fine enough, it even becomes impossible. The good news is, however, that we are not interested in modeling the temperature distribution on Ω in a resolution that resolves the microscopic heterogeneities in full detail. We will instead use homogenization to derive an upscaled version of the problem (3.6a)-(3.6b), posed on some new 'zoomed

out' homogenous (fictitious) domain. Thus, in this new system, the variable will not be the original temperature distribution u, defined on the original heterogenous domain, but instead some effective temperature distribution (usually denoted the same way) which is only valid on a scale much larger than the scale of the microscopic heterogeneities. In other words, the effective variable will only 'see' the new homogenous domain, and not the original heterogenous one. Furthermore, the effective (upscaled) thermal conductivity will no longer have rapid oscillations, but instead become a constant tensor. However, information about the microscopic heterogeneities will still be encoded into this effective tensor coefficient. Thus, the fundamental assumption we make when deriving the upscaled problem is that solving this should produce a temperature distribution which is *accurate enough*, but without the rapid oscillations coming from the microscopic heterogeneities.

3.2.2 Homogenization ansatz

Homogenization usually requires careful dimensional considerations of the (micro-scale) problem at hand, but we will here assume that this has been done, and simply denote the characteristic micro- and macro-scales by l, and L, respectively, and their ratio by $\varepsilon := l/L$. Thus, in this non-dimensional framework the macro-scale is of $\mathcal{O}(1)$, while the micro-scale is of $\mathcal{O}(\varepsilon)$. We now introduce the periodicity assumption: Assume we have a reference period Y, in which the reference heterogeneities are given, but rescaled such that $Y := [0,1]^d$. This means that the heterogeneities in Ω are now periodic with period εY. The figure 3.2.2 below illustrates the idea behind the reference period Y.

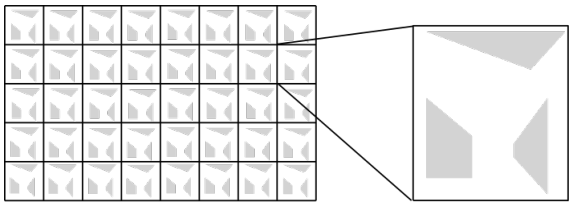


Figure 3.3: Periodicity of the micro-scale heterogeneities. Cells of periodicity indicated on the 'zoomed in' domain on the left. On the right is the reference period Y.

To now make the ε-dependence on the problem (3.6a)-(3.6b) explicit, we rewrite it as

$$-\nabla \cdot (\gamma^\varepsilon \nabla u^\varepsilon) = f, \qquad \text{in } \Omega^\varepsilon, \tag{3.7a}$$

$$u^\varepsilon = 0, \qquad \text{on } \partial\Omega^\varepsilon, \tag{3.7b}$$

where the ε-superscript on the variable and on the domain implies dependency on ε. Furthermore, the coefficient function γ^ε is now periodic with period εY, hence we can write $\gamma^\varepsilon(x) = \gamma(x/\varepsilon)$ for some Y-periodic function γ. We can thus think of Ω^ε and u^ε as being the 'resolved' domain and temperature distribution, respectively.

In order to facilitate the homogenization procedure, we introduce now the 'microscopic variable', y by

$$y := x/\varepsilon, \tag{3.8}$$

where x is here the 'macroscopic variable'. This is motivated by the fact that if $x \in \Omega^\epsilon$, then there exists $n \in \mathbb{Z}^d$ such that $x/\varepsilon = (n+y)$, where $y \in Y$. Hence, x gives the position of a point in the domain Ω, while y gives its position within the rescaled reference period Y. Note that the macroscopic variable x does not live in the 'resolved' domain Ω^ϵ, but rather in some new homogenized domain Ω, where we can think there is a reference cell εY behind each point.

We have now arrived at the first key juncture in the homogenization process; to invoke the *homogenization ansatz*, i.e., to postulate that the solution to the 'resolved' problem (3.7) admits the following two-scale asymptotic expansion

$$u^\epsilon(x) := u^0(x,y) + \varepsilon u^1(x,y) + \varepsilon^2 u^2(x,y) + \cdots, \tag{3.9}$$

where each u^j, $j = 1, 2, ...$, is defined for $x \in \Omega$ and $y \in Y$, and is Y-periodic with respect to y for each $x \in \Omega$. The aim now is to substitute the expansion (3.9) into the problem (3.7a)-(3.7b), and find some boundary value problem satisfied by u^0, where the dependency on micro-scale variable y has been eliminated. Due to (3.8), we must first reformulate the differential operators in (3.7a) according to the chain rule, i.e.,

$$\nabla = \nabla_x + \frac{1}{\varepsilon}\nabla_y, \tag{3.10}$$

and similarly for the divergence. Using this, together with (3.9) yields equation (3.7a) as

$$\begin{aligned} f =& -\varepsilon^{-2}\nabla_y \cdot (\gamma \nabla_y u^0) \\ &- \varepsilon^{-1}\left[\nabla_x \cdot (\gamma \nabla_y u^0) + \nabla_y \cdot (\gamma(\nabla_y u^1 + \nabla_x u^0))\right] \\ &- \varepsilon^{0}\left[\nabla_x \cdot (\gamma(\nabla_x u^0 + \nabla_y u^1) + \nabla_y \cdot (\gamma(\nabla_x u^1 + \nabla_y u^2))\right] + \mathcal{O}(\varepsilon). \end{aligned} \tag{3.11}$$

Note that we here discarded all terms of $\mathcal{O}(\varepsilon)$ and higher order. In general, the choice of where to truncate the expansion (3.9) depends on the specific problem at hand. Equating terms of equal ε-power in (3.11) yields the following set of boundary value problems (indexed by their respective ε-power):

$$\varepsilon^{-2}: \begin{cases} -\nabla_y \cdot (\gamma \nabla_y u^0) = 0, & \text{in } Y, \quad (3.12a) \\ u_0(x,\cdot) \text{ is } Y\text{-periodic for all } x \in \Omega, & (3.12b) \end{cases}$$

and

$$\varepsilon^{-1}: \begin{cases} -\nabla_x \cdot (\gamma \nabla_y u^0) - \nabla_y \cdot (\gamma(\nabla_y u^1 + \nabla_x u^0)) = 0, \text{ in } Y, & (3.13a) \\ u_1(x,\cdot) \text{ is } Y\text{-periodic for all } x \in \Omega, & (3.13b) \end{cases}$$

and

$$\varepsilon^{0}: \begin{cases} -\nabla_x \cdot (\gamma(\nabla_x u^0 + \nabla_y u^1) - \nabla_y \cdot (\gamma(\nabla_x u^1 + \nabla_y u^2)) = f, \text{ in } Y, & (3.14a) \\ u_2(x,\cdot) \text{ is } Y\text{-periodic for all } x \in \Omega. & (3.14b) \end{cases}$$

Note that the macroscopic variable x only acts as a parameter in the above boundary value problems.

3.2.3 The upscaled system

To derive the upscaled system, we must now successively solve the boundary value problems (3.12a)-(3.12b), (3.13a)-(3.13b), and (3.14a)-(3.14b). Firstly, observe that from (3.12a)-(3.12b) we have that u^0 must be independent of y, i.e.

$$u^0(x, y) = u^0(x). \tag{3.15}$$

Thus, we also get that the first term on the left hand side of (3.13a) vanishes, and the boundary value problem for ε^{-1} therefore becomes

$$\begin{cases} -\nabla_y \cdot (\gamma(\nabla_y u^1 + \nabla_x u^0)) = 0, \text{ in } Y, \\ u_1(x, \cdot) \quad \text{is } Y\text{-periodic for all } x \in \Omega. \end{cases} \tag{3.16}$$

In order to solve this, we postulate the following form for the solution u^1 using separation of variables:

$$u^1(x, y) = \sum_{j=1}^{d} \frac{\partial u^0}{\partial x_j}(x) U^j(y), \tag{3.17}$$

for some auxiliary functions $U^j(y)$, $j = 1, ..., d$, which are determined by solving the following set of auxiliary boundary value problems (for $j = 1, ..., d$)

$$\begin{cases} -\nabla_y \cdot (\gamma(\nabla_y U^j + \mathbf{e}_j)) = 0, \text{ in } Y, & \text{(3.18a)} \\ U^j \text{ is } Y\text{-periodic}, & \text{(3.18b)} \end{cases}$$

where $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d\}$ is the canonical orthonormal basis for $\mathbb{R}^d$. We continue with the boundary value problem (3.14a)-(3.14b), where we use both (3.15) and (3.17), and integrate over Y to obtain

$$-\sum_{j,k=1}^{d} \frac{\partial}{\partial x_j} \int_Y \gamma(y)(\nabla_y U^k(y) + \mathbf{e}_k) \mathrm{d}y \frac{\partial u^0(x)}{\partial x_k} = f(x)|Y| = f(x), \qquad \text{for } x \in \Omega. \tag{3.19}$$

Defining the second order tensor $\Gamma \in \mathbb{R}^{d \times d}$ componentwise as

$$[\Gamma]_{ij} := \int_Y \gamma(y)(\nabla_y U^j(y) + \mathbf{e}_j)_i \mathrm{d}y, \tag{3.20}$$

we can write (3.19) more compactly as

$$-\nabla_x \cdot (\Gamma \nabla_x u^0) = f, \qquad \text{for } x \in \Omega. \tag{3.21}$$

Thus, we have found an equation satisfied by u^0, and where the dependency on the microscale variable y has been eliminated. Furthermore, from (3.9) we see that u^0 must vanish on the boundary of Ω, and therefore the upscaled version of (3.7a)-(3.7b) is given by (omitting now subscripts on differential operators and superscripts on variables)

$$\begin{aligned} -\nabla \cdot (\Gamma \nabla u) &= f, \qquad \text{in } \Omega, & \text{(3.22a)} \\ u &= 0, \qquad \text{on } \partial\Omega. & \text{(3.22b)} \end{aligned}$$

Observe that the structure of this upscaled problem is the same as for the original problem (3.7a)-(3.7b) (i.e. an elliptic equation with a homogenous Dirichlet boundary condition), but now with a constant tensor valued coefficient Γ, instead of a rapidly oscillating scalar valued coefficient function γ. This means that the complexity in the upscaled problem is significantly reduced, and practical computations are now feasible. However, in order to determine the entries of Γ, one needs to solve the set of auxiliary boundary value problems (3.18a)-(3.18b), usually referred to as *cell problems* in the homogenization literature, on some given reference cell geometry (e.g., such as illustrated in figure 3.3).

3.2.4 Properties of the effective coefficient

In terms of the solvability of (3.22a)-(3.22b), there are still some things which can be said about Γ, even without solving the so-called cell-problems (3.18a)-(3.18b), and even without specifying the reference cell geometry (although some constraints related to the regularity is required, see e.g. [3] for more details). In particular, the symmetry and positive definiteness properties of Γ are readily shown.

We begin by showing the *symmetry* property: Multiplying (3.18a) by U^i and integrating over Y yields

$$\int_Y \gamma(y)(\nabla_y U^j(y) + \mathbf{e}_j) \cdot \nabla_y U^i(y) \mathrm{d}y = 0. \tag{3.23}$$

Using this, we can rewrite (3.20) as

$$[\Gamma]_{ij} = \int_Y \gamma(y)(\nabla_y U^j(y) + \mathbf{e}_j) \cdot (\nabla_y U^i + \mathbf{e}_i) \mathrm{d}y, \tag{3.24}$$

from which $[\Gamma]_{ij} = [\Gamma]_{ji}$ follows immediately, thus establishing the symmetry of Γ. Finally, we show the *positive definiteness* property: Let $0 \neq \alpha := [\alpha_1, ..., \alpha_d]^\top \in \mathbb{R}^d$ be non-zero constant vector. Then, using also (3.24) we can write

$$\Gamma\alpha \cdot \alpha = \sum_{i,j=1}^{d} [\Gamma]_{ij} \alpha_i \alpha_j \tag{3.25}$$

$$= \sum_{i,j=1}^{d} \int_Y \gamma(y)(\alpha_j(\nabla_y U^j(y) + \mathbf{e}_j)) \cdot (\alpha_i(\nabla_y U^i + \mathbf{e}_i)) \mathrm{d}y > 0, \tag{3.26}$$

from which the positive definiteness of Γ follows.

3.3 Limitations of homogenization

As already mentioned, homogenization is a powerful tool for deriving effective mathematical models. In many cases, effective models can also be derived using heuristic approaches, such as was originally done with e.g., Darcy's law, and Biot's model for quasi-static consolidation. These models were also derived using homogenization at a

later point in time, and thus homogenization in these situations, and in some sense, validated the previously taken heuristic approaches. There are however some notable differences in the upscaled models coming from heuristic/experimental approaches and from the homogenization approach; the coefficients in the upscaled model are more accurately understood in the case of homogenization. As the example in the previous section demonstrates, there appear always formulas for the effective coefficients, and their dependency on the micro-scale geometry is made explicit. Thus, at least in principle, as long as the coefficients of the micro-scale problem are known, and a geometry of the reference cell is specified, then the effective coefficients of the upscaled problem are also known. In the heuristic approach, however, it may be difficult to interpret the effective coefficients, and it may even be impossible to quantify their precise dependency on the micro-scale heterogeneities.

The drawback of homogenization, in particular within the periodic framework, is that strictly speaking, it is only applicable for a perfectly periodic arrangement of the micro-scale heterogeneities. For some human made composite material, which is designed to be periodic in its microstructure, the periodicity requirement is of course not a drawback, but for any naturally formed composite this will never fully be the case. Thus, although the effective coefficients can be calculated explicitly, the geometry of the reference cell of which this calculation depends, will in general only be a guess. Furthermore, even if a realistic geometry for the reference cell is chosen, the periodicity assumption might still be enough to produce unrealistic values for the effective coefficients.

In light of the above discussion, homogenization should not be viewed as a substitute for the heuristic and experimental approaches for deriving upscaled models, but rather as a supplement. One of the strengths of homogenization, even within the periodic framework, is that the upscaled system may be derived without specifying the geometry of the reference cell. In other words, the structure of the upscaled equations can be easily revealed even if the coefficients are unknown. Furthermore, from the auxiliary problems one can also try to deduce something about the relationship between the effective coefficients and the original (micro-scale) coefficients, which can be quite useful on its own. For an accurate simulation of real world phenomena, however, the formulas for the effective coefficients produced by homogenization should not be relied upon to give realistic values.

Chapter 4

Iterative numerical methods

Iterative numerical methods are, broadly speaking, sequential procedures where an initial guess is used to generate a sequence of improving solutions, approaching the true solution in an asymptotical fashion. Iterative methods can thus never be used to solve exactly any given problem, but can approximate it to any desired degree. *Iterative methods* are contrasted to *direct methods*, which attempts to solve the given problem by a finite numbers of operations. In the absence of rounding errors, the direct method is thus characterized by delivering the exact solution. However, in practical computations there are always rounding errors, hence the best result one can hope for is to be within machine precision of the true solution, direct method or not. In fact, good iterative methods can easily reach the true solution within machine precision, and in many cases do so much faster than a direct method can. The approximate nature of iterative methods is therefore irrelevant.

Another thing to consider is the fact that direct methods are rarely applicable to nonlinear problems, while iterative methods can be applied to both linear and nonlinear problems. Thus, for many types of nonlinear problems, iterative methods are the only option. Furthermore, iterative methods can also turn out to be the best choice for a linear problem, especially if the linear problem involves a large number of variables (making a direct method prohibitively computationally expensive). To elaborate on this point, consider the problem of inverting a square matrix A, or solving a linear system $A\mathbf{x} = \mathbf{b}$. Any direct method will necessarily require $\mathcal{O}(n^3)$ work (i.e., floating point operations) for a matrix A of rank n, since there are $\mathcal{O}(n)$ steps to be performed, each requiring $\mathcal{O}(n^2)$ amount of work. If a linear system is to be solved, then an iterative method may only need to compute $A\mathbf{x}$ for any given vector $\mathbf{x} \in \mathbb{R}^n$. If computing $A\mathbf{x}$ requires $\mathcal{O}(n^2)$ amount of work, and the method reaches machine precision in less than $\mathcal{O}(n)$ steps, then the iterative method will beat a direct method. This is for example the case with *Krylov subspace methods* [85, 92], where a basis of the form $\{A\mathbf{b}, A^2\mathbf{b}, A^3\mathbf{b}, ...\}$ is computed, and a solution sought in the span of this basis which minimizes the residual. It is evident that this method will converge in n steps, but if n is very large, the iterative procedure may reach sufficient accuracy long before that. Moreover, if the matrix under consideration

has some structure which can be exploited, the amount of work may be reduced even further. In particular, if the matrix A is *sparse*, with μ nonzero entries per row, the amount of work needed to compute $A\mathbf{x}$ is $\mathcal{O}(\mu n)$. The figure 4 below shows a schematic of how an iterative method might beat a direct method.

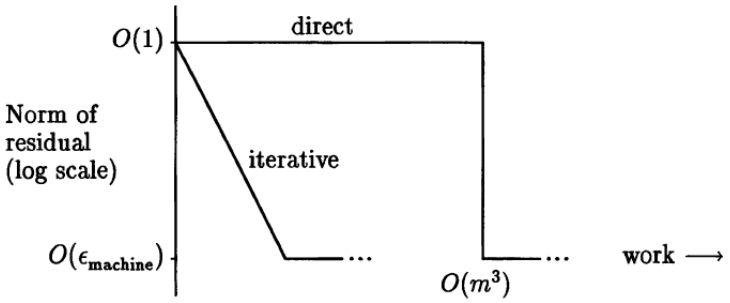


Figure 4.1: Illustration of convergence of direct and iterative methods for solving a large linear system. The direct method delivers the solution at machine precision when $\mathcal{O}(n^3)$ amount of work is performed, while the iterative method converges (geometrically) from the beginning, and reaches machine precision much faster than the direct method. Illustration is copied from [92].

The focus in this chapter will be on iterative numerical methods for solving partial differential equations (PDEs). Also in this case, an iterative method may be advantageous over a direct method, even if the PDE in question is linear. In fact, the situation is closely linked to the case of a linear system of equations described above: If the problem at hand is a coupled system of linear PDEs, with a large number of independent variables, then solving monolithically the whole discretized system may lead to a prohibitively large matrix. Solving instead each equation in a sequential fashion, while iteratively updating coupling terms, may turn out to be more accurate and faster than a direct monolithic computation. Furthermore, linear systems arising from discretizations of PDEs almost always has some exploitable structure (e.g., sparseness). There are several iterative methods which are often applied to PDEs. In the following sections, we will discuss the Newton method and the Fixed Stress Spitting / L-scheme method. Several examples will also be provided.

4.1 Newton's method...

Newton's method was originally designed to find zeroes of real valued functions, but can also be used in the context of PDEs. We describe first Newton's method for real valued functions, and then give an example with PDEs.

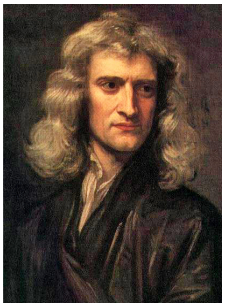


Figure 4.2: English 16th century physicist and mathematician Sir Isaac Newton (1642-1726), whose name among other things is attached to a popular iterative numerical method. Picture is copy of a portrait made by Sir Godfrey Kneller (1689).

4.1.1 ...for a real valued function

Let $f(x)$ be a real valued function, defined on all of $\mathbb{R}$. If x_0 is a (good enough) initial guess for a zero of f, then

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{4.1}$$

is a better approximation of that zero. By repeating this process according to

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{4.2}$$

the zero can be found to any degree of accuracy. Newton's method is famous for its quadratic convergence property, i.e., using Taylor's Theorem, one can show that the following estimate holds

$$\varepsilon_{n+1} \leq C\varepsilon_n^2, \tag{4.3}$$

for some generic constant $C > 0$ (independent of n), where $\varepsilon_n = |x_n - \hat{x}|$ and where $\hat{x} := \lim_{n\to\infty} x_n$ is the true zero. However, convergence of Newton's method is not guaranteed. The following conditions must be satisfied, where $I := [\hat{x}-r, \hat{x}+r]$ for some $r \geq |x_0+\hat{x}|$:

- $f'(x) \neq 0$ for all $x \in I$.
- $f''(x)$ is continuous for all $x \in I$.
- x_0 is chosen sufficiently close to $\hat{x}$.

Here, *sufficiently close* must be interpreted according to the situation, although it is possible to be more precise about this point (see e.g. [87]).

4.1.2 ...for PDEs

Consider now the following nonlinear PDE problem: Find $u(x)$ such that

$$-\nabla \cdot (A(u)\nabla u) = f, \quad \text{in } \Omega, \tag{4.4a}$$

$$u = 0, \text{ on } \partial\Omega. \tag{4.4b}$$

where $A(u)$ is some sufficiently smooth operator, and $f(x)$ is some given real valued function defined on some open and bounded subset $\Omega \subset \mathbb{R}^d$. Assuming now the above problem has been discretized, and taking some $V_h \subset H_0^1(\Omega)$ to be an appropriate discrete space (where h is the mesh size parameter), we readily obtain the discrete variational formulation as: Find $u_h \in V_h$ such that

$$\mathscr{F}(u_h)(v) := (A(u_h)\nabla u_h, \nabla v) = (f, v), \quad \forall v \in V_h. \tag{4.5}$$

In order to apply Newton's method to this problem, we need to calculate the Jacobian of the differential operator $\mathscr{F}$, i.e., we need to compute the limit

$$\begin{aligned}\mathscr{F}'(u_h)(\delta u_h, v) =& \lim_{s\to 0} \frac{1}{s}(\mathscr{F}(u_h + s\delta u_h)(v) - \mathscr{F}(u_h)(v)) \\ =& (A'(u_h)\nabla u_h, \nabla v) + (A(u_h)\nabla \delta u_h, \nabla v), \qquad \delta u_h \in V_h. \end{aligned} \tag{4.6}$$

Newton's method applied to the problem (4.5) (i.e., to the problem $\mathscr{F}(u_h)(v) - (f, v) = 0$) then reads as follows: (1) Let $u_h^0 \in V_h$ be an initial guess. (2) For the iteration steps $k = 1, 2, 3, ...$, we then solve the following (linear) problem: Find $\delta u_h^k \in V_h$ such that

$$\mathscr{F}'(u_h^k)(\delta u_h^k, v) = -\mathscr{F}(u_h^k)(v) - (f, v), \quad \forall v \in V_h. \tag{4.7}$$

(3) Update the solution by $u_h^{k+1} = u_h^k + \delta u_h^k$. Finally, either the algorithm is aborted if convergence is reached, or steps (2)-(3) are repeated. The criterion for convergence is given in terms of the specified tolerance TOL > 0, i.e., the method has converged if the following criterion is satisfied

$$\|\mathscr{F}(u_h^{k+1})(v)\| < \text{TOL}. \tag{4.8}$$

Newton's method can also be applied to a system of nonlinear PDEs, at the cost of the Jacobian becoming increasingly complex. When dealing with a nonlinear PDE, Newton's method is still often the first choice due to its quadratic convergence properties. Also, because modern PDE solving software usually has some built-in differentiation routine, calculating the Jacobian of the system by hand is rarely necessary (which can also be quite tedious). If, for some reason one of the above bullet points in the previous section are not fulfilled such that Newton's method does not converge, a more robust linearly convergent method can be chosen instead. For practitioners, usually some balance between robustness and efficiency is sough when choosing which numerical method to apply. If the emphasis is on efficiency, then Newton's method cannot be beat (assuming the method does in fact converge). If robustness is the most important factor, however, then Newton's method will rarely be a good choice. It is worth mentioning also that several modifications to Newton's method have been developed in order to increase its robustness, see e.g [39] for more details.

4.2 The L-scheme / Fixed Stress Splitting scheme

The L-scheme is an iterative scheme which generalizes the *Fixed Stress Split* and *Undrained Split* algorithms. These algorithms originate from the field of poroelastic-

ity, where they were originally designed for solving Biot's equations. In general, iterative splitting procedures for linear poroelasticity have been studied extensively (see e.g., [1, 15, 30, 50, 51, 55, 91]), and amongst them, the Fixed Stress Split and Undrained Split algorithms stand out as particularly robust choices. Both these algorithms involve decoupling the Biot system into two subproblems; mechanics and flow. These subproblems are then solved sequentially with mutually updated solution information. Specifically, the former method involves keeping a constant volumetric mean total stress during solution of flow problem, while the latter involves keeping constant fluid mass during the elastic structure deformation. The reason for the popularity of these two algorithms is the unconditional stability property, which was first shown in [51]. It is also worth mentioning that these algorithms are significantly easier to analyze than, say, a Newton method, which may involve complicated derivatives. In [67, 69] the convergence rates of these two algorithms were derived.

In the context of coupled problems, the L-scheme involves adding artificial stabilization terms to one or more of the subproblems with one or more stabilization parameters which are free to be chosen. Thus, in contrast to the two previously mentioned algorithms, the quantities held constant during solving of the subproblems need not have any physical interpretation. Furthermore, the L-scheme also has the property that it is easily analyzed. For these reasons, the L-scheme allows for further optimization than does the Undrained Split / Fixed Stress Split algorithms. However, good choices for the stabilization parameters is crucial for the L-scheme to perform efficiently and robustly. Therefore, some analysis is usually necessary in order to get an estimate for these. Another advantage of the L-scheme is that it can be used both as a stabilization technique, and as a linearization of nonlinear problems. In [63, 81] the L-scheme was used to solve Richard's equation. In [14, 15] it was used to solve both linear and nonlinear coupled flow and geomechanics, and in [52] variations of the L-scheme was used to solve nonlinear thermo-poroelasticity.

4.2.1 The L-scheme in practice

As a first application of the L-scheme, we employ the it on the discrete variational problem from the previous section (4.5): Let $u_h^0 \in V_h$ be some initial guess. Then, for the iteration steps $k = 1, 2, 3, ...$, we have given u_h^{k-1}, and seek u_h^k such that

$$L(u_h^k - u_h^{k-1}, v) + (A(u_h^{k-1})\nabla u_h^k, \nabla v) = (f, v), \quad \forall v \in V_h, \tag{4.9}$$

where $L > 0$ is some parameter which should be chosen with care. The first term on the left hand side is an *artificial stabilization term*, and tends to zero as the iterates approaches the solution. Before convergence is achieved, however, this term acts to stabilize the problem and may improve the convergence rate (depending on the choice of L).

To illustrate the L-scheme further, we employ it on a nonlinear coupled problem

(taken from [8]), which reads as follows:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} - AX(t) - Bu(1,t) = 0, \qquad t > 0, \tag{4.10a}$$

$$\frac{\partial u(x,t)}{\partial t} - \gamma \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad x \in (0,1], t > 0, \tag{4.10b}$$

$$u(0,t) = X(t), \qquad t > 0, \tag{4.10c}$$

$$\frac{\partial u(1,t)}{\partial x} = 0, \qquad t > 0, \tag{4.10d}$$

with initial data $u(x,0) = u_0(x)$ and $X(0) = X_0$. The coefficients A, B, γ are all assumed to be real positive constants. Before giving the discrete formulation of this problem, we define the following solution space $V := \{v \in H^1(0,1) : v|_{x=0} = 0\}$, and let $\tilde{u}(x,t) := u(x,t) - u_D(x,t)$, where $u_D(x,t) := X(t)$ (i.e., u_D is defined to be constant in space). The variational formulation of (4.10b) is then given by: Find $u(t) = \tilde{u}(t) + u_D(t)$ such that $\tilde{u}(t) \in V$ satisfying for all $t > 0$

$$(\partial_t \tilde{u}(t), v) + \gamma(\partial_x \tilde{u}(t), v) = -(\partial_t u_D(t), v), \quad \forall v \in V. \tag{4.11}$$

Here, the Dirichlet boundary condition (4.10c) depends on $X(t)$, which is given by solving the ODE (4.10a), which again depends on $u(1,t)$. To solve this system numerically, we will employ a variant of the L-scheme, which involves adding artificial stabilization terms (with stabilization parameters $L_u, L_X > 0$) to both the PDE (4.10b) and ODE (4.10a), in addition to a linearization of the nonlinear coupling between these equations.

Before presenting the iterative scheme, we discretize the above coupled problem in space and time, in particular we let $V_h \subset V$ be an appropriate finite dimensional space for the spatial discretization of the PDE (4.10b), where h is the mesh size of a uniform partition of $[0,1]$, and we employ a backward Euler method in time for both the PDE (4.10b) and ODE (4.10a) where $\tau > 0$ is the size of the time increment. With these definitions, the iterative scheme reads as follows: At time step $n \geq 1$, let $u^{n,0}$ and $X^{n,0}$ be initial guesses. At the iteration $k \geq 1$, let $\tilde{u}_h^{n,k-1}$ and $X^{n,k-1}$ be given, and find $\tilde{u}_h^{n,k}$ such that

$$\begin{aligned} L_u(\tilde{u}_h^{n,k} - \tilde{u}_h^{n,k-1}, v) + (\tilde{u}_h^{n,k}, v) + \tau\gamma(\partial_x \tilde{u}_h^{n,k}, \partial_x v) \\ = (u_h^{n-1}, v) - (u_D^{n,k-1}, v) + (u_D^{n-1}, v), \qquad \forall v \in V_h. \end{aligned} \tag{4.12}$$

Then, given $\tilde{u}^k$, find X^k such that

$$L_X(X^{n,k} - X^{n,k-1}) + X^{n,k} - \tau A X^{n,k} - \tau B X^{n,k} = X^{n-1} + \tau B \tilde{u}_h^{n,k}(x = 1). \tag{4.13}$$

The above procedure is then repeated until convergence, and then either the time step is incremented or the computation is terminated. If we let $\mathrm{aTOL}, \mathrm{rTOL} > 0$ be the chosen absolute and relative tolerances, respectively, then the method has converged (for the current time step) if the following criterion is satisfied

$$\|\tilde{u}_h^{n,k} - \tilde{u}_h^{n,k-1}\| \leq \mathrm{aTOL} + \mathrm{rTOL}\|\tilde{u}_h^{n,k}\|, \tag{4.14}$$

$$\text{and } |X^{n,k} - X^{n,k-1}| \leq \mathrm{aTOL} + \mathrm{rTOL}|X^{n,k}|. \tag{4.15}$$

4.2.2 Convergence rates

A convergence proof for the algorithm described above is readily obtained. In particular, if $\tilde{u}^n$ and X^n are the exact solutions to (4.10a)-(4.10d) at time step n, then we will show a contraction of successive differences, defined by $e^k_{\tilde{u}} := \tilde{u}^{n,k}_h - \tilde{u}^n$ and $e^k_X := X^{n,k} - X^n$, which will imply convergence of the scheme by the Banach Fixed Point Theorem (see e.g., [32]). This analysis will also reveal the convergence rate of the scheme. We state it as the following theorem:

Theorem 4.2.1 (Convergence)**.** *If $L_u, L_X > 0$ and if the time step τ satisfies*

$$0 < \tau < \frac{3}{4}\left(A + BC + \frac{B^2 c_{\mathrm{tr}}}{2\gamma}\right)^{-1}, \tag{4.16}$$

then the following contraction estimate can be obtained from the scheme (4.12)-(4.13)

$$\begin{aligned}\left(\frac{L_u}{2} + \tau\frac{\gamma}{2c_P}\right)\|e^k_{\tilde{u}_h}\|^2 + \left(\frac{L_X}{2} + 1 - \tau\left(A + B + \frac{B^2 c_{\mathrm{tr}}}{2\gamma}\right)\right)|e^k_X|^2 \\ \leq \frac{L_u}{2}\|e^{k-1}_{\tilde{u}_h}\|^2 + \left(\frac{L_X}{2} + \frac{1}{4}\right)|e^{k-1}_X|^2,\end{aligned} \tag{4.17}$$

where $c_{\mathrm{tr}} > 0$ and $c_P > 0$ are domain specific constants coming from the trace and Poincaré inequalities, respectively.

Proof. We begin by subtraction equations (4.12)-(4.13) solved by the exact solutions $(\tilde{u}^n, X^n)$ from the same equations solved by the iterate solutions. Defining the following difference functions $e^k_{\tilde{u}} := \tilde{u}^{n,k}_h - \tilde{u}^n$, $e^k_D := u^{n,k}_D - u^n_D$, and $e^k_X := X^{n,k} - X^n$, we obtain for all $n \geq 1$ the corresponding set of difference equations as

$$L_u(e^k_{\tilde{u}} - e^{k-1}_{\tilde{u}}, v) + (e^k_{\tilde{u}}, v) + \tau\gamma(\partial_x e^k_{\tilde{u}}, \partial_x v) = (e^{k-1}_D, v), \qquad \forall v \in V, \tag{4.18a}$$

$$L_X(e^k_X - e^{k-1}_X) + e^k_X - \tau(A + B)e^k_X = \tau B e^k_{\tilde{u}}(x = 1). \tag{4.18b}$$

Taking $v = e^k_{\tilde{u}}$ in equation (4.18a), and multiplying equation (4.18b) by e^k_X, and adding together the resulting equations yields the following inequality

$$\begin{aligned}\left(\frac{L_u}{2} + 1\right)\|e^k_{\tilde{u}}\|^2 + \tau\gamma\|\partial_x e^k_{\tilde{u}}\|^2 + \left(\frac{L_X}{2} + 1 - \tau(A + B)\right)|e^k_X|^2 \\ \leq \frac{L_u}{2}\|e^{k-1}_{\tilde{u}}\|^2 + \frac{L_X}{2}|e^{k-1}_X|^2 + \|e^{k-1}_D\|\|e^k_{\tilde{u}}\| + \tau B|e^k_X||e^k_{\tilde{u}}(x = 1)|,\end{aligned} \tag{4.19}$$

where we also used the Cauchy-Schwarz inequality. Since by the trace-inequality we have

$$|e^k_{\tilde{u}}(x = 1)|^2 \leq c_{\mathrm{tr}}\|\partial_x e^k_{\tilde{u}}\|^2, \tag{4.20}$$

for some generic (domain-specific) constant $c_{\mathrm{tr}} > 0$, and since $\|e^{k-1}_D\| = |e^{k-1}_X|$ (due to

the spatial domain being the unit interval), we can write (4.19) as

$$\left(\frac{L_u}{2}+1-\frac{\delta_1}{2}\right)\|e_{\tilde{u}}^k\|^2+\left(\tau\gamma-\frac{1}{2\delta_2}\tau B c_{\mathrm{tr}}\right)\|\partial_x e_{\tilde{u}}^k\|^2 + \left(\frac{L_X}{2}+1-\tau\left(A+B+B\frac{\delta_2}{2}\right)\right)|e_X^k|^2 \leq \frac{L_u}{2}\|e_{\tilde{u}}^{k-1}\|^2+\left(\frac{L_X}{2}+\frac{1}{2\delta_1}\right)|e_X^{k-1}|^2, \tag{4.21}$$

where the constants $\delta_1, \delta_2 > 0$ are coming from application of the Young inequality. Choosing $\delta_1 = 2$ and $\delta_2 = Bc_{\mathrm{tr}}/\gamma$ yields

$$\frac{L_u}{2}\|e_{\tilde{u}}^k\|^2+\frac{\tau\gamma}{2}\|\partial_x e_{\tilde{u}}^k\|^2+\left(\frac{L_X}{2}+1-\tau\left(A+B+\frac{B^2 c_{\mathrm{tr}}}{2\gamma}\right)\right)|e_X^k|^2 \leq \frac{L_u}{2}\|e_{\tilde{u}}^{k-1}\|^2+\left(\frac{L_X}{2}+\frac{1}{4}\right)|e_X^{k-1}|^2. \tag{4.22}$$

Finally, by employing the Poincaré inequality on the second term on the left hand side, i.e.

$$\|e_{\tilde{u}}^k\|^2 \leq c_P\|\partial_x e_{\tilde{u}}^k\|^2, \tag{4.23}$$

for some generic (domain-specific) constant $c_P > 0$, we obtain

$$\left(\frac{L_u}{2}+\tau\frac{\gamma}{2c_P}\right)\|e_{\tilde{u}}^k\|^2+\left(\frac{L_X}{2}+1-\tau\left(A+B+\frac{B^2 c_{\mathrm{tr}}}{2\gamma}\right)\right)|e_X^k|^2 \leq \frac{L_u}{2}\|e_{\tilde{u}}^{k-1}\|^2+\left(\frac{L_X}{2}+\frac{1}{4}\right)|e_X^{k-1}|^2. \tag{4.24}$$

Thus, if the time step satisfies the constraint (4.16), the estimate (4.24) is a contraction, and convergence of the sequences $\{\tilde{u}_h^{n,k}\}_k$ and $\{X^{n,k}\}_k$ for each $n \geq 1$ follows. □

It is worth noting that convergence of the L-scheme in this case is only *conditional*, i.e., if the time increment τ is too big, the method will not converge. This is a typical situation for nonlinear coupled problems.

Chapter 5

Introduction to the papers

This chapter provides an introduction to the included papers, all of which are either published or submitted for publication in scientific journals.

5.1 Paper A

Title: *Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium*

Authors: Brun, Mats Kirkesæther and Berre, Inga and Nordbotten, Jan Martin and Radu, Florin Adrian

Journal: Transport in Porous Media 124, 1 (2018).

Pages: 137–158

Publisher: Springer

This paper concerns the upscaling of a thermal fluid-structure interaction problem in the context of porous media. The resulting system of equations on the macro-scale (i.e., the scale at which the fluid saturated elastic matrix can be replaced by a homogenized 'fictitious' material) extends the well-known linear poroelastic model known as *Biot's quasi-static consolidation model* to the non-isothermal case. This derivation provides a precise understanding of the coupling terms at the macro-scale, and forms a justification for previous heuristically derived models which are present in the literature [36, 45, 89]. In particular, we undertake a formal derivation of a *poro-thermo-elastic system* within the framework of *quasi-static deformation*. This work is based upon the well-known derivation of the quasi-static poroelastic equations (i.e., the previously mentioned Biot model) by homogenization of the fluid-structure interaction at the pore-scale [23, 34, 86]. However, compared to these works, we now include energy conservation at the pore-scale, which is coupled to the fluid-structure model by using linear thermoelasticity for the solid structure, coupled with thermal flow in the fluid saturated void space. The resulting upscaled model is similar to the Biot model, but with an added conservation of

energy equation, fully coupled to the momentum and mass conservation equations. In particular, we obtain a system of equations on the macro-scale accounting for the effects of elastic mechanical deformation, heat transfer, and fluid flow within a fully saturated porous material. Moreover, two different scaling regimes of the pore-scale system are considered: One where the Péclet number is small (resulting in heat transfer dominated by thermal diffusion), and another where it is unity (resulting in heat transfer both by thermal diffusion and thermal convection). The upscaled models corresponding to both choices of scaling are presented.

The most important limitation of the formal approach taken herein is the assumption of a perfectly periodic geometry within the porous medium. This is a common assumption in applications of homogenization (especially in the context of porous media), although this is rarely satisfied in practice. However, it has been shown for similar problems that the periodicity assumption can be relaxed, and we expect that these results are possible to extend to the present setting as well. As such, we expect the structure of the equations derived in this paper to be valid, in particular for non-periodic *natural* porous media, at least when there is some uniformity on the sizes and shapes of the solid grains making up the porous structure. Furthermore, the combined Lagrangian-Eulerian system employed at the pore-scale is transformed into a purely Lagrangian formulation. Thus, due to the use of linear thermoelasticity as the governing equations for the solid at the pore-scale, it is only the elastic mechanical strain which is assumed to be small, and not the displacement itself. Alternatively, if an Eulerian framework was used, it would be necessary to assume that the displacement of the solid structure is small relative to the pore-scale, which would then preclude meaningful macroscopic deformations.

In this work we have chosen to a large extent to linearize the governing equations already at the pore-scale. This is in accordance with the pore-scale model leading to the (isothermal) Biot system, and in part explains the linear structure of the majority of terms on the macro-scale. Nonlinear constitutive relationships could be accommodated at the cost of technical and notational complexity, varying from relatively straight-forward (i.e., nonlinear constitutive laws for fluid density) to complex (nonlinear elastic or plastic constitutive laws for material deformation). However, in the case of a Péclet number of order one, there is introduced a nonlinear coupling in the upscaled model, which accounts for the non-negligible thermal convection. Its presence makes the upscaled energy conservation equation a nonlinear one, and therefore complicates the model compared to the isothermal situation (i.e., to Biot's model). Thus, the multitude of results regarding the isothermal system (e.g. well-posedness, numerical schemes, preconditioners, etc.) can not be directly transferred to the thermal situation as long as there appears thermal convection. On the other hand, if thermal convection is neglected, the upscaled system becomes fully linear, and very similar to the isothermal system. In fact, by defining a new variable $\xi := p - T$, where p and T are the pressure and temperature variables, respectively, the three-field thermo-poroelastic system is reduced to a two-field system, which is formally equivalent to the isothermal system.

The derived effective coefficients of the upscaled system can be explicitly calculated in terms of the microstructure of the porous material, by solving so-called *auxiliary*

problems. It is here that the periodicity assumption becomes most relevant; in order to solve these *auxiliary problems*, one must specify a reference geometry, which then in part determines the calculated value of the effective coefficients. Thus, the periodicity assumption results in having constant effective coefficients, and not spatially dependent ones. The derived effective coefficients herein are not computed, but we do establish the symmetry and positive definiteness properties for the ones which are new in the literature. We mention also that while finalizing this work, we were made aware that a similar derivation was undertaken simultaneously by other authors [93].

5.2 Paper B

Title: *Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport*

Authors: Brun, Mats Kirkesæther and Ahmed, Elyes and Nordbotten, Jan Martin and Radu, Florin Adrian

Journal: Journal of Mathematical Analysis and Applications 471, 1–2 (2019).

Pages: 239–266

Publisher: Elsevier

This paper concerns the mathematical analysis of *thermo-poroelasticy* within the context of *quasi-static deformation*. This model problem is nonlinear and includes thermal effects compared to the classical quasi-static poroelastic model, i.e. to *Biot's consolidation model*. Biot's model constitutes (in primal form) a two-field elliptic-parabolic system, of which there is an extensive literature, both in terms of mathematical analysis and numerical approximation. To mention a few, the well-posedness based on the canonical two-field formulation was first carried out in [88], while three and four-field formulations have also been analyzed (introducing Darcy flux and/or total poroelastic stress as independent variables), and can be found in several studies, e.g. [2, 80, 97]. Robust discretizations and preconditioners can be found in e.g. [6, 61]. A key feature of Biot's model, one which greatly facilitates its analysis, is the symmetric coupling between the two equations.

Compared to Biot's model, we now have a three-field coupled system, consisting of a momentum balance equation, a mass balance equation, and an energy balance equation, fully coupled and nonlinear due to a convective transport term in the energy balance equation. In this model, there also appears symmetric couplings between the heat and flow, and between the heat and mechanics, similar to the couplings found in the isothermal system. However, there also appears the non-symmetric coupling between the heat and flow, which is the previously mentioned thermal convective term. The aim of this article is to investigate, in the framework of mixed formulations, the existence and uniqueness of a weak solution to this model problem. The primary variables in these formulations are the *fluid pressure*, *temperature* and *elastic displacement* as well as the

Darcy flux, *heat flux* and *total (thermo-poroelastic) stress*. The well-posedness of a linearized formulation is addressed first, using the theory of DAEs (Differential Algebraic Equations), a *Galerkin method*, and suitable *a priori* estimates. This is used next to study the well-posedness of an iterative solution procedure, based on the previous linearized formulation, in order to approximate the full nonlinear problem. A convergence proof for this iterative algorithm is then inferred for small time intervals by a contraction of successive difference functions of the iterates using suitable norms, and by application of the Banach Fixed Point Theorem. Having obtained local solutions in time for the nonlinear problem, and due to the continuity in time of the convergent (local) solutions, we can infer a (global) convergence proof of the iterative procedure, thus establishing the well-posedness of the original nonlinear problem for arbitrary (finite) final time.

The main difficulty we encounter in this analysis is the regularity of the nonlinear term. In particular, in the fully mixed formulation, the convective term takes the form $\mathbf{w} \cdot \Theta^{-1}\mathbf{r}$, where $\mathbf{w}$ is the Darcy flux, $\mathbf{r}$ is the heat flux, and Θ is the thermal conductivity coefficient. A priori, the regularity of the fluxes satisfy $\|\mathbf{w}(t)\|_{H(\mathrm{div})} + \|\Theta^{-1}\mathbf{r}(t)\|_{H(\mathrm{div})} \leq C$, where $C > 0$ is a generic constant. However, in the variational formulation of the energy balance equation there appears the L^2-inner product $(\mathbf{w} \cdot \Theta^{-1}\mathbf{r}, S)$, where $S \in L^2$ is a test function for the temperature. Thus, for this inner product to be well-defined, there is needed either $\mathbf{w}(t) \in L^\infty$ or $\Theta^{-1}\mathbf{r}(t) \in L^\infty$, or $\mathbf{w}(t), \Theta^{-1}\mathbf{r}(t) \in L^4$. In order to deal with this issue, we consider two approaches: First, we consider the required L^∞-regularity for either of the fluxes to be given *a priori*. Second, consider sufficiently regular initial and source data is given, such that the solution to the linearized formulation admits L^4-regularity for both of the fluxes, thus allowing the same regularity to be inferred for the converged solution. We chose to include the latter approach only as an appendix, since with the former approach, only 'standard' energy estimates are necessary.

Although the literature on (isothermal) poroelasticity is extensive, there is not much literature on the analysis of thermo-poroelastic models; in [93] a corresponding energy functional for the thermo-poroelastic model was derived. This functional was then shown to be monotonically decreasing in time for a small enough characteristic temperature difference. For this reason, the analysis undertaken in Paper B addresses an important gap in the literature on thermo-poroelasticity.

5.3 Paper C

Title: *Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport*

Authors: Brun, Mats Kirkesæther and Ahmed, Elyes and Berre, Inga and Nordbotten, Jan Martin and Radu, Florin Adrian

Journal: In review (2019).

Preprint: `https://arxiv.org/abs/1902.05783`

This paper concerns *monolithic* and *splitting-based* iterative numerical procedures for

the coupled nonlinear thermo-poroelasticity model problem as described in [21, 59, 93]. This model problem is formulated as a three-field system of partial differential equations (PDEs), consisting of an energy balance equation, a mass balance equation and a momentum balance equation, where the primary variables are temperature, fluid pressure, and elastic displacement. However, due to the presence of a nonlinear convective transport term in the energy balance equation, it is convenient to have access to both the pressure and temperature gradients. Hence, we introduce these as two additional variables and extend the original three-field model to a five-field model. For the numerical solution of this five-field formulation, we compare six approaches that differ by how we treat the coupling/decoupling between the flow and/from heat and/from the mechanics, suitable for varying coupling strength between the three physical processes. These approaches have in common a simultaneous application of the linearization and stabilization treatments warranted by the structure of the problem. More precisely, the derived procedures transform a nonlinear and fully coupled problem into a set of simpler subproblems to be solved sequentially in an iterative fashion. We provide convergence proofs for the derived algorithms, and demonstrate their performance through several numerical examples. In particular, we pay special attention to investigating different strengths of the coupling between the flow, mechanics and heat.

The proposed six algorithms are all based on recent developments on iterative splitting schemes coming from linear poroelasticity, extended here to accommodate for nonlinear thermo-poroelasticity. These algorithms use stabilization and linearization techniques similar to [15, 63], which is known in the literature as the 'L-scheme'. The L-scheme can itself be seen as a generalization of the Undrained and Fixed-Stress Split algorithms [1, 15, 30, 50, 51, 55, 67, 69, 91], and works both to stabilize iterative splitting as well as to linearize nonlinear problems. The thermo-poroelastic problem we consider can be viewed as a coupling of three physical processes (or subproblems): Flow, mechanics and heat. Thus, solving this system either monolithically (all three subproblems simultaneously), partially decoupled (two subproblems simultaneously), or fully decoupled (each subproblem separately), yields six possible combinations of coupling/decoupling, which serves as the backdrop for the design of the six algorithms. All of these involve a linearization of the convective term and added stabilization terms to both the flow and heat subproblems. In this sense, our use of the L-scheme is both as a stabilization for iterative splitting, and as a linearization of nonlinear problems.

For any given situation the coupling strength between the three subproblems may vary. A-priori, the expectation is that solving together subproblems that are strongly coupled will yield better efficiency properties than does splitting. On the other hand, if the coupling between two or more subproblems is weak, a splitting procedure might be beneficial. For this reason, and due to the fact that splitting the three-way coupled multi-physics problem into smaller subproblems allows for combining existing codes that separately can handle any of the three processes involved (or two of them combined), six different algorithms are presented. In this sense, we provide a complete framework for splitting-type solution strategies for thermo-poroelasticity. Furthermore, using the well-posedness of the continuous problem, we obtain lower bounds on the stabilization

parameters, and prove the convergence of our proposed algorithms under a constraint on the size of the time step. In practice, however, we find that this bound is not tight; as long as the fluxes are not becoming unbounded (e.g., due to a singularity), a 'reasonable' time step can safely be chosen.

Our algorithms are tested in detail with several numerical examples. In particular, we find that all six algorithms are performing robustly with respect to both mesh refinement and different parameter regimes (i.e., strong/weak coupling between the subproblems and strong/weak nonlinear effects), using the stabilization revealed by our analysis. We also find that using no stabilization results in the algorithms being more sensitive to the parameter regimes, i.e. splitting subproblems that are strongly coupled yields high iteration numbers compared to solving these subproblems together. This phenomena is also observed in the stabilized algorithms, but to a significantly lesser extent. In particular, our conclusion is that with no stabilization, each of the algorithms is suitable only for a certain parameter regime (i.e., one that corresponds to the coupling/decoupling structure present in the algorithm), in contrast to the stabilized algorithms, which can handle a much wider range of different parameter regimes.

5.4 Paper D

Title: *An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters*

Authors: Brun, Mats Kirkesæther and Wick, Thomas and Berre, Inga and Nordbotten, Jan Martin and Radu, Florin Adrian

Journal: In review (2019).

Preprint: `https://arxiv.org/abs/1903.08717`

This paper concerns the analysis and implementation of a novel iterative staggered scheme for brittle fracture propagation within a quasi-static elastic medium, where the fracture evolution is tracked by a phase field variable. Herein, the phase field creates a diffusive transition zone around fracture surfaces with (half-)thickness $\varepsilon > 0$. The proposed algorithm employs stabilization and linearization techniques known in the literature as the 'L-scheme', which is a generalization of the 'Fixed Stress Splitting' algorithm coming from the field of poroelasticity. The model problem we consider is a two-field variational inequality system, with the phase field function $\varphi(x,t)$, and the elastic displacements of the solid material $u(x,t)$, as independent variables. Using a penalization strategy, this variational inequality system is transformed into a variational equality system, which is the formulation we take as the starting point for our algorithmic developments. The proposed scheme involves a partitioning of this model into two subproblems; phase field and mechanics, and with added stabilization terms to both subproblems (with stabilization parameters $L_\varphi, L_u > 0$) for improved efficiency and robustness.

Under the natural assumptions that the elastic mechanical energy remains bounded, and that the model parameter $\varepsilon > 0$ is sufficiently large (i.e., that the diffusive transi-

tion zone around crack surfaces must be sufficiently thick), we show that a contraction of successive difference functions in energy norms can be obtained from the proposed scheme. This result implies that the algorithm is converging monotonically with a linear convergence rate. However, in the convergence analysis there appears some unknown constants which makes the precise convergence rate, as well as the precise lower bound on ε, difficult to determine in practice.

We provide several numerical tests where our proposed scheme is tested in detail. In particular, the proposed scheme is employed on several numerical benchmark problems within the context of phase field brittle fracture propagation. Moreover, for each numerical example we provide results for different values of stabilization parameters, i.e., for most cases we let $L_u = L_\varphi > 0$, but for comparison we include also the stabilization configurations $L_u = 0$ with $L_\varphi > 0$, and $L_u = L_\varphi = 0$. These tests reveal that stabilizing both subproblems is necessary for a robust algorithm, which confirms our theoretical results coming from the convergence analysis. However, further work is needed to find an optimal configuration of L_u, L_φ. Furthermore, for all numerical tests we provide computational justification for the assumption of bounded elastic mechanical energy (i.e. we provide plots tracking the elastic mechanical energy with respect to iteration numbers).

A slight dependency on the mesh parameter h in the iteration numbers is observed in the numerical tests, but this is to be expected since we use $\varepsilon = 2h$, and as our analysis demonstrates, the convergence rate is dependent on ε. The variation in iteration numbers with mesh refinement is in any case sufficiently small enough that we can conclude our algorithm is robust with respect to mesh refinement. Furthermore, it is well known that in numerical simulations of brittle fracture propagation using phase fields, there appear spikes in iteration numbers (typical iteration counts are on the order of 1000 [46, 64, 95, 96]) at the critical loading steps, i.e. when the crack starts to propagate or is further propagating. We also observe such iteration spikes at the critical loading steps using the proposed scheme. For this reason, we have also included, for comparison, several results in which the iteration is truncated before convergence is reached. Due to the monotonic convergence property of the proposed scheme, this strategy still produces acceptable results (even when truncating the scheme at iteration numbers as low as 20), while effectively avoiding the iteration spikes. We therefore conclude, at least for the particular examples presented here, that a truncation of the proposed L-scheme may be employed for greatly improved efficiency with only minor loss of accuracy.

Chapter 6

Summary and outlook

Within the overarching framework of *thermo-mechanical subsurface energy storage*, two primary topics have been discussed in this dissertation: *Thermo-poroelasticity* and *phase field brittle fracture propagation*. While both are important subjects in their own right, and have a multitude of applications in various engineering fields, they are both particularly relevant for *geothermal applications in the subsurface*.

The part of this dissertation devoted to thermo-poroelasticity concerns both the *derivation*, *analysis*, and *numerical implementation* of a thermo-poroelastic system. In particular, a thermo-poroelastic system was derived with the purpose of extending the (isothermal) linear Biot system for quasi-static deformation to the non-isothermal case, i.e., to couple the Biot system to an energy conservation equation, with an additional variable representing the temperature distribution of the medium. This derivation was done using formal upscaling, i.e., the homogenization method of two-scale asymptotic expansions within the periodic framework. The emphasis here was on the structure of the upscaled model as a fully coupled and nonlinear system of PDEs. Formulas for the effective coefficients were derived, but not calculated explicitly. However, for the derived effective coefficients which were new in the literature, i.e., not part of the original isothermal Biot-system, their symmetry and positive definiteness properties were shown. In particular, in the upscaled system, these appeared as the coupling coefficients between the heat and flow, and between the heat and mechanics. The coupling coefficient between the flow and mechanics remained unchanged from the isothermal system.

The thermo-poroelastic model problem was then analyzed in the context of mixed formulations, i.e., the original three-field model (with *temperature*, *fluid pressure* and *elastic displacements* as variables) was extended to a six-field model (introducing the *heat flux*, *Darcy flux* and *total stress* as new variables). This was done in order to facilitate a future mixed finite element discretization of the thermo-poroelastic problem, since it is well-known that mixed formulations of elliptic and parabolic problems are locally mass conservative when discretized using mixed finite elements, e.g., the Raviart-Thomas [7, 44] and Arnold-Winther [4, 5, 29] elements. Another reason to consider the fully mixed formulation for the analysis was to take advantage of the recent developments in the

literature on fully mixed formulations of the Biot system [2, 60, 97]. This analysis was done in two steps: First, a linearized system was analyzed using a Galerkin technique together with weak compactness. Then, this linearized system was used to design an iterative procedure which was shown to converge to the solution of the original nonlinear problem.

Finally, six iterative numerical schemes were proposed for the thermo-poroelastic model problem. These schemes are all based on the linearization technique employed in the analysis part, and on the iterative scheme known as the L-scheme/Fixed Stress Splitting scheme. Thermo-poroelasticity can be viewed as a coupling of three physical processes; heat, flow and mechanics. This is also reflected in the equations which make up the system, which are readily divided into three corresponding subproblems. A natural approach to consider when faced with solving such a problem is therefore to decouple the three subproblems and solve them sequentially, while at the same time updating coupling terms. Thus, in order to provide a complete framework for this type of solution strategy in the context of thermo-poroelasticity, the proposed six algorithms were designed to cover all possibilities of coupling/decoupling of the three subproblems. In practice, this means either all three subproblems are decoupled and solved sequentially, or two subproblems are solved together decoupled from the third, or finally a linearized system is solved monolithically.

The second primary topic of this dissertation is phase field descriptions of brittle fracture propagation within a quasi-static elastic material. In particular, the numerical solution algorithms of such models. Based on developments on iterative splitting procedures coming from poroelasticity, a novel iterative splitting scheme for phase field brittle fracture propagation was proposed. This algorithm also employs stabilization and linearization techniques which are based on the L-scheme, as well as a decoupling of the two subproblems involved; phase field and mechanics. The convergence of this algorithm was proved under a natural condition that the diffusive transition zone around fracture surfaces must be sufficiently thick, and that the elastic mechanical energy remains bounded. Detailed numerical examples confirm these theoretical findings, and demonstrate the robustness of the proposed algorithm in practice.

A natural extension of this research project is to combine the results on *thermo-poroelasticity* with the results on *phase field brittle fracture propagation*. Specifically, the phase field description of brittle fracture propagation can be extended from a quasi-static elastic material to a quasi-static thermo-poroelastic material. Using phase field descriptions of brittle fracture propagation in connection with poroelasticity is nothing new [62, 70, 71], but so far there does not exist in the literature a phase field description of brittle fracture propagation in a thermo-poroelastic medium. In this context, the free energy functional related to the phase field should be of the form $E(\varphi, T, p, \mathbf{u})$. This will lead to a more complicated phase field equation, but the same linearization techniques used for the quasi-static elastic model will in all likelihood be applicable with some modifications. Moreover, the six iterative schemes proposed for thermo-poroelasticity can then be extended to the phase field thermo-poroelastic model, giving further possibilities of coupling/decoupling of the now four subproblems. A convergence proof of

such algorithms might also be achieved by combining the convergence proofs from the thermo-poroelasticity schemes with the phase field brittle fracture propagation scheme, since the same linearization and stabilization techniques are employed in both contexts.

Bibliography

[1] Ahmed, E., Nordbotten, J. M., and Radu, F. A. (2019). Adaptive asynchronous time-stepping, stopping criteria, and a posteriori error estimates for fixed-stress iterative schemes for coupled poromechanics problems. *arXiv preprint arXiv:1901.01206* .

[2] Ahmed, E., Radu, F. A., and Nordbotten, J. M. (2019). Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model. *Comput. Methods Appl. Mech. Engrg. 347*, 264–294. doi: 10.1016/j.cma.2018.12.016.

[3] Allaire, G. (1992). Homogenization and two-scale convergence. *SIAM J. Math. Anal. 23*(6), 1482–1518. doi: 10.1137/0523084.

[4] Arnold, D. N., Falk, R. S., and Winther, R. (2007). Mixed finite element methods for linear elasticity with weakly imposed symmetry. *Math. Comp. 76*(260), 1699–1723. doi: 10.1090/S0025-5718-07-01998-9.

[5] Arnold, D. N. and Winther, R. (2002). Mixed finite elements for elasticity. *Numer. Math. 92*(3), 401–419. doi: 10.1007/s002110100348.

[6] Bærland, T., Lee, J. J., Mardal, K.-A., and Winther, R. (2017). Weakly imposed symmetry and robust preconditioners for Biot's consolidation model. *Comput. Methods Appl. Math. 17*(3), 377–396. doi: 10.1515/cmam-2017-0016.

[7] Bahriawati, C. and Carstensen, C. (2005). Three MATLAB implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control. *CMAM 5*(4), 333–361.

[8] Baudouin, L., Seuret, A., and Gouaisbaut, F. (2019). Stability analysis of a system coupled to a heat equation. *Automatica J. IFAC 99*, 195–202. doi: 10.1016/j.automatica.2018.10.021.

[9] Bejan, A. and Kraus, A. D. (2003). Heat transfer handbook, volume 1. John Wiley & Sons.

[10] Biner, S. B. (2017). Programming phase-field modeling. Springer, Cham. ISBN 978-3-319-41194-1; 978-3-319-41196-5. doi: 10.1007/978-3-319-41196-5.

[11] Biot, M. A. (1941). General theory of three-dimensional consolidation. *Journal of Applied Physics 12*(2), 155–164.

[12] Biot, M. A. (1965). Mechanics of incremental deformations. Theory of elasticity and viscoelasticity of initially stressed solids and fluids, including thermodynamic foundations and applications to finite strain. John Wiley & Sons, Inc., New York-London-Sydney.

[13] Biot, M. A. (1972). Theory of finite deformations of porous solids. *Indiana University Mathematics Journal 21*(7), 597–620.

[14] Borregales, M., Radu, F. A., Kumar, K., and Nordbotten, J. M. (2018). Robust iterative schemes for non-linear poromechanics. *Comput. Geosci. 22*(4), 1021–1038. doi: 10.1007/s10596-018-9736-6.

[15] Both, J. W., Borregales, M., Nordbotten, J. M., Kumar, K., and Radu, F. A. (2017). Robust fixed stress splitting for Biot's equations in heterogeneous media. *Appl. Math. Lett. 68*, 101–108. doi: 10.1016/j.aml.2016.12.019.

[16] Bourdin, B., Francfort, G. A., and Marigo, J.-J. (2008). The variational approach to fracture. Springer, New York. ISBN 978-1-4020-6394-7. doi: 10.1007/978-1-4020-6395-4. Reprinted from J. Elasticity **91** (2008), no. 1-3 [MR2390547], With a foreword by Roger Fosdick.

[17] Brezzi, F. (1974). On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge 8*(R-2), 129–151.

[18] Bringedal, C., Berre, I., Pop, I. S., and Radu, F. A. (2015). A model for non-isothermal flow and mineral precipitation and dissolution in a thin strip. *J. Comput. Appl. Math. 289*, 346–355. doi: 10.1016/j.cam.2014.12.009.

[19] Bringedal, C., Berre, I., Pop, I. S., and Radu, F. A. (2016). Upscaling of non-isothermal reactive porous media flow with changing porosity. *Transp. Porous Media 114*(2), 371–393. doi: 10.1007/s11242-015-0530-9.

[20] Bringedal, C., Berre, I., Pop, I. S., and Radu, F. A. (2016). Upscaling of nonisothermal reactive porous media flow under dominant Péclet number: the effect of changing porosity. *Multiscale Model. Simul. 14*(1), 502–533. doi: 10.1137/15M1022781.

[21] Brun, M. K., Berre, I., Nordbotten, J. M., and Radu, F. A. (2018). Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium. *Transp. Porous Media 124*(1), 137–158. doi: 10.1007/s11242-018-1056-8.

[22] Brun, M. K. t., Ahmed, E., Nordbotten, J. M., and Radu, F. A. (2019). Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport. *J. Math. Anal. Appl. 471*(1-2), 239–266. doi: 10.1016/j.jmaa.2018.10.074.

[23] Burridge, R. and Keller, J. B. (1981). Poroelasticity equations derived from microstructure. *The Journal of the Acoustical Society of America 70*(4), 1140–1146.

[24] Caginalp, G. and Fife, P. (1986). Higher-order phase field models and detailed anisotropy. *Phys. Rev. B (3) 34*(7), 4940–4943. doi: 10.1103/PhysRevB.34.4940.

[25] Caginalp, G. and Jones, J. (1995). A derivation and analysis of phase field models of thermal alloys. *Annals of physics 237*(1), 66–107.

[26] Caginalp, G. and Socolovsky, E. A. (1991). Computation of sharp phase boundaries by spreading: the planar and spherically symmetric cases. *J. Comput. Phys. 95*(1), 85–100. doi: 10.1016/0021-9991(91)90254-I.

[27] Cahn, J. W. and Hilliard, J. E. (1958). Free energy of a nonuniform system. I. Interfacial free energy. *The Journal of chemical physics 28*(2), 258–267.

[28] Cahn, J. W. and Hilliard, J. E. (1959). Free energy of a nonuniform system. III. Nucleation in a two-component incompressible fluid. *The Journal of chemical physics 31*(3), 688–699.

[29] Carstensen, C., Günther, D., Reininghaus, J., and Thiele, J. (2008). The Arnold-Winther mixed FEM in linear elasticity. Part I: Implementation and numerical verification. *Comput. Methods Appl. Mech. Engrg. 197*, 3014–3023.

[30] Castelletto, N., White, J., and Tchelepi, H. (2015). Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics. *International Journal for Numerical and Analytical Methods in Geomechanics 39*(14), 1593–1618.

[31] Chen, L. and Yang, W. (1994). Computer simulation of the domain dynamics of quenched system with a large number of non-conserved order parameters: the grain-growth kinetic. *Phys Rev B 50*(15752).

[32] Cheney, W. (2001). Analysis for applied mathematics, volume 208 of *Graduate Texts in Mathematics*. Springer-Verlag, New York. ISBN 0-387-95279-9. doi: 10.1007/978-1-4757-3559-8.

[33] Cioranescu, D. and Donato, P. (1999). An introduction to homogenization, volume 17 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press, Oxford University Press, New York. ISBN 0-19-856554-2.

[34] Clopeau, T., Ferrín, J. L., Gilbert, R. P., and Mikelić, A. (2001). Homogenizing the acoustic properties of the seabed. II. *Math. Comput. Modelling 33*(8-9), 821–841. doi: 10.1016/S0895-7177(00)00283-1.

[35] Collins, J. B. and Levine, H. (1985). Diffuse interface model of diffusion-limited crystal growth. *Physical Review B 31*(9), 6119.

[36] Coussy, O. (2004). Poromechanics. John Wiley & Sons.

[37] Dawson, C. N., Klíe, H., Wheeler, M. F., and Woodward, C. S. (1997). A parallel, implicit, cell-centered method for two-phase flow with a preconditioned Newton-Krylov solver. *Comput. Geosci. 1*(3-4), 215–249 (1998). doi: 10.1023/A:1011521413158.

[38] Detournay, E. and Cheng, A. H.-D. (1993). Fundamentals of poroelasticity. In *Analysis and design methods*, pages 113–171. Elsevier.

[39] Deuflhard, P. (2011). Newton methods for nonlinear problems, volume 35 of *Springer Series in Computational Mathematics*. Springer, Heidelberg. ISBN 978-3-642-23898-7. doi: 10.1007/978-3-642-23899-4. Affine invariance and adaptive algorithms, First softcover printing of the 2006 corrected printing.

[40] Ferrín, J. L. and Mikelić, A. (2003). Homogenizing the acoustic properties of a porous matrix containing an incompressible inviscid fluid. *Math. Methods Appl. Sci. 26*(10), 831–859. doi: 10.1002/mma.398.

[41] Fix, G. J. (1982). Phase field methods for free boundary problems. *Research Notes in Mathematics 78–79*, 580–589.

[42] Francfort, G. A. and Marigo, J.-J. (1998). Revisiting brittle fracture as an energy minimization problem. *J. Mech. Phys. Solids 46*(8), 1319–1342. doi: 10.1016/S0022-5096(98)00034-9.

[43] Frénod, E. (2012). Two-scale convergence. In *CEMRACS'11: Multiscale coupling of complex models in scientific computing*, volume 38 of *ESAIM Proc.*, pages 1–35. EDP Sci., Les Ulis. doi: 10.1051/proc/201238002.

[44] Gatica, G. N. (2014). A simple introduction to the mixed finite element method. SpringerBriefs in Mathematics. Springer, Cham. ISBN 978-3-319-03694-6; 978-3-319-03695-3. doi: 10.1007/978-3-319-03695-3. Theory and applications.

[45] Gatmiri, B. and Delage, P. (1997). A formulation of fully coupled thermal–hydraulic–mechanical behaviour of saturated porous media—numerical approach. *International Journal for Numerical and Analytical Methods in Geomechanics 21*(3), 199–225.

[46] Gerasimov, T. and De Lorenzis, L. (2016). A line search assisted monolithic approach for phase-field computing of brittle fracture. *Comput. Methods Appl. Mech. Engrg. 312*, 276–303. doi: 10.1016/j.cma.2015.12.017.

[47] Gilbert, R. P. and Mikelić, A. (2000). Homogenizing the acoustic properties of the seabed. I. *Nonlinear Anal. 40*(1-8, Ser. A: Theory Methods), 185–212. doi: 10.1016/S0362-546X(00)85011-7. Lakshmikantham's legacy: a tribute on his 75th birthday.

[48] Griffith, A. A. and Eng, M. (1921). VI. The phenomena of rupture and flow in solids. *Phil. Trans. R. Soc. Lond. A 221*(582-593), 163–198.

[49] Hornung, U. (2012). Homogenization and porous media, volume 6. Springer Science & Business Media.

[50] Iliev, O. P., Kolesov, A. E., and Vabishchevich, P. N. (2016). Numerical solution of plate poroelasticity problems. *Transp. Porous Media 115*(3), 563–580. doi: 10.1007/s11242-016-0726-7.

[51] Kim, J., Tchelepi, H. A., Juanes, R., et al. (2009). Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics. In *SPE reservoir simulation symposium*. Society of Petroleum Engineers.

[52] Kirkesæther Brun, M., Ahmed, E., Berre, I., Nordbotten, J. M., and Radu, F. A. (2019). Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. *arXiv e-prints* arXiv:1902.05783.

[53] Kirkesæther Brun, M., Wick, T., Berre, I., Nordbotten, J. M., and Radu, F. A. (2019). An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters. *arXiv e-prints* arXiv:1903.08717.

[54] Kobayashi, R., Warren, J. A., and Carter, W. C. (2000). A continuum model of grain boundaries. *Phys. D 140*(1-2), 141–150. doi: 10.1016/S0167-2789(00)00023-3.

[55] Kolesov, A. E. and Vabishchevich, P. N. (2017). Splitting schemes with respect to physical processes for double-porosity poroelasticity problems. *Russian J. Numer. Anal. Math. Modelling 32*(2), 99–113. doi: 10.1515/rnam-2017-0009.

[56] Kundu, P. K., Cohen, I. M., and Dowling, D. (2008). Fluid Mechanics 4th. Elsevier.

[57] Landa-Marbán, D., Radu, F. A., and Nordbotten, J. M. (2017). Modeling and simulation of microbial enhanced oil recovery including interfacial area. *Transp. Porous Media 120*(2), 395–413. doi: 10.1007/s11242-017-0929-6.

[58] Langer, J. S. (1986). Models of pattern formation in first-order phase transitions. In *Directions in condensed matter physics*, volume 1 of *World Sci. Ser. Dir. Condensed Matter Phys.*, pages 165–186. World Sci. Publishing, Singapore. doi: 10.1142/9789814415309-0005.

[59] Lee, C. K. and Mei, C. C. (1997). Thermal consolidation in porous media by homogenization theory—I. Derivation of macroscale equations. *Advances in water resources 20*(2-3), 127–144.

[60] Lee, J. J. (2016). Robust error analysis of coupled mixed methods for Biot's consolidation model. *J. Sci. Comput. 69*(2), 610–632. doi: 10.1007/s10915-016-0210-0.

[61] Lee, J. J., Mardal, K.-A., and Winther, R. (2017). Parameter-robust discretization and preconditioning of Biot's consolidation model. *SIAM J. Sci. Comput. 39*(1), A1–A24. doi: 10.1137/15M1029473.

[62] Lee, S., Mikelić, A., Wheeler, M. F., and Wick, T. (2016). Phase-field modeling of proppant-filled fractures in a poroelastic medium. *Comput. Methods Appl. Mech. Engrg. 312*, 509–541. doi: 10.1016/j.cma.2016.02.008.

[63] List, F. and Radu, F. A. (2016). A study on iterative methods for solving Richards' equation. *Comput. Geosci. 20*(2), 341–353. doi: 10.1007/s10596-016-9566-3.

[64] Mesgarnejad, A., Bourdin, B., and Khonsari, M. M. (2015). Validation simulations for the variational approach to fracture. *Comput. Methods Appl. Mech. Engrg. 290*, 420–437. doi: 10.1016/j.cma.2014.10.052.

[65] Miehe, C., Hofacker, M., and Welschinger, F. (2010). A phase field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits. *Comput. Methods Appl. Mech. Engrg. 199*(45-48), 2765–2778. doi: 10.1016/j.cma.2010.04.011.

[66] Mikelić, A. (2003). Recent developments in multiscale problems coming from fluid mechanics. In *Trends in nonlinear analysis*, pages 225–267. Springer, Berlin.

[67] Mikelić, A., Wang, B., and Wheeler, M. F. (2014). Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Comput. Geosci. 18*(3-4), 325–341. doi: 10.1007/s10596-013-9393-8.

[68] Mikelić, A. and Wheeler, M. F. (2012). Theory of the dynamic Biot-Allard equations and their link to the quasi-static Biot system. *J. Math. Phys. 53*(12), 123702, 15. doi: 10.1063/1.4764887.

[69] Mikelić, A. and Wheeler, M. F. (2013). Convergence of iterative coupling for coupled flow and geomechanics. *Comput. Geosci. 17*(3), 455–461. doi: 10.1007/s10596-012-9318-y.

[70] Mikelic, A., Wheeler, M. F., and Wick, T. (2014). Phase-field modeling of pressurized fractures in a poroelastic medium. *ICES Report* pages 14–18.

[71] Mikelić, A., Wheeler, M. F., and Wick, T. (2015). A phase-field method for propagating fluid-filled fractures coupled to a surrounding porous medium. *Multiscale Model. Simul. 13*(1), 367–398. doi: 10.1137/140967118.

[72] Mikelić, A., Wheeler, M. F., and Wick, T. (2015). A quasi-static phase-field approach to pressurized fractures. *Nonlinearity 28*(5), 1371–1399. doi: 10.1088/0951-7715/28/5/1371.

[73] Mikelić, A., Wheeler, M. F., and Wick, T. (2019). Phase-field modeling through iterative splitting of hydraulic fractures in a poroelastic medium. *GEM Int. J. Geomath. 10*(1), Art. 2, 33. doi: 10.1007/s13137-019-0113-y.

[74] Morin, B., Elder, K., Sutton, M., and Grant, M. (1995). Model of the kinetics of polymorphous crystallization. *Phys Rev Lett 75*(2156).

[75] Neitzel, I., Wick, T., and Wollner, W. (2017). An optimal control problem governed by a regularized phase-field fracture propagation model. *SIAM J. Control Optim. 55*(4), 2271–2288. doi: 10.1137/16M1062375.

[76] Nguetseng, G. (1989). A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal. 20*(3), 608–623. doi: 10.1137/0520043.

[77] Nguetseng, G. (1990). Asymptotic analysis for a stiff variational problem arising in mechanics. *SIAM J. Math. Anal. 21*(6), 1394–1414. doi: 10.1137/0521078.

[78] Penrose, O. and Fife, P. C. (1990). Thermodynamically consistent models of phase-field type for the kinetics of phase transitions. *Phys. D 43*(1), 44–62. doi: 10.1016/0167-2789(90)90015-H.

[79] Penrose, O. and Fife, P. C. (1993). On the relation between the standard phase-field model and a "thermodynamically consistent" phase-field model. *Phys. D 69*(1-2), 107–113. doi: 10.1016/0167-2789(93)90183-2.

[80] Phillips, P. J. and Wheeler, M. F. (2008). A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity. *Comput. Geosci. 12*(4), 417–435. doi: 10.1007/s10596-008-9082-1.

[81] Pop, I. S., Radu, F., and Knabner, P. (2004). Mixed finite elements for the Richards' equation: linearization procedure. *J. Comput. Appl. Math. 168*(1-2), 365–373. doi: 10.1016/j.cam.2003.04.008.

[82] Qin, R. and Bhadeshia, H. (2009). Phase field method. *Materials Science and Technology 26*(7).

[83] Radu, F. A., Kumar, K., Nordbotten, J. M., and Pop, I. S. (2018). A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities. *IMA J. Numer. Anal. 38*(2), 884–920. doi: 10.1093/imanum/drx032.

[84] Radu, F. A., Nordbotten, J. M., Pop, I. S., and Kumar, K. (2015). A robust linearization scheme for finite volume based discretizations for simulation

of two-phase flow in porous media. *J. Comput. Appl. Math. 289*, 134–141. doi: 10.1016/j.cam.2015.02.051.

[85] Saad, Y. (2003). Iterative methods for sparse linear systems. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition. ISBN 0-89871-534-2. doi: 10.1137/1.9780898718003.

[86] Sánchez-Palencia, E. (1980). Nonhomogeneous media and vibration theory, volume 127 of *Lecture Notes in Physics*. Springer-Verlag, Berlin-New York. ISBN 3-540-10000-8.

[87] Sauer, T. (2012). Numerical Analysis (2nd).

[88] Showalter, R. E. (2000). Diffusion in poro-elastic media. *J. Math. Anal. Appl. 251*(1), 310–340. doi: 10.1006/jmaa.2000.7048.

[89] Suvorov, A. P. and Selvadurai, A. P. S. (2010). Macroscopic constitutive equations of thermo-poroviscoelasticity derived using eigenstrains. *J. Mech. Phys. Solids 58*(10), 1461–1473. doi: 10.1016/j.jmps.2010.07.016.

[90] Terzaghi, K. (1944). Theoretical soil mechanics. Chapman And Hali, Limited John Wiler And Sons, Inc; New York.

[91] Tran, D., Nghiem, L., Buchanan, L., et al. (2005). An overview of iterative coupling between geomechanical deformation and reservoir flow. In *SPE International Thermal Operations and Heavy Oil Symposium*. Society of Petroleum Engineers.

[92] Trefethen, L. N. and Bau, D., III (1997). Numerical linear algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. ISBN 0-89871-361-7. doi: 10.1137/1.9780898719574.

[93] van Duijn, C. J., Mikelić, A., Wheeler, M. F., and Wick, T. (2019). Thermoporoelasticity via homogenization: Modeling and formal two-scale expansions. *Internat. J. Engrg. Sci. 138*, 1–25. doi: 10.1016/j.ijengsci.2019.02.005.

[94] Wang, H. F. (2017). Theory of linear poroelasticity with applications to geomechanics and hydrogeology. Princeton University Press.

[95] Wick, T. (2017). An Error-Oriented Newton/Inexact Augmented Lagrangian Approach for Fully Monolithic Phase-Field Fracture Propagation. *SIAM Journal on Scientific Computing 39*(4), B589–B617. doi: 10.1137/16M1063873.

[96] Wick, T. (2017). Modified Newton methods for solving fully monolithic phase-field quasi-static brittle fracture propagation. *Comput. Methods Appl. Mech. Engrg. 325*, 577–611. doi: 10.1016/j.cma.2017.07.026.

[97] Yi, S.-Y. (2014). Convergence analysis of a new mixed finite element method for Biot's consolidation model. *Numer. Methods Partial Differential Equations 30*(4), 1189–1210. doi: 10.1002/num.21865.

Part II

Scientific Results

Paper A

Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium

M. K. Brun, I. Berre, J. M. Nordbotten, F. A. Radu

Transport in Porous Media **124(1)** (2018), p. 137–158.

doi: 10.1007/s11242-018-1056-8

Transp Porous Med (2018) 124:137–158
https://doi.org/10.1007/s11242-018-1056-8

Upscaling of the Coupling of Hydromechanical and Thermal Processes in a Quasi-static Poroelastic Medium

Mats K. Brun[1] · **Inga Berre**[1] · **Jan M. Nordbotten**[1] · **Florin A. Radu**[1]

Received: 23 November 2017 / Accepted: 5 April 2018 / Published online: 12 May 2018

Abstract We undertake a formal derivation of a linear poro-thermo-elastic system within the framework of quasi-static deformation. This work is based upon the well-known derivation of the quasi-static poroelastic equations (also known as the Biot consolidation model) by homogenization of the fluid-structure interaction at the microscale. We now include energy, which is coupled to the fluid-structure model by using linear thermoelasticity, with the full system transformed to a Lagrangian coordinate system. The resulting upscaled system is similar to the linear poroelastic equations, but with an added conservation of energy equation, fully coupled to the momentum and mass conservation equations. In the end, we obtain a system of equations on the macroscale accounting for the effects of mechanical deformation, heat transfer, and fluid flow within a fully saturated porous material, wherein the coefficients can be explicitly defined in terms of the microstructure of the material. For the heat transfer we consider two different scaling regimes, one where the Péclet number is small, and another where it is unity. We also establish the symmetry and positivity for the homogenized coefficients.

Keywords Homogenization · Two-scale expansions · Porous media · Thermoelasticity

✉ Mats K. Brun
mats.brun@uib.no

Inga Berre
inga.berre@uib.no

Jan M. Nordbotten
jan.nordbotten@uib.no

Florin A. Radu
florin.radu@uib.no

[1] Department of Mathematics, University of Bergen, Bergen, Norway

1 Introduction

The theory of consolidation of soils goes back to the work of Terzaghi (1944) and Biot (1941, 1972, 1977), and since then numerous authors have contributed to the field, extending the models to different situations and providing more rigorous results for the equations. Today, this field is better known as 'poroelasticity', and is of great importance in a range of different engineering disciplines, such as reservoir engineering and biomechanics. Notable contributions are Burridge and Keller (1981), where a formal upscaling leading to the quasi-static Biot-model was undertaken, and the book Sanchez-Palencia (1980) where a rigorous derivation can be found. In Clopeau et al. (2001) and Gilbert and Mikelić (2000) the rigorous derivation of a dynamic Biot-model corresponding to different choices of scalings of the microstructure is undertaken, and in Ferrin and Mikelić (2003) the case of an inviscid fluid filling the pore space is treated. In Lévy (1979) elastic wave propagation is considered. Additional cases and results can also be found in the references of these works.

The motivation for the present article is to better understand how thermal stresses in the solid structure of a porous medium are influenced by the forces exerted on the pore walls by the fluid. We consider a porous medium on the macroscopic scale such that the continuum hypothesis is valid, and derive the pointwise continuum model by upscaling the fluid-structure interaction at the microscopic scale where the complex geometry is resolved. We shall focus on a natural system, such as the subsurface, where flow velocity, mechanical strain, and temperature changes are small. This also allows for linearization of the constitutive laws of thermoelasticity, as well as linearization of the fluid-structure coupling conditions. Topics such as nonlinear deformation and high flow rates are beyond the scope of this article. Previously, the homogenization of a similar model problem was undertaken by Lee and Mei (1997), but with a different scaling, and with the fine-scale model defined in terms of Eulerian coordinates. This approach leads to relatively strict conditions on the allowable deformations. It also makes a direct comparison of models difficult. In Bringedal et al. (2016) a formal upscaling of non-isothermal reactive flow in porous media was undertaken, but the solid matrix was assumed rigid. In Eden and Muntean (2017) homogenization of a fully coupled thermoelasticity problem was undertaken, but not in the context of fluid-structure interaction. In the book Coussy (1995) there is also a section on linear thermo-poroelasticity, where the macroscale equations are derived using principles from continuum mechanics and thermodynamics. While finalizing this work, we have been made aware that a similar derivation has been undertaken simultaneously by the authors van Duijn et al. Their work is currently under review and exists as a preprint (Van Duijn et al. 2017).

Our microscale model consists of a fluid-structure interaction model, and energy conservation for both phases (the solid and fluid), where we scale the fluid-structure equations corresponding to the biphasic macroscopic behavior of the system (i.e., fluid pressure in balance with the normal forces coming from the solid matrix, and small viscous forces in the fluid). A rigorous study of this situation in the isothermal case can be found in Clopeau et al. (2001). Different scalings are of course possible, and for different values of the reference quantities, the homogenization process may result in vastly different macroscale models. A discussion around the characterization of the behavior of porous media according to the values of such reference quantities can be found in Auriault (1991). Regarding the energy conservation, we consider two different scaling regimes; one corresponding to a Péclet number of order one, giving a (nonlinear) convective term in the upscaled energy conservation equation, and one corresponding to a small Péclet number, resulting in no convective term,

giving a fully linear upscaled system. Depending on the flow rate and the thermal conductive properties of the fluid, both may be relevant.

The upscaling procedure is done via the formal two-scale asymptotic expansion method of homogenization. This is a well-known technique for qualitatively assessing the structure of the upscaled equations. For a detailed explanation of this method we refer to the books Hornung (2012) and Cioranescu and Donato (2000). For an accurate physical model, the values of the homogenized coefficients should be confirmed by experiments, as the asymptotic expansion method only provides formulas for these in the case of simple microscale geometries. Our justification for the upscaled model comes from the similarity with the isothermal poroelastic equations, and the analogy to the thermoelasticity equations in mechanics.

2 The Pore-Scale Model

2.1 Notation

A short remark on the notation used in this article is in order. We denote by : the scalar product of two second-order tensors, i.e., $\mathbf{A} : \mathbf{B} = \sum_{i,j=1}^{3} A_{ij} B_{ij}$, and by $\otimes$ the vector outer product, which given two vectors produce a second-order tensor, i.e., $(u \otimes v)_{ij} = u_i v_j$. Note also that we shall reserve the use of bold fonts for tensors of second order or more.

2.2 Presentation of the Equations

In this and the next section we present the governing equations to be used throughout the rest of this article. This will include a brief discussion of the constitutive relations of linear thermoelasticity for an anisotropic solid, relevant for the present work. For a detailed derivation of the equations of linear thermoelasticity, we refer to the book Silhavy (2013). We also mention Pabst (2005) where a more compact presentation is given.

Our physical domain is $\Omega = (0, L)^3$, which consists of a solid skeleton, Ω_s, and a fluid filled void space, Ω_f, where the internal boundary between the solid and void parts is denoted by Γ, i.e., in the reference configuration we have: $\Omega = \Omega_s \cup \Omega_f \cup \Gamma$ where $\Omega_s \cap \Omega_f = \varnothing$, and $\Gamma = \partial\Omega_s \cap \partial\Omega_f$. We let $J = (0, T_{\text{end}}]$ be the time interval, where $T_{\text{end}} > 0$ is the final time. We denote by $x = (x_1, x_2, x_3)$ the coordinates of the reference configuration, and by t the time coordinate.

We let w be the displacement vector of the solid, defined on the reference configuration, and assume it can be decomposed as $w(x,t) = \hat{w}(x,t) + w_0(t)$, where $\hat{w}$ corresponds to the local deformation, and w_0 corresponds to a rigid body motion. We let the v be the flow velocity of the fluid, defined on the current configuration, which we then can write as $v(x + \hat{w}, t) = \hat{v}(x + \hat{w}, t) + \dfrac{\mathrm{d}w_0}{\mathrm{d}t}(t)$, using the reference coordinates.

Given a body force b, the linear momentum balance for an elastic solid is given by

$$\rho_s \frac{\partial^2 w}{\partial t^2} - \nabla \cdot \boldsymbol{\sigma} = b \quad \text{in} \quad \Omega_s \times J, \tag{1}$$

where ρ_s is the solid density. In the non-isothermal case, the constitutive equation for the stress is

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{F}, T_s), \tag{2}$$

where T_s is the temperature distribution of the solid, and $\mathbf{F}$ is the deformation gradient.

Denoting by c_s the specific heat capacity of the solid, the conservation of energy is given by

$$\rho_s c_s \frac{\partial T_s}{\partial t} = \boldsymbol{\sigma} : \mathbf{e}(\partial_t w) - \nabla \cdot h_s, \tag{3}$$

where h_s is the heat flux within the solid.

In the pore space, the flow is governed by the Navier–Stokes equations

$$\rho_f \left(\frac{\partial v}{\partial t} + v \cdot \nabla v \right) - \nabla p + \mu \Delta v = b, \quad \text{in} \quad \Omega_f(t) \times J, \tag{4}$$

with the mass conservation

$$\nabla \cdot v = 0, \quad \text{in} \quad \Omega_f(t) \times J, \tag{5}$$

where ρ_f is the fluid density, p is the fluid pressure, and μ is the fluid viscosity.

Since there is no heat generation from dissipative effects in the fluid, we use a simple convection-diffusion equation for the energy conservation

$$\rho_f c_f \left(\frac{\partial T_f}{\partial t} + v \cdot \nabla T_f \right) - \nabla \cdot h_f = 0, \quad \text{in} \quad \Omega_f(t) \times J, \tag{6}$$

where T_f is the temperature distribution of the fluid, c_f is the specific heat capacity of the fluid, and h_f is the heat flux within the fluid.

We now turn to the fluid-structure coupling conditions at the internal interface and denote by ν the outward unit normal field of Ω_f (i.e., pointing into the solid).

By Newton's third law we must have a balance of normal forces coming from both sides

$$(p\mathbf{I} + 2\mu\, \mathbf{e}(v))\,|_{x+\hat{w}}\, \nu = \boldsymbol{\sigma}\, \nu, \quad \text{on} \quad \Gamma \times J, \tag{7}$$

where $\mathbf{I}$ denotes the 3×3 identity tensor.

The no-flow condition at the internal interface now takes the form

$$v|_{x+\hat{w}} = \partial_t w, \quad \text{on} \quad \Gamma \times J. \tag{8}$$

Finally, continuity of heat flux and continuity of temperature at the internal interface gives (due to our assumption of the two phases being in local thermodynamic equilibrium)

$$h_f|_{x+\hat{w}} \cdot \nu = h_s \cdot \nu, \quad \text{on} \quad \Gamma \times J, \tag{9}$$

and

$$T_f|_{x+\hat{w}} = T_s, \quad \text{on} \quad \Gamma \times J. \tag{10}$$

2.3 Constitutive Equations

We let $(\mathbf{F}_0, \theta_0)$ denote the reference values of the deformation gradient and the temperature of the medium (considered here to be uniform, i.e., constant), and assume that for all $t \in J$ the deviations from this reference state are small. Within this framework, a physical linearization of the constitutive equations is justified (see Pabst (2005) for more details). The deformation gradient $\mathbf{F}$, however, is still a nonlinear measure of deformation; hence, we assume in addition that the displacement gradients, or more precisely the local deformations, $\hat{w}$, are small such that $\mathbf{F}$ can be considered approximately identity. This amounts to a geometric linearization of the kinematic measures, and consequently, the first Piola–Kirchoff stress and the Cauchy stress tensors coincide. The strains in the solid are therefore given by the symmetric gradient

of the displacements, i.e., $\mathbf{e}(w) = \frac{1}{2}(\nabla w + (\nabla w)^T)$. The constitutive equation for the stress, Eq. (2) then takes the form

$$\boldsymbol{\sigma}(w, T_s) = \mathbf{C}\,\mathbf{e}(w) - \mathbf{M}(T_s - \theta_0), \tag{11}$$

which is a generalized Hooke's law, extended to include thermal effects. The stiffness tensor of the material (or more precisely, the referential tensor of isothermal elasticities) is given by $\mathbf{C} = (C_{ijkl})_{i,j,k,l=1}^3$, which satisfies $C_{ijkl} = C_{klij} = C_{jikl} = C_{ijlk}$, and the thermal stress tensor (or the referential coefficient of thermal stress) is $\mathbf{M} = (M_{ij})_{i,j=1}^3$, satisfying $M_{ij} = M_{ji}$. In order to have symmetric positive definite coefficients in the upscaled problem, the same must be true for $\mathbf{C}$ and $\mathbf{M}$, i.e.,

$$\mathbf{C}\,\mathbf{e} : \mathbf{e} > 0, \forall\, \mathbf{e} \in \mathbb{R}^{3\times 3} \setminus \{0\}, \text{ and } \mathbf{M}\,\mathbf{x} \cdot \mathbf{x} > 0, \forall\, \mathbf{x} \in \mathbb{R}^3 \setminus \{0\}. \tag{12}$$

Further, we assume the heat fluxes within the solid and the fluid obey Fourier's law of heat conduction, i.e.,

$$h_s = -\,\mathbf{K}_s\,\nabla T_s \quad \text{and} \quad h_f = -\,\mathbf{K}_f\,\nabla T_f, \tag{13}$$

where $\mathbf{K}_s = (K_{ij}^s)_{i,j=1}^3$ and $\mathbf{K}_f = (K_{ij}^f)_{i,j=1}^3$ are the thermal conductivity tensors of the solid and fluid, respectively, which are assumed to be both symmetric and positive definite, i.e., $K_{ij}^{s,f} = K_{ji}^{s,f}$, and $\mathbf{K}_{s,f}\,\mathbf{x} \cdot \mathbf{x} > 0, \forall\, \mathbf{x} \in \mathbb{R}^3 \setminus \{0\}$.

Within this completely linearized framework (physically and geometrically), the densities, ρ_s and ρ_f, are constants taking the values of the reference densities. We assume in addition that the fluid viscosity, μ, is a constant.

We note that in a consistent linear theory, the material coefficients cannot depend on the current temperature, but only on the (uniform) reference temperature, θ_0.

2.4 The Domain

Before undertaking the scaling analysis, we provide a more detailed description of the domain, specifically that it is made up of a periodic repetition of a single pore, such that the geometry of the whole solid skeleton is determined by the geometry inside a singe microscopic cell. This is a valid assumption since we are modeling a fine grained porous media with microstructure on a scale much smaller than the continuum scale of interest. Thus, although the material is heterogenous on the microscale, it appears locally homogenous on the macroscale. We follow Allaire (1989) in this description.

Let l be a typical pore size, and let L be the size of the macroscale domain, and define as usual $\varepsilon = l/L$. We let $\Omega^\varepsilon = \frac{1}{L}\Omega$ be the dimensionless domain, which now is $\Omega^\varepsilon = (0, 1)^3$, such that Ω_s^ε and Ω_f^ε are the corresponding dimensionless solid and void parts, respectively, and Γ^ε is the corresponding dimensionless internal interface. We continue with the notation J for the time interval; keeping in mind time is now also dimensionless.

Let $Y = (0, 1)^3$ be the rescaled unit cube in $\mathbb{R}^3$, consisting of a solid part, $Y_s \subset \bar{Y}$, which is a closed subset of strictly positive measure, and a void space, $Y_f = Y \setminus Y_s$, which is an open and connected subset of strictly positive measure. We let now $\Gamma = \partial Y_s \cap \partial Y_f$ denote the internal interface of the unit pore cell, and assume the configuration is such that Γ is a smooth surface. We make a periodic repetition of Y_s over $\mathbb{R}^3$, and set $Y_{s,k}^\varepsilon = \varepsilon(Y_s + k)$, where $k \in \mathbb{Z}^3$. Let $K = \{k \in \mathbb{Z}^3 : Y_{s,k}^\varepsilon \subset \bar{\Omega}^\varepsilon\}$, such that $\bar{\Omega}_s^\varepsilon = \bigcup_{k\in K} Y_{s,k}^\varepsilon$ is the solid skeleton, and $\Omega_f^\varepsilon = \Omega^\varepsilon \setminus \Omega_s^\varepsilon$ is the fluid filled void space. The fluid/solid internal interface can now be written $\Gamma^\varepsilon = \partial\Omega_s^\varepsilon \setminus \partial\Omega^\varepsilon$. By construction, both Ω_s^ε and Ω_f^ε are now connected sets of strictly positive measure, and Γ^ε is a smooth surface.

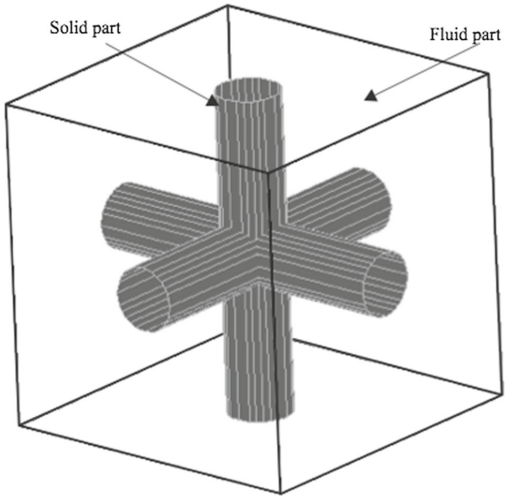


Fig. 1 Example of geometry inside unit cell. (Picture from Mikelić and Wheeler 2012)

The epsilon-superscript on Ω^ε implies the implicit dependence of the domain on both length scales, l and L, but later when we impose the homogenization ansatz, we separate these two scales, and let the size of the domain become arbitrarily large (that is, we let $\varepsilon \to 0$). Then, behind each infinitesimal point x (seen from the macro scale) there is a pore cell with its own geometry which can only be seen by the fast variable y. When this scale separation is done, we shall denote the macro scale domain simply by Ω, which is no longer possible to separate into solid and void parts because the porous structure is now seen as a single (fictitious) uniform material. An example of a pore cell geometry which satisfies the above assumptions is shown in Fig. 1.

2.5 Scaling Analysis

In this section, we introduce dimensionless variables and scale the system according to the quasi-static biphasic macroscale behavior (see Auriault (1991) for more details). In short, this means the fluid pressure should be of the same order as the normal stress coming from the elastic matrix and that the viscous forces are small.

The modeling of fluid and elastic solid structure interaction is in general challenging, since for an elastic solid Lagrangian coordinates are the preferred reference frame, while for the fluid it is the Eulerian one. Thus, when coupling the two processes at the mutual interface, one needs to take into account the movement of the interface itself, as seen in Sect. 2.2. For more details on this type of modeling we refer to Iliev et al. (2008). In the present work we shall avoid this difficulty by linearizing the fluid-structure coupling conditions, and thereby transform the fluid problem into Lagrangian coordinates based on the material deformation.

We let $l = 10^{-5}\,\mathrm{m}$ and $L = 10\,\mathrm{m}$, which results in $\varepsilon = l/L = 10^{-6}$. With a slight abuse of notation we write now the dimensional variables with a tilde, and the new dimensionless quantities in the same way as before

Table 1 Reference values. *Source* https://www.engineeringtoolbox.com

Quantity	Value	Unit
Material stiffness (Young's modulus)	$C_{\text{ref}} = 4 \times 10^{10}$	N/m^2
Fluid density	$\rho_f = 10^3$	kg/m^3
Solid grain density	$\rho_s = 2.65 \times 10^3$	kg/m^3
Fluid viscosity	$\mu_{\text{ref}} = 10^{-3}$	Pa·s
Thermal stress coefficient	$M_{\text{ref}} = 4 \times 10^5$	$\text{N/m}^2 K$
Solid grain specific heat	$c_s = 920$	J/kgK
Fluid specific heat	$c_f = 4182$	J/kgK
Thermal conductivity solid grain	$K^s_{\text{ref}} = 1.7$	W/mK
Thermal conductivity fluid	$K^f_{\text{ref}} = 0.58$	W/mK

$$\begin{array}{llll} \tilde{x} = Lx, & \tilde{t} = \tau t, & \tilde{\hat{w}} = l\hat{w}, & \tilde{w}_0 = W_{\text{ref}} w_0, \quad \tilde{\hat{v}} = V_{\text{ref}} v, \\ \tilde{T}_s = T_{\text{diff}} T_s, & \tilde{T}_f = T_{\text{diff}} T_f, & \tilde{p} = P_{\text{ref}} p, & \tilde{\mu} = \mu_{\text{ref}} \mu, \\ \tilde{\mathbf{C}} = C_{\text{ref}} \mathbf{C}, & \tilde{\mathbf{M}} = M_{\text{ref}} \mathbf{M}, & \tilde{\mathbf{K}}_f = K^f_{\text{ref}} \mathbf{K}_f, & \tilde{\mathbf{K}}_s = K^s_{\text{ref}} \mathbf{K}_s, \end{array}$$

where we choose the time scale, $\tau = \frac{L}{V_{\text{ref}}}$, as the characteristic transport time. As we consider a system in equilibrium with only natural convection, we set $\tau = 10^4$ s, which is the time it takes a fluid particle to traverse the distance L. This gives the reference velocity as $V_{\text{ref}} = 10^{-3}$ m/s. This is a realistic value, as flow velocities coming from natural convection in a geological permeable layer can be as low as 1 m/year Wood and Hewett (1982). We let the size of the rigid displacements be $W_{\text{ref}} = 1$ m, while the local deformation is no larger than the pore size. We set the characteristic temperature as the maximum difference between the reference temperature and the current temperature as $T_{\text{diff}} = 10$ K. Linearizing the fluid/solid coupling conditions then amounts to discarding terms of $\hat{w} \cdot \nabla v$ and $\hat{w} \cdot \nabla T_f$ and higher order ones (this can be seen by expanding the fluid side about $x + \hat{w} = x$). Thus, since the spatial differential operators are acting with respect to the scale L, we are introducing errors of the order $\varepsilon V_{\text{ref}} = 10^{-9}$ m/s and $\varepsilon T_{\text{diff}} = 10^{-5}$ K, which we deem negligible.

The table below shows reference values for the coefficients, in the case of water and sandstone at room temperature (we note that some are only approximate, but good enough for our purposes, as we are only interested in identifying terms which differ by an order of ε or more) (Table 1).

The balance of contact forces at the interface, Eq. (7), in dimensionless variables reads

$$\left(-P_{\text{ref}} p \mathbf{I} + \frac{V_{\text{ref}} \mu_{\text{ref}}}{L} 2\mu\, \mathbf{e}(v)\right) \nu = (C_{\text{ref}} \varepsilon\, \mathbf{C}\, \mathbf{e}(w) - M_{\text{ref}} T_{\text{diff}}\, \mathbf{M}\, T_s) \nu, \text{ on } \Gamma^\varepsilon \times J. \quad (14)$$

Due to the biphasic scaling regime, we have $P_{\text{ref}} \sim C_{\text{ref}} \varepsilon \sim M_{\text{ref}} T_{\text{diff}} \sim 10^5 N/m^2$. Dividing by this factor, we get $\frac{V_{\text{ref}} \mu_{\text{ref}}}{L P_{\text{ref}}} \sim \mathcal{O}(\varepsilon^2)$ in front of the viscous term. Thus, we can simplify the above equation and get the dimensionless form of Eq. (7) as

$$\left(p\mathbf{I} + 2\mu\varepsilon^2\, \mathbf{e}(v)\right) \nu = (\mathbf{C}\, \mathbf{e}(w) - \mathbf{M}\, T_s)) \nu, \quad \text{on} \quad \Gamma^\varepsilon \times J. \quad (15)$$

The momentum equation for the solid (1) in dimensionless form is

$$\frac{\rho_s}{\tau^2}\left(l\frac{\partial^2\hat{w}}{\partial t^2}+W_{\text{ref}}\frac{\mathrm{d}^2 w_0}{\mathrm{d}t^2}\right)-\frac{1}{L}\nabla\cdot(C_{\text{ref}}\varepsilon\,\mathbf{C}\,\mathbf{e}(w)-M_{\text{ref}}T_{\text{diff}}\,\mathbf{M}\,T_s)=b,\ \text{on}\ \Gamma^\varepsilon\times J. \tag{16}$$

Multiplying with $\frac{L}{P_{\text{ref}}}$, we get a new dimensionless body force (still denoted b) and the dimensionless constants $\frac{\rho_s l L}{\tau^2 P_{\text{ref}}}\sim\mathcal{O}(\varepsilon^2)$ and $\frac{\rho_s W_{\text{ref}} L}{\tau^2 P_{\text{ref}}}\sim\mathcal{O}(\varepsilon)$ multiplying the acceleration terms. Thus, we discard these and get the dimensionless form of Eq. (1) as

$$-\nabla\cdot(\mathbf{C}\,\mathbf{e}(w)-\mathbf{M}\,T_s)=b,\quad \text{on}\quad \Gamma^\varepsilon\times J. \tag{17}$$

The dimensionless form of the momentum equation for the fluid, (4), is

$$\begin{aligned}\rho_f\frac{L}{\tau^2}\left(\frac{\partial\hat{v}}{\partial t}+\hat{v}\cdot\nabla\hat{v}\right)+\rho_f\frac{W_{\text{ref}}}{\tau^2}\left(\frac{\mathrm{d}^2 w_0}{\mathrm{d}t^2}+\frac{\mathrm{d}w_0}{\mathrm{d}t}\cdot\nabla\hat{v}\right)\\ -\frac{P_{\text{ref}}}{L}\nabla p+\frac{V_{\text{ref}}\mu_{\text{ref}}}{L^2}\mu\Delta v=b,\quad \text{in}\quad \Omega_f^\varepsilon\times J.\end{aligned} \tag{18}$$

Multiplying by $\frac{L}{P_{\text{ref}}}$, we again get a new dimensionless body force (still denoted by b) and the dimensionless constants $\frac{\rho_f L^2}{\tau^2 P_{\text{ref}}}\sim\mathcal{O}(\varepsilon)$ and $\frac{\rho_f W_{\text{ref}} L}{\tau^2 P_{\text{ref}}}\sim\mathcal{O}(\varepsilon)$ multiplying the inertial terms. Multiplying the viscous term we have $\frac{V_{\text{ref}}\mu_{\text{ref}}}{L P_{\text{ref}}}\sim\mathcal{O}(\varepsilon^2)$. Thus, we discard the inertial terms and get the dimensionless form of Eq. (4) as

$$-\nabla\cdot(p\mathbf{I}-2\mu\varepsilon^2\,\mathbf{e}(v))=b,\quad \text{in}\quad \Omega_f^\varepsilon\times J. \tag{19}$$

The mass conservation equation for the fluid, Eq. (5) and the no-flow condition at the boundary, Eq. (8), in dimensionless variables are

$$\nabla\cdot v=0,\quad \text{in}\quad \Omega_f^\varepsilon\times J, \tag{20}$$

and

$$v=\partial_t w,\quad \text{on}\quad \Gamma^\varepsilon\times J, \tag{21}$$

respectively.

We now turn to the energy conservation equations. The one for the fluid, Eq. (6), in dimensionless variables reads as

$$\rho_f c_f\left(\frac{1}{\tau_D}\frac{\partial T_f}{\partial t}+\frac{1}{L}\left(V_{\text{ref}}\hat{v}+\frac{W_{\text{ref}}}{\tau}\frac{\mathrm{d}w_0}{\mathrm{d}t}\right)\cdot\nabla T_f\right)-\frac{K^f_{\text{ref}}}{L^2}\nabla\cdot(\mathbf{K}_f\,\nabla T_f)=0,\ \text{in}\ \Omega_f^\varepsilon\times J, \tag{22}$$

where $\tau_D=\tau/\varepsilon$ is the characteristic heat diffusion time. Multiplying by $\frac{L^2}{K^f_{\text{ref}}}$, we get the Péclet number in front of the convective term, which in this case is given by $Pe=\frac{c_f\rho_f V_{\text{ref}} L}{K^f_{\text{ref}}}\sim\frac{c_f\rho_f W_{\text{ref}} L}{\tau K^f_{\text{ref}}}\sim\mathcal{O}(1)$, meaning heat is transported within the fluid by convection and diffusion at an approximately equal rate. In the following, we shall also look at the case $Pe=\mathcal{O}(\varepsilon)$, such that heat is mainly transported within the fluid through diffusion. This could be realized, e.g., in a system with a lower flow velocity or a fluid with a higher thermal conductivity. However, when we undertake the upscaling procedure, it will become clear that this choice can be seen just as a special case of the more general $Pe\sim\mathcal{O}(1)$. In the concluding section we will however present the homogenized model corresponding to both choices of scaling.

In front of the time derivative term, we get the dimensionless constant $\frac{\rho_f c_f L^2}{\tau_D K^f_{\text{ref}}} \sim \mathcal{O}(\varepsilon)$. Thus, we discard this term and get the dimensionless form of Eq. (6) as

$$v \cdot \nabla T_f - \nabla \cdot (\mathbf{K}_f \nabla T_f) = 0, \quad \text{in} \quad \Omega_f^\varepsilon \times J. \tag{23}$$

In the dissipative term of the energy conservation equation for the solid, Eq. (3), we neglect the contribution from the mechanical stress (which is second order in the gradient of w), and assume the heat generation from the thermal stress can be approximated by using a constant value for the temperature difference, i.e., $-\mathbf{M}(T_s - \theta_0) : \mathbf{e}(\partial_t w) \approx -T_{\text{diff}}\mathbf{M} : \mathbf{e}(\partial_t w)$ (see, e.g., Kupradze et al. 1979). Thus, we get Eq. (3) in dimensionless variables as

$$\rho_s c_s \frac{1}{\tau_D}\frac{\partial T_s}{\partial t} + \frac{M_{\text{ref}} l}{\tau L}\mathbf{M} : \mathbf{e}(\partial_t w) - \frac{K^s_{\text{ref}}}{L^2}\nabla \cdot (\mathbf{K}_s \nabla T_s) = 0, \quad \text{in } \Omega_s^\varepsilon \times J. \tag{24}$$

We multiply by $\frac{L^2}{K^s_{\text{ref}}}$, and get the dimensionless constants, $\frac{\rho_s c_s L^2}{\tau_D K^s_{\text{ref}}} \sim \mathcal{O}(\varepsilon)$, multiplying the time derivative term, and $\frac{M_{\text{ref}} l L}{\tau K^s_{\text{ref}}} \sim \mathcal{O}(1)$, multiplying the dissipative term. Thus, we discard the time derivative term and can write Eq. (24) as

$$\mathbf{M} : \mathbf{e}(\partial_t w) - \nabla \cdot (\mathbf{K}_s \nabla T_s) = 0, \quad \text{in} \quad \Omega_s^\varepsilon \times J. \tag{25}$$

The reference values of the thermal conductivities of the two phases can be regarded as approximately the same order (i.e., $K^f_{\text{ref}} \sim K^s_{\text{ref}}$), and we therefore write the dimensionless form of Eqs. (9) and (10) as

$$\mathbf{K}_f \nabla T_f \cdot \nu = \mathbf{K}_s \nabla T_s \cdot \nu, \quad \text{on} \quad \Gamma^\varepsilon \times J, \tag{26}$$

and

$$T_f = T_s, \quad \text{on} \quad \Gamma^\varepsilon \times J, \tag{27}$$

respectively.

2.6 The Complete Dimensionless Pore-Scale Model

For convenience, we summarize the dimensionless equations at the microscale below:

$$-\nabla \cdot (\mathbf{C}\,\mathbf{e}(w^\varepsilon) - \mathbf{M}\,T_s^\varepsilon) = b, \quad \text{in} \quad \Omega_s^\varepsilon \times J, \tag{28a}$$

$$-\nabla \cdot (p^\varepsilon \mathbf{I} - 2\mu\varepsilon^2\,\mathbf{e}(v^\varepsilon)) = b, \quad \text{in} \quad \Omega_f^\varepsilon \times J, \tag{28b}$$

$$\nabla \cdot v^\varepsilon = 0, \quad \text{in} \quad \Omega_f^\varepsilon \times J, \tag{28c}$$

$$(-p^\varepsilon \mathbf{I} + 2\mu\varepsilon^2\,\mathbf{e}(v^\varepsilon))\nu = (\mathbf{C}\,\mathbf{e}(w^\varepsilon) - \mathbf{M}\,T_s^\varepsilon)\nu, \quad \text{on} \quad \Gamma^\varepsilon \times J, \tag{28d}$$

$$v^\varepsilon = \partial_t w^\varepsilon, \quad \text{on} \quad \Gamma^\varepsilon \times J, \tag{28e}$$

$$\mathbf{M} : \mathbf{e}(\partial_t w^\varepsilon) - \nabla \cdot (\mathbf{K}_s \nabla T_s^\varepsilon) = 0, \quad \text{in} \quad \Omega_s^\varepsilon \times J, \tag{28f}$$

$$v^\varepsilon \cdot \nabla T_f^\varepsilon - \nabla \cdot (\mathbf{K}_f \nabla T_f^\varepsilon) = 0, \quad \text{in} \quad \Omega_f^\varepsilon \times J, \tag{28g}$$

$$\mathbf{K}_f \nabla T_f^\varepsilon \cdot \nu = \mathbf{K}_s \nabla T_s^\varepsilon \cdot \nu, \quad \text{on} \quad \Gamma^\varepsilon \times J, \tag{28h}$$

$$T_s^\varepsilon = T_f^\varepsilon, \quad \text{on} \quad \Gamma^\varepsilon \times J. \tag{28i}$$

We impose periodic boundary conditions on the outer boundary (i.e., $\partial\Omega^\varepsilon$) and omit initial conditions since they are not important for the homogenization procedure. Note also that we have included an epsilon-superscript on the dependent variables to emphasize the implicit dependence on both the slow and fast scales.

Finally, we mention that the upscaling of Eqs. (28a) (without the thermal stress term), (28b), (28c), (28d), and (28e), leads to the quasi-static poroelastic equations as described in, e.g., Biot (1941) and Coussy (1995).

3 Two-Scale Asymptotic Expansions

3.1 Homogenization Ansatz

We now undertake the separation of scales and introduce the homogenization ansatz for the unknowns

$$
\begin{aligned}
v^\varepsilon(x,t) &= v^0(x,y,t) + \varepsilon v^1(x,y,t) + \varepsilon^2 v^2(x,y,t) + \cdots,\\
w^\varepsilon(x,t) &= w^0(x,y,t) + \varepsilon w^1(x,y,t) + \varepsilon^2 w^2(x,y,t) + \cdots,\\
T_f^\varepsilon(x,t) &= T_f^0(x,y,t) + \varepsilon T_f^1(x,y,t) + \varepsilon^2 T_f^2(x,y,t) + \cdots,\\
T_s^\varepsilon(x,t) &= T_s^0(x,y,t) + \varepsilon T_s^1(x,y,t) + \varepsilon^2 T_s^2(x,y,t) + \cdots,\\
p^\varepsilon(x,t) &= p^0(x,y,t) + \varepsilon p^1(x,y,t) + \varepsilon^2 p^2(x,y,t) + \cdots.
\end{aligned}
$$

Note that we now have an added dependence on the spatial variable $y \in Y$, in which all terms of the above expansions are Y-periodic due to the scaling and the periodic arrangement of the porous structure. This is the key step in the two-scale asymptotic expansion method of homogenization. For a detailed review of this method and its applications to porous media, we refer to the books Hornung (2012) and Cioranescu and Donato (2000).

Since $y = x/\varepsilon$, we reformulate the differential operators according to the chain rule, i.e., $\nabla = \nabla_x + \varepsilon^{-1}\nabla_y$, and $\mathbf{e}(\cdot) = \mathbf{e}_x(\cdot) + \varepsilon^{-1}\mathbf{e}_y(\cdot)$.

We now insert the asymptotic expansions into the governing equations and discard all therms of $\mathcal{O}(\varepsilon)$ or higher. We furthermore assume that the governing equations of the last section are applicable in the product domain. We start with Eq. (28a) for the elastic solid structure

$$
\begin{aligned}
b = &-\varepsilon^{-2}\nabla_y\cdot(\mathbf{C}\,\mathbf{e}_y(w^0))\\
&-\varepsilon^{-1}\left[\nabla_y\cdot\left(\mathbf{C}\left(\mathbf{e}_x(w^0)+\mathbf{e}_y(w^1)\right)-\mathbf{M}\,T_s^0\right)+\nabla_x\cdot\left(\mathbf{C}\,\mathbf{e}_y(w^0)\right)\right]\\
&-\varepsilon^{0}\left[\nabla_y\cdot\left(\mathbf{C}\left(\mathbf{e}_x(w^1)+\mathbf{e}_y(w^2)\right)-\mathbf{M}\,T_s^1\right)\right.\\
&\left.+\nabla_x\cdot\left(\mathbf{C}\left(\mathbf{e}_x(w^0)+\mathbf{e}_y(w^1)\right)-\mathbf{M}\,T_s^0\right)\right]\\
&+\mathcal{O}(\varepsilon), \quad \text{in} \quad \Omega\times Y_s\times J,
\end{aligned}
\tag{29}
$$

The conservation of momentum and mass for the fluid, Eqs. (28b) and (28c), yields

$$
b = \varepsilon^{-1}\nabla_y p^0 + \varepsilon^0\left[\nabla_x p^0 + \nabla_y p^1 - \mu\Delta_y v^0\right] + \mathcal{O}(\varepsilon), \quad \text{in} \quad \Omega\times Y_f\times J, \tag{30}
$$

and

$$
0 = \varepsilon^{-1}\nabla_y\cdot v^0 + \varepsilon^0\left[\nabla_x\cdot v^0 + \nabla_y\cdot v^1\right] + \mathcal{O}(\varepsilon), \quad \text{in} \quad \Omega\times Y_f\times J. \tag{31}
$$

At the internal interface, continuity of contact forces and continuity of displacement velocity, Eqs. (28d) and (28e), gives

$$
\begin{aligned}
0 = &\,\varepsilon^{-1}\,\mathbf{C}\,\mathbf{e}_y(w^0)\nu\\
&+\varepsilon^0\left[\mathbf{C}(\mathbf{e}_x(w^0)+\mathbf{e}_y(w^1))-\mathbf{M}\,T_s^0+p^0\mathbf{I}\right]\nu
\end{aligned}
$$

$$+\varepsilon^1\left[\mathbf{C}(\mathbf{e}_x(w^1)+\mathbf{e}_y(w^2))-\mathbf{M}\,T_s^1+p^1\mathbf{I}-2\,\mathbf{e}_y(v^0)\right]\nu$$
$$+\mathcal{O}(\varepsilon^2),\quad \text{on}\quad \Omega\times\Gamma\times J, \tag{32}$$

and

$$0=(v^0-\partial_t w^0)+\varepsilon(v^1-\partial_t w^1)+\mathcal{O}(\varepsilon^2)\quad \text{on}\quad \Omega\times\Gamma\times J. \tag{33}$$

The energy conservation equations for the solid and fluid, Eqs. (28f) and (28g), yields

$$\begin{aligned}0=&-\varepsilon^{-2}\nabla_y\cdot(\mathbf{K}_s\,\nabla_y T_s^0)\\&+\varepsilon^{-1}\left[\mathbf{M}:\mathbf{e}_y(\partial_t w^0)-\nabla_y\cdot\left(\mathbf{K}_s(\nabla_x T_s^0+\nabla_y T_s^1)\right)-\nabla_x\cdot\left(\mathbf{K}_s\,\nabla_y T_s^0\right)\right]\\&+\varepsilon^0\left[\mathbf{M}:\left(\mathbf{e}_x(\partial_t w^0)+\mathbf{e}_y(\partial_t w^1)\right)-\nabla_x\cdot\left(\mathbf{K}_s(\nabla_x T_s^0+\nabla_y T_s^1)\right)\right.\\&\left.-\nabla_y\cdot\left(\mathbf{K}_s(\nabla_x T_s^1+\nabla_y T_s^2)\right)\right]+\mathcal{O}(\varepsilon),\quad \text{in}\quad \Omega\times Y_s\times J,\end{aligned} \tag{34}$$

and

$$\begin{aligned}0=&-\varepsilon^{-2}\nabla_y\cdot(\mathbf{K}_f\,\nabla_y T_f^0)\\&+\varepsilon^{-1}\left[v^0\cdot\nabla_y T_f^0-\nabla_y\cdot\left(\mathbf{K}_f(\nabla_x T_f^0+\nabla_y T_f^1)\right)-\nabla_x\cdot(\mathbf{K}_f\,\nabla_y T_f^0)\right]\\&+\varepsilon^0\left[\partial_t u^0\cdot\nabla_x T_f^0-\nabla_x\cdot\left(\mathbf{K}_f(\nabla_x T_f^0+\nabla_y T_f^1)\right)+v^0\cdot\nabla_y T_f^1\right.\\&\left.+v^1\cdot\nabla_y T_f^0-\nabla_y\cdot\left(\mathbf{K}_f(\nabla_x T_f^1+\nabla_y T_f^2)\right)\right]+\mathcal{O}(\varepsilon),\quad \text{in}\quad \Omega\times Y_s\times J.\end{aligned} \tag{35}$$

At the internal interface, continuity of energy and temperature, Eqs. (28h) and (28i), gives

$$\begin{aligned}0=&\varepsilon^{-1}\left[\mathbf{K}_f\,\nabla_y T_f^0-\mathbf{K}_s\,\nabla_y T_s^0\right]\cdot\nu\\&+\varepsilon^0\left[\mathbf{K}_f(\nabla_x T_f^0+\nabla_y T_f^1)-\mathbf{K}_s(\nabla_x T_s^0+\nabla_y T_s^1)\right]\cdot\nu\\&+\varepsilon^1\left[\mathbf{K}_f(\nabla_x T_f^1+\nabla_y T_f^2)-\mathbf{K}_s(\nabla_x T_s^1+\nabla_y T_s^2)\right]\cdot\nu\\&+\mathcal{O}(\varepsilon^2),\quad \text{on}\ \Omega\times\Gamma\times J,\end{aligned} \tag{36}$$

and

$$(T_s^0-T_f^0)+\varepsilon(T_s-T_f^1)+\mathcal{O}(\varepsilon^2)=0\quad \text{on}\ \Omega\times\Gamma\times J. \tag{37}$$

It is evident from the above equations that at the lowest order, the displacement, pressure, and temperature has no y-dependence, so we write

$$p^0(x,y,t)=p^0(x,t),\qquad \text{in}\quad \Omega\times J, \tag{38a}$$
$$w^0(x,y,t)=w^0(x,t),\qquad \text{in}\quad \Omega\times J, \tag{38b}$$
$$T_s^0(x,y)=T_f^0(x,y)=T^0(x,t),\qquad \text{in}\quad \Omega\times J. \tag{38c}$$

However, as seen from Eqs. (30) and (31), at the lowest order, there is still a y-dependence in the fluid velocity.

3.2 The Flow

We now consider Eq. (30) at order $\mathcal{O}(\varepsilon^0)$, Eq. (31) at order $\mathcal{O}(\varepsilon^{-1})$, and the boundary condition, Eq. (33) at order $\mathcal{O}(\varepsilon^0)$, which gives the problem

$$\nabla_x p^0-\mu\Delta_y v^0=b-\nabla_y p^1,\qquad \text{in}\quad Y_f,$$
$$\nabla_y\cdot v^0=0,\qquad \text{in}\quad Y_f,$$

$$
\begin{aligned}
&v^0 = \partial_t w^0, && \text{on} \quad \Gamma,\\
&v^0(x,\cdot,t) \text{ and } p^1(x,\cdot,t) \text{ are } Y\text{-periodic}, && \forall (x,t) \in \Omega \times J.
\end{aligned}
$$

Note that since $\nabla_y v^0 = 0$ for $y \in Y_f$, we have: $\nabla_y \cdot 2\,\mathbf{e}_y(v^0) = \Delta_y v^0$. By defining $q = v^0 - \partial_t w^0$, we can rewrite the above problem as

$$
\begin{aligned}
&\nabla_x p^0 - \mu \Delta_y q = b - \nabla_y p^1, && \text{in} \quad Y_f, && \text{(39a)}\\
&\nabla_y \cdot q = 0, && \text{in} \quad Y_f, && \text{(39b)}\\
&q = 0, && \text{on} \quad \Gamma, && \text{(39c)}\\
&q(x,\cdot,t) \text{ and } p^1(x,\cdot,t) \text{ are } Y\text{-periodic}, && \forall (x,t) \in \Omega \times J, && \text{(39d)}
\end{aligned}
$$

which is the well-known cell problem in the homogenization of the filtration through rigid porous media, see, e.g., Hornung (2012) pp. 16–18. By using the identities $b = \sum_{j=1}^3 b_j e_j$, and $\nabla_x p^0 = \sum_{j=1}^3 \dfrac{\partial p}{\partial x_j} e_j$, where b_j is the j'th component of the body force, and $\{e_j\}_{j=1,2,3}$ the canonical basis of $\mathbb{R}^3$, we can solve for q and p^1 as follows

$$
q(x,y,t) = \frac{1}{\mu} \sum_{j=1}^3 \Lambda^j(y) \left(b_j(x,t) - \frac{\partial p^0}{\partial x_j}(x,t) \right), \tag{40}
$$

$$
p^1(x,y,t) = \sum_{j=1}^3 \Pi^j(y) \left(b_j(x,t) - \frac{\partial p^0}{\partial x_j}(x,t) \right), \tag{41}
$$

where Λ^j and Π^j ($\Lambda^j(y) \in \mathbb{R}^3$, $\Pi^j(y) \in \mathbb{R}$), are determined by the following cell problems (for $j = 1, 2, 3$)

$$
\begin{aligned}
&-\Delta_y \Lambda^j + \nabla_y \Pi^j = e_j, && \text{in} \quad Y_f, && \text{(42a)}\\
&\nabla_y \cdot \Lambda^j = 0, && \text{in} \quad Y_f, && \text{(42b)}\\
&\Lambda^j = 0, && \text{on} \quad \Gamma, && \text{(42c)}\\
&\Lambda^j \text{ and } \Pi^j \text{ are } Y\text{-periodic}. && && \text{(42d)}
\end{aligned}
$$

We integrate over Y_f and obtain the Darcy flux

$$
q_D(x,t) := \int_{Y_f} q(x,y,t)\mathrm{d}y = -\frac{1}{\mu} \mathbf{K}^H (\nabla_x p^0(x,t) - b(x,t)), \tag{43}
$$

where the effective coefficient, $\mathbf{K}^H$, (known as the permeability tensor) is given by:

$$
\left(\mathbf{K}^H\right)_{ij} = \int_{Y_f} (\Lambda^j(y))_i \mathrm{d}y, \quad i, j = 1, 2, 3. \tag{44}
$$

We get a similar expression for the average of v^0

$$
\int_{Y_f} v^0(x,y,t)\mathrm{d}y = \partial_t w^0(x,t)|Y_f| - \frac{1}{\mu} \mathbf{K}^H (\nabla_x p^0(x,t) - b(x,t)). \tag{45}
$$

It can be shown that the tensor $\mathbf{K}^H$ is symmetric and positive definite. We refer to Mikelić (1994) for a proof.

By using the expressions for q^0 and p^1, we also obtain the following which will be useful later

$$\left(2\mu\,\mathbf{e}_y(v^0) - p^1\mathbf{I}\right)\cdot\nu = \sum_{j=1}^{3}(2\,\mathbf{e}_y(\Lambda^j) - \Pi^j\mathbf{I})\left(b_j - \frac{\partial p^0}{\partial x_j}\right)\cdot\nu \quad \text{on } \Omega\times\Gamma\times J. \tag{46}$$

3.3 Momentum Conservation

From Eqs. (29) and (32) at order $\mathcal{O}(\varepsilon^{-1})$ we obtain

$$\nabla_y\cdot(\mathbf{C}(\mathbf{e}_x(w^0)+\mathbf{e}_y(w^1)) - \mathbf{M}\,T^0) = 0, \qquad \text{in} \quad \Omega\times Y_s\times J \tag{47a}$$

$$(\mathbf{C}(\mathbf{e}_x(w^0)+\mathbf{e}_y(w^1)) - \mathbf{M}\,T^0)\nu = -p^0\,\mathbf{I}\,\nu, \qquad \text{on} \quad \Omega\times\Gamma\times J, \tag{47b}$$

$$w^1(x,\cdot,t) \text{ is } Y\text{-periodic}, \qquad \forall (x,t)\in\Omega\times J. \tag{47c}$$

Using the tensor outer product (denoted "⊗"), we now make use of the following identity

$$\mathbf{e}_x(w^0) = \sum_{i,j=1}^{3}\frac{1}{2}\frac{\partial w_i^0}{\partial x_j}(e_i\otimes e_j + e_j\otimes e_i),$$

such that we can use $\dfrac{\partial w_i^0}{\partial x_j}$ as scalars in the expression for w^1

$$w^1(x,y,t) = \sum_{i,j=1}^{3}\frac{\partial w_i^0}{\partial x_j}(x,t)U^{ij}(y) + T^0(x,t)V(y) + p^0(x,t)W(y), \tag{48}$$

where the functions U^{ij}, V, and W, ($U^{ij}(y)$, $V(y)$, $W(y)\in\mathbb{R}^3$), are determined by the following cell problems (for $i,j=1,2,3$)

$$\nabla_y\cdot\left(\mathbf{C}\,\mathbf{e}_y(U^{ij})\right) = 0, \qquad \text{in} \quad Y_s, \tag{49a}$$

$$\mathbf{C}\left(\mathbf{e}_y(U^{ij}) + \frac{e_i\otimes e_j + e_j\otimes e_i}{2}\right)\nu = 0, \qquad \text{on} \quad \Gamma, \tag{49b}$$

$$U^{ij} \text{ is } Y\text{-periodic}, \tag{49c}$$

and

$$\nabla_y\cdot\left(\mathbf{C}\,\mathbf{e}_y(V)\right) = 0, \qquad \text{in} \quad Y_s, \tag{50a}$$

$$\mathbf{C}\,\mathbf{e}_y(V)\nu = \mathbf{M}\,\nu, \qquad \text{on} \quad \Gamma, \tag{50b}$$

$$V \text{ is } Y\text{-periodic}, \tag{50c}$$

and

$$\nabla_y\cdot\left(\mathbf{C}\,\mathbf{e}_y(W)\right) = 0, \qquad \text{in} \quad Y_s, \tag{51a}$$

$$\mathbf{C}\,\mathbf{e}_y(W)\nu = -\mathbf{I}\nu, \qquad \text{on} \quad \Gamma, \tag{51b}$$

$$W \text{ is } Y\text{-periodic}. \tag{51c}$$

We now continue with the solid at order $\mathcal{O}(\varepsilon^0)$, where we make use of the expression (46) from the last section. We thus obtain the following problem

$$\nabla_x \cdot (\mathbf{C}(\mathbf{e}_x(w^0) + \mathbf{e}_y(w^1)) - \mathbf{M}\,T^0) + b = -\nabla_y \cdot (\mathbf{C}(\mathbf{e}_x(w^1) + \mathbf{e}_y(w^2)) - \mathbf{M}\,T_s^1), \quad \text{in } \Omega \times Y_s \times J,$$

$$(\mathbf{C}(\mathbf{e}_x(w^1) + \mathbf{e}_y(w^2)) - \mathbf{M}\,T_s^1)\nu = \sum_{j=1}^{3}(2\,\mathbf{e}_y(\Lambda^j) - \Pi^j\mathbf{I})\left(b_j - \frac{\partial p^0}{\partial x_j}\right)\nu, \quad \text{on } \Omega \times \Gamma \times J,$$

$$w^2(x,\cdot,t) \text{ is } Y\text{-periodic}, \quad \forall (x,t) \in \Omega \times J.$$

Integrating the right hand side of the first equation over Y_s, using also Eq. (42a), yields

$$\begin{aligned}
-\int_{Y_s} \nabla_y \cdot (\mathbf{C}(\mathbf{e}_x(w^1) + \mathbf{e}_y(w^2)) - \mathbf{M}\,T_s^1)\mathrm{d}y &= \int_{\Gamma} (\mathbf{C}(\mathbf{e}_x(w^1) + \mathbf{e}_y(w^2)) - \mathbf{M}\,T_s^1)\nu \mathrm{d}s_y \\
&= \sum_{j=1}^{3} \int_{Y_f} \nabla_y \Pi^j - \Delta_y \Lambda^j \mathrm{d}y \left(\frac{\partial p^0}{\partial x_j} - b_j\right) \\
&= \sum_{j=1}^{3} \int_{Y_f} e_j \mathrm{d}y \left(\frac{\partial p^0}{\partial x_j} - b_j\right) \\
&= (\nabla_x p^0 - b)|Y_f|.
\end{aligned}$$

Using the expression for w^1, Eq. (48), we get for the left hand side

$$\begin{aligned}
& b|Y_s| + \nabla_x \cdot \int_{Y_s} (\mathbf{C}(\mathbf{e}_x(w^0) + \mathbf{e}_y(w^1)) - \mathbf{M}\,T^0)\mathrm{d}y \\
&= b|Y_s| + \nabla_x \cdot \left(p^0 \int_{Y_s} \mathbf{C}\,\mathbf{e}_y(W)\mathrm{d}y\right) + \nabla_x \cdot \left(T^0 \int_{Y_s} \mathbf{C}\,\mathbf{e}_y(V) - \mathbf{M}\,\mathrm{d}y\right) \\
&+ \nabla_x \cdot \left(\sum_{i,j=1}^{3} \frac{\partial w_i^0}{\partial x_j} \int_{Y_s} \mathbf{C}\left(\mathbf{e}_y(U^{ij}) + \frac{e_i \otimes e_j + e_j \otimes e_i}{2}\right)\mathrm{d}y\right).
\end{aligned}$$

Putting the two sides together gives the upscaled momentum equation

$$-\nabla_x \cdot \left(\mathbf{A}^H\,\mathbf{e}_x(w^0) - (|Y_f|\,\mathbf{I} - \mathbf{B}^H)p^0 - (|Y_s|\,\mathbf{M} - \mathbf{U}^H)T^0\right) = b, \quad \text{in } \Omega \times J, \tag{52}$$

where (for $i, j, k, l, = 1, 2, 3$)

$$(\mathbf{A}^H)_{ijkl} = \int_{Y_s} \left(\mathbf{C}\left(\mathbf{e}_y(U^{ij}(y)) + \frac{e_i \otimes e_j + e_j \otimes e_i}{2}\right)\right)_{kl} \mathrm{d}y, \tag{53}$$

$$(\mathbf{B}^H)_{ij} = \int_{Y_s} \left(\mathbf{C}\,\mathbf{e}_y(W(y))\right)_{ij} \mathrm{d}y, \tag{54}$$

$$(\mathbf{U}^H)_{ij} = \int_{Y_s} \left(\mathbf{C}\,\mathbf{e}_y(V(y))\right)_{ij} \mathrm{d}y. \tag{55}$$

The effective tensors $\mathbf{A}^H$ and $\mathbf{B}^H$ are symmetric and positive definite. We refer to Sanchez-Palencia (1980) for a proof. That $\mathbf{U}^H$ is symmetric and positive definite is shown the same way as for $\mathbf{B}^H$, except that it now relies on the same properties for $\mathbf{M}$.

3.4 Mass Conservation

In order to derive the upscaled mass conservation equation, we take terms of $\mathcal{O}(\varepsilon^0)$ from Eq. (31), together with $\mathcal{O}(\varepsilon^1)$ terms from the boundary condition (33), and obtain the following problem

$$\nabla_y \cdot v^1 = -\nabla_x \cdot v^0, \qquad \text{in} \quad \Omega \times Y_f \times J, \tag{56a}$$

$$v^1 = \partial_t w^1, \qquad \text{on} \quad \Omega \times \Gamma \times J, \tag{56b}$$

$$v^1(x, \cdot, t) \ \text{ is } Y\text{-periodic}, \qquad \forall (x,t) \in \Omega \times J. \tag{56c}$$

Integrating the left hand side of the first equation over Y_f, and using the expression for w^1, Eq. (48), yields

$$\begin{aligned}
&\int_{Y_f} \nabla_y \cdot v^1 \mathrm{d}y = -\int_{Y_s} \nabla_y \cdot \partial_t w^1 \mathrm{d}y \\
&= -\partial_t \left(\sum_{i,j=1}^{3} \frac{\partial w_i^0}{\partial x_j} \int_{Y_s} \nabla_y \cdot U^{ij} \mathrm{d}y + T^0 \int_{Y_s} \nabla_y \cdot V \mathrm{d}y + p^0 \int_{Y_s} \nabla_y \cdot W \mathrm{d}y \right) \\
&= -\mathbf{D}^H : \mathbf{e}_x(\partial_t w^0) - \partial_t T^0 E^H + \partial_t p^0 G^H,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{D}_{ij}^H &= \int_{Y_s} \nabla_y \cdot U^{ij} \mathrm{d}y, \\
E^H &= \int_{Y_s} \nabla_y \cdot V \mathrm{d}y, \\
G^H &= -\int_{Y_s} \nabla_y \cdot W \mathrm{d}y.
\end{aligned}$$

Integrating the right hand side of (56a) over Y_f, and using the expression for the average of v^0, Eq. (45), yields

$$-\nabla_x \cdot \left(\int_{Y_f} v^0 \mathrm{d}y \right) = -\nabla_x \cdot \left(\partial_t w^0 |Y_f| + q_D \right).$$

Putting the two sides together, we obtain the upscaled mass conservation equation

$$\mathbf{D}^H : \mathbf{e}_x(\partial_t w^0) + \partial_t T^0 E^H - \partial_t p^0 G^H = \nabla_x \cdot \left(\partial_t w^0 |Y_f| + q_D \right). \tag{57}$$

By testing with W in the cell problems (50) it is easily shown that $G^H > 0$, i.e.,

$$G^H = -\int_{Y_s} \nabla_y \cdot W \mathrm{d}y = \int_{Y_s} \mathbf{C}\, \mathbf{e}_y(W) : \mathbf{e}_y(W) \mathrm{d}y > 0 \tag{58}$$

The identification $\mathbf{D}^H = \mathbf{B}^H$ is shown by testing first with U^{ij} in the cell problem (50) to obtain

$$\mathbf{D}_{ij}^H = \int_{Y_s} \nabla_y \cdot U^{ij} \mathrm{d}y = -\int_{Y_s} \mathbf{C}\, \mathbf{e}_y(W) : \mathbf{e}_y(U^{ij}) \mathrm{d}y, \tag{59}$$

and on the other hand, by testing with W in cell problem (49)

$$\mathbf{B}_{ij}^H = \int_{Y_s} (\mathbf{C}\, \mathbf{e}_y(W))_{ij} \mathrm{d}y = -\int_{Y_s} \mathbf{C}\, \mathbf{e}_y(U^{ij}) : \mathbf{e}_y(W) \mathrm{d}y. \tag{60}$$

Using this, we can rewrite Eq. (57) as

$$\partial_t \left(p^0 G^H - T^0 E^H \right) + \nabla_x \cdot \left((|Y_f| \mathbf{I} - \mathbf{B}^H) \partial_t w^0 + q_D \right) = 0. \tag{61}$$

3.5 Energy Conservation

In this section we derive the upscaled energy conservation equation.

We consider the terms of order $\mathcal{O}(\varepsilon^{-1})$ from Eqs. (34), (35) and (36), and obtain the following problem

$$\nabla_y \cdot \left(\mathbf{K}_f \left(\nabla_x T^0 + \nabla_y T_f^1 \right) \right) = 0, \qquad \text{in} \quad \Omega \times Y_f \times J, \tag{62a}$$

$$\nabla_y \cdot \left(\mathbf{K}_s \left(\nabla_x T^0 + \nabla_y T_s^1 \right) \right) = 0, \qquad \text{in} \quad \Omega \times Y_s \times J, \tag{62b}$$

$$\mathbf{K}_f \left(\nabla_x T^0 + \nabla_y T_f^1 \right) \cdot \nu = \mathbf{K}_s \left(\nabla_x T^0 + \nabla_y T_s^1 \right) \cdot \nu, \qquad \text{on} \quad \Omega \times \Gamma \times J, \tag{62c}$$

$$T_f^1 = T_s^1 \qquad \text{on} \quad \Omega \times \Gamma \times J, \tag{62d}$$

$$T_f^1(x, \cdot, t) \text{ and } T_s^1(x, \cdot, t) \text{ are } Y\text{-periodic}, \qquad \forall (x, t) \in \Omega \times J. \tag{62e}$$

Using the identity $\nabla_x T^0 = \sum_{j=1}^3 \dfrac{\partial T^0}{\partial x_j} e_j$, we can solve for T_f^1 and T_s^1 as

$$T_f^1(x, y, t) = \sum_{j=1}^{3} \frac{\partial T^0(x, t)}{\partial x_j} \theta_f^j(y) \quad \text{and} \quad T_s^1(x, y, t) = \sum_{j=1}^{3} \frac{\partial T^0(x, t)}{\partial x_j} \theta_s^j(y), \tag{63}$$

where θ_f^j and θ_s^j ($\theta_f^j(y), \theta_s^j(y) \in \mathbb{R}$) are determined by (for $j = 1, 2, 3$)

$$\nabla_y \cdot (\mathbf{K}_f \nabla_y \theta_f^j) = 0, \qquad \text{in} \quad Y_f,$$

$$\nabla_y \cdot (\mathbf{K}_s \nabla_y \theta_s^j) = 0, \qquad \text{in} \quad Y_s,$$

$$\mathbf{K}_f (e_j + \nabla_y \theta_f^j) \cdot \nu = \mathbf{K}_s (e_j + \nabla_y \theta_s^j) \cdot \nu, \qquad \text{on} \quad \Gamma,$$

$$\theta_f^j = \theta_s^j, \qquad \text{on} \quad \Gamma,$$

$$\theta_f^j \text{ and } \theta_s^j \text{ are } Y\text{-periodic}.$$

By defining

$$\theta^j(y) = \begin{cases} \theta_f^j(y), & \text{if } y \in Y_f, \\ \theta_s^j(y), & \text{if } y \in Y_s, \end{cases} \tag{64}$$

due to the boundary condition, and using the properties of $\mathbf{K}_s$ and $\mathbf{K}_f$, we can write the more convenient problem

$$\Delta_y \theta^j = 0, \qquad \text{in} \quad Y_s \cup Y_f, \tag{65a}$$

$$\left(e_j + \nabla_y \theta^j \right) \cdot \nu = 0, \qquad \text{on} \quad \Gamma, \tag{65b}$$

$$\theta^j \text{ is } Y\text{-periodic}. \tag{65c}$$

Continuing to the next order, $\mathcal{O}(\varepsilon^0)$, we obtain the problem

$$
\begin{aligned}
&v^0 \cdot \nabla_x T^0 + v^0 \cdot \nabla_y T_f^1 - \nabla_x \cdot \left(\mathbf{K}_f \left(\nabla_x T^0 + \nabla_y T_f^1 \right) \right) \\
&\quad = \nabla_y \cdot \left(\mathbf{K}_f \left(\nabla_x T_f^1 + \nabla_y T_f^2 \right) \right), && \text{in } \Omega \times Y_f \times J, \\
&\mathbf{M} : \left(\mathbf{e}_x \left(\partial_t w^0 \right) + \mathbf{e}_y \left(\partial_t w^1 \right) \right) - \nabla_x \cdot \left(\mathbf{K}_s \left(\nabla_x T^0 + \nabla_y T_s^1 \right) \right) \\
&\quad = \nabla_y \cdot \left(\mathbf{K}_s \left(\nabla_x T_s^1 + \nabla_y T_s^2 \right) \right), && \text{in } \Omega \times Y_s \times J, \\
&\mathbf{K}_f \left(\nabla_x T_f^1 + \nabla_y T_f^2 \right) \cdot \nu = \mathbf{K}_s \left(\nabla_x T_s^1 + \nabla_y T_s^2 \right) \cdot \nu, && \text{on } \Omega \times \Gamma \times J, \\
&T_f^2 = T_s^2 && \text{on } \Omega \times \Gamma \times J, \\
&T_f^2(x, \cdot, t) \text{ and } T_s^2(x, \cdot, t) \text{ are } Y\text{-periodic}, && \forall (x, t) \in \Omega \times J.
\end{aligned}
$$

Integrating the first equation over Y_f, and using the expressions for T_f^1 and the average of v^0, Eqs. (63) and (45), together with the boundary conditions (62d) and (39a) yields

$$
\begin{aligned}
&\int_\Gamma \mathbf{K}_f \left(\nabla_x T_f^1 + \nabla_y T_f^2 \right) \cdot \nu \mathrm{d}s \\
&= \left(|Y_f| \partial_t w^0 + \mathbf{q} \right) \cdot \nabla_x T^0 + \int_{Y_f} v^0 \cdot \nabla_y T_f^1 \mathrm{d}y - \nabla_x \cdot \left(\int_{Y_f} \mathbf{K}_f \left(\nabla_x T^0 + \nabla_y T_f^1 \right) \mathrm{d}y \right) \\
&= q_D \cdot \nabla_x T^0 + \partial_t w^0 \cdot \sum_{j=1}^{3} \frac{\partial T^0}{\partial x_j} \int_{Y_f} e_j + \nabla_y \theta_f^j \mathrm{d}y - \nabla_x \cdot \left(\sum_{j=1}^{3} \frac{\partial T^0}{\partial x_j} \int_{Y_f} \mathbf{K}_f \left(e_j + \nabla_y \theta_f^j \right) \mathrm{d}y \right).
\end{aligned}
$$

Integrating the second equation over Y_s, using also the expressions for T_s^1 and w^1, Eqs. (63) and (48), yields

$$
\begin{aligned}
&\int_\Gamma \mathbf{K}_s \left(\nabla_x T_f^1 + \nabla_y T_f^2 \right) \cdot \nu \mathrm{d}s \\
&= \nabla_x \cdot \left(\int_{Y_s} \mathbf{K}_s \left(\nabla_x T^0 + \nabla_y T_s^1 \right) \mathrm{d}y \right) - |Y_s| \, \mathbf{M} : \mathbf{e}_x \left(\partial_t w^0 \right) - \int_{Y_s} \mathbf{M} : \mathbf{e}_y \left(\partial_t w^1 \right) \mathrm{d}y \\
&= \nabla_x \cdot \left(\sum_{j=1}^{3} \frac{\partial T^0}{\partial x_j} \int_{Y_s} \mathbf{K}_s \left(e_j + \nabla_y \theta_s^j \right) \mathrm{d}y \right) - |Y_s| \, \mathbf{M} : \mathbf{e}_x \left(\partial_t w^0 \right) \\
&\quad - \mathbf{e}_x \left(\partial_t w^0 \right) : \sum_{i,j=1}^{3} \int_{Y_s} \mathbf{M} : \mathbf{e}_y \left(U^{ij} \right) \mathrm{d}y - \partial_t T^0 \int_{Y_s} \mathbf{M} : \mathbf{e}_y(V) \mathrm{d}y - \partial_t p^0 \int_{Y_s} \mathbf{M} : \mathbf{e}_y(W) \mathrm{d}y.
\end{aligned}
$$

Since the left hand sides of the two above equations are equal, we put them together and obtain the upscaled energy conservation equation

$$
\begin{aligned}
&\partial_t T^0 M^H + \partial_t p^0 N^H + \partial_t w^0 \cdot \boldsymbol{\Xi}^H \nabla_x T^0 + q_D \cdot \nabla_x T^0 \\
&\quad + \nabla_x \cdot \left((\mathbf{R}^H + |Y_s| \, \mathbf{M}) \partial_t w^0 - \boldsymbol{\Theta}^H \nabla_x T^0 \right) = 0
\end{aligned}
$$

where (for $i, j = 1, 2, 3$)

$$(\boldsymbol{\Theta}^H)_{ij} = \int_{Y_f} \left(\mathbf{K}_f(e_j + \nabla_y \theta_f^j)\right)_i \mathrm{d}y + \int_{Y_s} \left(\mathbf{K}_s(e_j + \nabla_y \theta_s^j)\right)_i \mathrm{d}y,$$

$$(\boldsymbol{\Xi}^H)_{ij} = \int_{Y_f} (e_j + \nabla_y \theta_f^j)_i \mathrm{d}y,$$

$$(\mathbf{R}^H)_{ij} = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(U^{ij}) \mathrm{d}y,$$

$$M^H = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(V) \mathrm{d}y,$$

$$N^H = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(W) \mathrm{d}y.$$

Again, some properties of the coefficients can be established. By testing with V in the cell problems (50) we obtain

$$M^H = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(V) \mathrm{d}y = \int_{Y_s} \mathbf{C}\,\mathbf{e}_y(V) : \mathbf{e}_y(V) \mathrm{d}y > 0. \tag{66}$$

The identification $\mathbf{R}^H = -\mathbf{U}^H$ can also be shown by testing first with U^{ij} in the cell problem (50) to obtain

$$(\mathbf{R}^H)_{ij} = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(U^{ij}) \mathrm{d}y = \int_{Y_s} \mathbf{C}\,\mathbf{e}_y(V) : \mathbf{e}_y(U^{ij}) \mathrm{d}y, \tag{67}$$

and then with V in the cell problem (49)

$$(\mathbf{U}^H)_{ij} = \int_{Y_s} \left(\mathbf{C}\,\mathbf{e}_y(V)\right)_{ij} \mathrm{d}y = -\int_{Y_s} \mathbf{C}\,\mathbf{e}_y(U^{ij}) : \mathbf{e}_y(V) \mathrm{d}y. \tag{68}$$

Further, we can also show $N^H = -E^H$, by testing first with V in the cell problem (51)

$$E^H = \int_{Y_s} \nabla_y \cdot V \mathrm{d}y = -\int_{Y_s} \mathbf{C}\,\mathbf{e}_y(W) : \mathbf{e}_y(V) \mathrm{d}y, \tag{69}$$

and with W in the cell problem (51)

$$N^H = \int_{Y_s} \mathbf{M} : \mathbf{e}_y(W) \mathrm{d}y = \int_{Y_s} \mathbf{C}\,\mathbf{e}_y(V) : \mathbf{e}_y(W) \mathrm{d}y. \tag{70}$$

Lemma 1 *$\boldsymbol{\Theta}^H$ and $\boldsymbol{\Xi}^H$ are symmetric and positive definite.*

Proof Test with θ^i in the j'th cell problem (65), and by θ^j in the i'th problem to obtain

$$\int_{Y_f} \mathbf{K}_f \left(e_j + \nabla_y \theta^j\right) \cdot \nabla_y \theta^i \mathrm{d}y = \int_{Y_f} \mathbf{K}_f \left(e_i + \nabla_y \theta^i\right) \cdot \nabla_y \theta^j \mathrm{d}y = 0,$$

and

$$\int_{Y_s} \mathbf{K}_s \left(e_j + \nabla_y \theta^j\right) \cdot \nabla_y \theta^i \mathrm{d}y = \int_{Y_s} \mathbf{K}_s \left(e_i + \nabla_y \theta^i\right) \cdot \nabla_y \theta^j \mathrm{d}y = 0.$$

Thus, we can write $\boldsymbol{\Theta}^H$ as:

$$(\boldsymbol{\Theta}^H)_{ij} = \int_{Y_f} \mathbf{K}_f \left(e_j + \nabla_y \theta^j\right) \cdot \left(e_i + \nabla_y \theta^i\right) \mathrm{d}y + \int_{Y_s} \mathbf{K}_s \left(e_j + \nabla_y \theta^j\right) \cdot \left(e_i + \nabla_y \theta^i\right) \mathrm{d}y,$$

and it follows that $\boldsymbol{\Theta}^H$ is symmetric. For the positive definiteness, observe that for nonnegative $\alpha_{1,2,3} \in \mathbb{R}$ not all equal to zero we have

$$\sum_{i,j=1}^{3} (\boldsymbol{\Theta}^H)_{ij}\alpha_i\alpha_j = \sum_{i,j=1}^{3} \int_{Y_f} \mathbf{K}_f \nabla_y \left(\alpha_j(y_j + \theta^j)\right) \cdot \nabla_y \left(\alpha_i(y_i + \theta^i)\right) \mathrm{d}y + \sum_{i,j=1}^{3} \int_{Y_s} \mathbf{K}_s \nabla_y \left(\alpha_j(y_j + \theta^j)\right) \cdot \nabla_y \left(\alpha_i(y_i + \theta^i)\right) \mathrm{d}y > 0.$$

That $\boldsymbol{\Xi}^H$ is symmetric and positive definite is shown in the same way. □

We can now rewrite the upscaled energy conservation equation as

$$\partial_t(T^0 M^H - p^0 E^H) + (\boldsymbol{\Xi}^H \partial_t w^0 + q_D)\cdot\nabla_x T^0 + \nabla_x \cdot \left((|Y_s|\,\mathbf{M} - \mathbf{U}^H)\partial_t w^0 - \boldsymbol{\Theta}^H \nabla_x T^0\right) = 0. \tag{71}$$

4 Summary

4.1 The Upscaled Quasi-static Thermo-poroelastic System

We now summarize the upscaled equations derived in the previous sections. We omit all superscripts in the variables, subscripts in the differential operators (with the understanding they are now all taken with respect to the slow variable x), and introduce a more familiar notation for the coefficients, similar to what is commonly used in the literature on the quasi-static poroelastic equations:

$$\begin{aligned} &\boldsymbol{\alpha} := (|Y_f|\,\mathbf{I} - \mathbf{B}^H), && \boldsymbol{\beta} := (|Y_f|\,\mathbf{M} - \mathbf{U}^H), && \mathbf{A} := \mathbf{A}^H \\ &\mathbf{K} := \mathbf{K}^H, && \boldsymbol{\Theta} := \boldsymbol{\Theta}^H, && \boldsymbol{\Xi} := \boldsymbol{\Xi}^H, \\ &c_0 := G^H, && a_0 := M^H, && b_0 := E^H, \end{aligned}$$

where $\boldsymbol{\alpha}$ is the Biot–Willis constant, c_0 is the specific storage coefficient, and $\mathbf{A}$ is the effective elastic moduli, containing the elastic coefficients of the porous medium.

Thus, we write the upscaled system as:

$$q_D = -\frac{1}{\mu}\mathbf{K}(\nabla p - \mathbf{b}), \qquad \text{in } \Omega \times J, \tag{72a}$$

$$-\nabla \cdot (\mathbf{A}\,\mathbf{e}(w) - \boldsymbol{\alpha}\, p - \boldsymbol{\beta}\, T) = \mathbf{b}, \qquad \text{in } \Omega \times J, \tag{72b}$$

$$\partial_t(c_0 p - b_0 T + \nabla \cdot \boldsymbol{\alpha}\, w) + \nabla \cdot q_D = 0, \qquad \text{in } \Omega \times J, \tag{72c}$$

$$\partial_t(a_0 T - b_0 p + \nabla \cdot \boldsymbol{\beta}\, w) + (\boldsymbol{\Xi}\, \partial_t w + q_D) \cdot \nabla T - \nabla \cdot (\boldsymbol{\Theta}\, \nabla T) = 0, \qquad \text{in } \Omega \times J. \tag{72d}$$

Compared to the linear poroelastic equations, we see that the stress in the momentum Eq. (72b) now has an additional linear dependency on the temperature of the medium, i.e., $\boldsymbol{\sigma} = \boldsymbol{\sigma}(w, p, T) = \mathbf{A}\,\mathbf{e}(w) - \boldsymbol{\alpha}\, p - \boldsymbol{\beta}\, T$. This is completely analogous to the linear thermoelastic equations in mechanics, see, e.g., Pabst (2005). The homogenized tensor $\boldsymbol{\beta}$ can be interpreted

as an upscaled thermal stress coefficient, giving the induced thermal stress coming from a unit temperature gradient. In the mass conservation Eq. (72c) we see that the porosity (denoted by η) is also linearly dependent on the temperature, i.e., $\eta = \eta(w, p, T) = c_0 p - b_0 T + \nabla \cdot \boldsymbol{\alpha}\, w$. In other words, the amount of fluid that can be injected into an arbitrary fixed control volume is now given by: $c_0 p - b_0 T$, where the homogenized coefficient b_0 can be interpreted as a thermal expansion coefficient.

It remains to discuss the energy conservation Eq. (72d). If we had used the different scaling corresponding to a small Péclet number, i.e., $Pe \sim \mathcal{O}(\varepsilon)$, the dimensionless energy conservation equation for the fluid at the microscale, Eq. (28g), would take the form:

$$\varepsilon \left(\frac{\partial T_f^\varepsilon}{\partial t} + \partial_t u^\varepsilon \cdot \nabla T_f^\varepsilon \right) - \nabla \cdot \left(\mathbf{K}_f \nabla T_f^\varepsilon \right) = 0, \qquad \text{in} \quad \Omega_f^\varepsilon \times J. \tag{73}$$

Then, after separating the scales, the terms: $\varepsilon(\frac{\partial T_f^\varepsilon}{\partial t} + \partial_t u^\varepsilon \cdot \nabla T_f^\varepsilon)$ give no contribution to the $\mathcal{O}(\varepsilon^{-1})$-problem, and for the $\mathcal{O}(\varepsilon^0)$-problem, we only retain the term: $\partial_t u^0 \cdot \nabla_y T^0$, which is evidently equal to zero. Thus, the upscaled energy conservation equation corresponding to a small Péclet number is:

$$\partial_t (a_0 T - b_0 p + \nabla \cdot \boldsymbol{\beta}\, w) - \nabla \cdot (\boldsymbol{\Theta} \nabla T) = 0, \qquad \text{in} \quad \Omega \times J, \tag{74}$$

and we have a fully linear upscaled system.

Denoting by: $\xi = \xi(w, p, T) = a_0 T - b_0 p + \nabla \cdot \boldsymbol{\beta}\, w$, the energy present in some arbitrary control volume, we see from Eq. (72d) that the rate of change of energy present, $\partial_t \xi$, is balanced by the net energy flux into the same control volume, either by conduction: $-\nabla \cdot (\boldsymbol{\Theta} \nabla T)$, or convection: $(\boldsymbol{\Xi}\, \partial_t w + q_D) \cdot \nabla T$. We see also from Eq. (73) that in the case of a small Péclet number, the rate of change in energy present is balanced only by the conduction. The homogenized tensors $\boldsymbol{\Theta}$ and $\boldsymbol{\Xi}$ can be interpreted as a kind of upscaled thermal conductivities, while a_0 gives the energy present by a unit temperature rate of change.

An important property of the linear poroelastic equations is that the Biot–Willis coefficient appears both in front of the pressure term in the momentum Eq. (72b) (i.e., $\boldsymbol{\alpha}\, p$), and in front of the volumetric term in the mass conservation Eq. (72c) (i.e., $\nabla \cdot \boldsymbol{\alpha}\, w$). As described in Coussy (1995) p. 75, a similar situation is expected with the temperature term in the momentum equation and the volumetric term in the energy equation. As we see from Eqs. (72b) and (72d), this is indeed the case, as the thermal stress coefficient, $\boldsymbol{\beta}$, appears in both places. Another interesting fact is the coefficient b_0 which appear both in front of the temperature term in the mass conservation Eq. (72c) and in front of the pressure term in the energy Eq. (72d). This is indeed also the case in Coussy (1995), where they refer to the coefficient b_0 as the "volumetric thermal dilation coefficient related to the porosity."

In the article Lee and Mei (1997), the allowable deformations at the microscale are much smaller than the microscale length (i.e., l), which makes a direct comparison of our models difficult. However, we note that also here there is a linear dependency on temperature in the upscaled solid stress, and thus, Eq. (72c) matches that of Lee and Mei (1997). Our energy equations differ significantly, on the other hand, as in Lee and Mei (1997) there appears no pressure term.

4.2 Conclusions

We have presented a formal upscaling of the microscale thermal fluid-structure problem in porous media, leading to thermal Biot equations at the macroscale. This derivation gives a

precise understanding of the coupling terms at the macroscale and forms a justification for heuristically derived models.

The most important limitation of the formal approach taken herein is the use of a perfectly periodic geometry. This limitation will not affect the applicability of the results to some human-made porous media, but will invalidate the approach for natural porous media. However, it has been shown for similar problems that the periodicity assumption can be relaxed, and we expect that these results would be possible to extend to the present setting. As such we expect the structure of the equations summarized in the preceding section to be valid also for non-periodic porous media, at least when there is some uniformity on the sizes of the solid grains.

All homogenization results are based on a series of "smallness" assumptions. We emphasize the important understanding that the results presented herein are based on ε being "small," but not tending to zero. This distinction is important, since some of the parameters defined in the homogenization procedure (such as i.e., permeability) may depend on ε, depending on the choice of characteristic macroscopic length scales. A further comment in this regard is that the Lagrangian formulation used herein implies that we only assume that the strain is small, and not that the displacement itself is small. Alternatively, if an Eulerian framework was used, it would be necessary to assume that the displacement is small relative to the microscale, which would preclude meaningful macroscopic deformations.

In this work we have chosen to a large extent to linearize the governing equations already at the microscale. This in part explains the linear structure of the majority of terms on the macroscale. Nonlinear constitutive relationships could be accommodated at the cost of technical and notational complexity, varying from relatively straight-forward (i.e., nonlinear constitutive laws for fluid density) to complex (nonlinear elastic or plastic constitutive laws for material deformation).

Acknowledgements This work is partly supported by the Research Council of Norway Project 250223. The authors would like to thank the anonymous reviewers, Prof. Andro Mikelić, and Carina Bringedal for reading the manuscript and providing helpful comments, all of which contributed to the improvement of the paper. The authors also acknowledge the support from the University of Bergen.

References

Allaire, G.: Homogenization of the Stokes flow in a connected porous medium. Asymptot. Anal. **2**(3), 203–222 (1989)

Auriault, J.L.: Heterogeneous medium. Is an equivalent macroscopic description possible? Int. J. Eng. Sci. **29**(7), 785–795 (1991)

Biot, M.A.: General theory of three-dimensional consolidation. J. Appl. Phys. **12**(2), 155–164 (1941)

Biot, M.A.: Theory of finite deformations of porous solids. Indiana Univ. Math. J. **21**(7), 597–620 (1972)

Biot, M.A.: Variational Lagrangian-thermodynamics of nonisothermal finite strain mechanics of porous solids and thermomolecular diffusion. Int. J. Solids Struct. **13**(6), 579–597 (1977)

Bringedal, C., Berre, I., Pop, I.S., Radu, F.A.: Upscaling of non-isothermal reactive porous media flow with changing porosity. Transp. Porous Media **114**(2), 371–393 (2016)

Burridge, R., Keller, J.B.: Poroelasticity equations derived from microstructure. J. Acoust. Soc. Am. **70**(4), 1140–1146 (1981)

Cioranescu, D., Donato, P.: Introduction to Homogenization. Oxford University Press, Oxford (2000)

Clopeau, T., Ferrin, J.L., Gilbert, R.P., Mikelić, A.: Homogenizing the acoustic properties of the seabed, part II. Math. Comput. Model. **33**(8–9), 821–841 (2001)
Coussy, O.: Mechanics of Porous Continua, pp. 71–108. Wiley, Hoboken (1995)
Eden, M., Muntean, A.: Homogenization of a fully coupled thermoelasticity problem for a highly heterogeneous medium with a priori known phase transformations. Math. Methods Appl. Sci. **40**, 3955–3972 (2017)
Ferrin, J.L., Mikelić, A.: Homogenizing the acoustic properties of a porous matrix containing an incompressible inviscid fluid. Math. Methods Appl. Sci. **26**(10), 831–859 (2003)
Gilbert, R.P., Mikelić, A.: Homogenizing the acoustic properties of the seabed: Part I. Nonlinear Anal. Theory Methods Appl. **40**(1–8), 185–212 (2000)
Hornung, U.: Homogenization and porous media, vol. 6. Springer, Berlin (2012)
Iliev, O., Mikelić, A., Popov, P.: On upscaling certain flows in deformable porous media. Multiscale Model. Simul. **7**(1), 93–123 (2008)
Kupradze, V.D., Gegelia, T.G., Basheleishvili, M.O., Burchuladze, T.V.: Three-dimensional problems of the mathematical theory of elasticity and thermoelasticity. Translated from the second Russian edition. Edited by VD Kupradze, North-Holland Series in Applied Mathematics and Mechanics (1979)
Lee, C.K., Mei, C.C.: Thermal consolidation in porous media by homogenization theory—I. Derivation of macroscale equations. Adv. Water Resour. **20**(2), 127–144 (1997)
Lévy, T.: Propagation of waves in a fluid-saturated porous elastic solid. Int. J. Eng. Sci. **17**(9), 1005–1014 (1979)
Mikelić, A.: Mathematical derivation of the Darcy-type law with memory effects, governing transient flow through porous medium. Glasnik Matematicki **29**(49), 57–77 (1994)
Mikelić, A., Wheeler, M.F.: Theory of the dynamic Biot–Allard equations and their link to the quasi-static Biot system. J. Math. Phys. **53**(12), 123, 702 (2012)
Pabst, W.: The linear theory of thermoelasticity from the viewpoint of rational thermomechanics. Ceram Silikaty **49**(4), 242–251 (2005)
Sanchez-Palencia, E.: Non-homogeneous media and vibration theory. Lect. Notes Phys. **127**, 158–190 (1980)
Silhavy, M.: The Mechanics and Thermodynamics of Continuous Media. Springer, Berlin (2013)
Terzaghi, K.: Theoretical Soil Mechanics. Chapman and Hali, Limited John Wiler and Sons Inc, New York (1944)
Van Duijn, C., Mikelic, A., Wheeler, M., Wick, T.: Thermoporoelasticity via homogenization I. Modeling and formal two-scale expansions. In: Working Paper or Preprint (2017). URL https://hal.archives-ouvertes.fr/hal-01650194
Wood, J.R., Hewett, T.A.: Fluid convection and mass transfer in porous sandstones? A theoretical model. Geochimica et Cosmochimica Acta **46**(10), 1707–1713 (1982)

Paper B

Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport

M. K. Brun, E. Ahmed, J. M. Nordbotten, F. A. Radu

Journal of Mathematical Analysis and Applications **471(1–2)** (2018), p. 239–266.
doi: 10.1016/j.jmaa.2018.10.074

J. Math. Anal. Appl. 471 (2019) 239–266

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport ☆

Mats Kirkesæther Brun [a,*], Elyes Ahmed [a], Jan Martin Nordbotten [a,b], Florin Adrian Radu [a]

[a] *Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020 Bergen, Norway*
[b] *Princeton Environmental Institute, Princeton University, Princeton, N.J., USA*

ARTICLE INFO

Article history:
Received 23 July 2018
Available online 29 October 2018
Submitted by A. Mazzucato

Keywords:
Thermo-poroelasticity
Nonlinear convective transport
Biot's model
Well-posedness
Galerkin's method
Convergence analysis

ABSTRACT

This paper is concerned with the analysis of the quasi-static thermo-poroelastic model. This model is nonlinear and includes thermal effects compared to the classical quasi-static poroelastic model (also known as Biot's model). It consists of a momentum balance equation, a mass balance equation, and an energy balance equation, fully coupled and nonlinear due to a convective transport term in the energy balance equation. The aim of this article is to investigate, in the context of mixed formulations, the existence and uniqueness of a weak solution to this model problem. The primary variables in these formulations are the fluid pressure, temperature and elastic displacement as well as the Darcy flux, heat flux and total stress. The well-posedness of a linearized formulation is addressed first through the use of a Galerkin method and suitable *a priori* estimates. This is used next to study the well-posedness of an iterative solution procedure for the full nonlinear problem. A convergence proof for this algorithm is then inferred by a contraction of successive difference functions of the iterates using suitable norms.

1. Introduction

The field of *poroelasticity* is concerned with describing the interaction between viscous fluid flow and elastic solid deformation within a porous material, and goes back to the works of K. Terzhagi [32] and M.A. Biot [6,7]. Porous materials are by definition solid materials comprising a great number of interconnected pores, typically at the order of micrometers, where the interconnectivity of the pores is sufficient to allow for fluid flow through the material. For this reason, porous materials are usually modeled at the

☆ This work forms part of Norwegian Research Council project 250223.
* Corresponding author.
E-mail addresses: mats.brun@uib.no (M.K. Brun), elyes.ahmed@uib.no (E. Ahmed), jan.nordbotten@uib.no (J.M. Nordbotten), florin.radu@uib.no (F.A. Radu).

https://doi.org/10.1016/j.jmaa.2018.10.074
0022-247X/

continuum scale, such that the complex micro-structure needs not be explicitly accounted for in the modeling, but rather implicitly through so-called *effective parameters* such as e.g. *porosity* and *permeability*. Porous materials are primarily associated with objects such as rocks and clays, but biological tissue, foams and paper products also fall within this category. Consequently, the field of *poroelasticity* is of great importance in a range of different engineering disciplines, such as petroleum engineering, agricultural science and biomedicine, among others. A number of comprehensive text books related to the field exists; see e.g. [13,14,36].

Mathematical modeling of fluid saturated deformable porous media on the continuum scale relies on the theory of linear elasticity, adapted to porous materials by using the so-called *total stress tensor* instead of the Cauchy stress in the momentum balance equation. In particular, the total stress tensor is a linear combination of the Cauchy stress for the empty elastic skeleton and the isotropic stress coming from the fluid, i.e. the pore pressure. Within the quasi-static framework inertial terms are ignored, thus giving a purely elliptic equation for the momentum balance. A second equation of parabolic type accounts for the mass balance as fluid is displaced by the deformation of the solid, and relates change in porosity to volumetric fluid flow, i.e. the *Darcy flux*. This is essentially Biot's poroelastic model for quasi-static deformation (see e.g. [6,13]). There is an extensive literature on this model problem and on its numerical approximation. To mention a few, the well-posedness based on the canonical two-field formulation with displacement and pressure as variables was carried out in [30], while three and four-field formulations have also been analyzed (taking Darcy flux and total stress as independent variables), and can be found in several studies, e.g. [1, 26,37]. A key feature of this model, one which greatly facilitates the analysis, is the symmetric coupling between the equations.

In many important applications, such as geothermal energy extraction, nuclear waste disposal and carbon storage, temperature also plays a vital role and must therefore be included in the modeling. Using the method of formal two-scale expansions (see e.g. [12,18] for a detailed review of this method), a thermo-poroelastic model was derived in [10], which accounts for fluid pressure, elastic displacement, and temperature distribution within a fine-grained, fully saturated poroelastic material within the framework of quasi-static deformation. This model is similar to other thermo-poroelastic models which exists in the literature; see e.g. [13,16,22,31,34], although there are also some notable differences among these works, in particular from the modeling point of view; i.e. allowable flow rates and deformation, choice of coordinate frames etc. (see [10,34] for a comparison of existing thermo-poroelastic models). However, from the point of view of analysis the important factor is the coupling structure between the equations, and the model we analyze exhibits a fully coupled structure.

The aim of the present work is to establish the well-posedness of the nonlinear thermo-poroelastic model as described in [10], where we also provide *a priori* energy estimates and regularity properties of the solutions. We restrict our attention to an isotropic material such that the elastic coefficients are given by the Lamé parameters, and the Biot coefficient and thermal stress coefficient are given by scalar quantities. Some algebraic constraints on these coefficients must be imposed in order to obtain our results. Although the literature on the analysis of poroelastic models is quite extensive, there is not much literature on the analysis of thermo-poroelastic models; in [34] a corresponding energy functional for the thermo-poroelastic model was derived. This functional was then shown to be monotonically decreasing in time for a small enough characteristic temperature difference.

We undertake our analysis with a future mixed finite-element implementation in mind, and therefore double the number of variables from three to six, and investigate the existence and uniqueness of a weak solution corresponding to this fully coupled six-field model. The primary variables in this model are; fluid pressure, temperature, elastic displacement, Darcy flux, heat flux, and total stress. This makes the problem suitable for combinations of well-known stable finite-elements, such as Raviart-Thomas(-Nédélec) [25,29] and Arnold–Winther [2,3]. From an implementation point of view there are several advantages of a mixed formulation over the canonical three-field formulation; the discretization respects mass and energy conser-

vation, produces continuous normal fluxes regardless of mesh quality, and in general a mixed formulation is advantageous for domain decomposition techniques. We restrict our attention to two spatial dimensions, as this will be the most relevant case for the subsequent work, although the results we present can be extended to higher dimensions in a straightforward manner. In particular, the definition of the isotropic compliance tensor must reflect the choice of spatial dimension.

The main difficulty we face in the following analysis is the nonlinear coupling between the equations, i.e. the nonlinear convective transport term in the energy balance equation, which takes the form $\nabla T \cdot \mathbf{w}$, where $\mathbf{w}$ is the Darcy flux, and T is the temperature distribution. The first part of the paper is therefore concerned with analyzing a linearized version of the model, where we write the convective transport term as $\boldsymbol{\eta} \cdot \mathbf{w}$, for some given $\boldsymbol{\eta} \in L^\infty$ (the remaining coupling terms are retained). Once we have obtained the existence and uniqueness of a weak solution to this problem, we introduce an iterative algorithm where we approximate the convective transport term as $\nabla T^{m-1} \cdot \mathbf{w}^m$, where $m \geq 1$ is the iteration index. Due to the results we obtained for the linearized problem, and by a natural assumption that the temperature gradient admits L^∞-regularity in space, we construct a well defined sequence of iterates as $m \to \infty$. This we show to converge in adequate norms to the solution of the original nonlinear problem, thus establishing the existence and uniqueness of its weak solution. The convergence proof relies on the Banach Fixed Point Theorem, which we use to obtain local solutions in time. Here, the time interval is supposed to be small to ensure a contraction of the successive difference functions of the iterates. Then, using piecewise continuation in time, we extend these local solutions to global solutions for any finite final time. The idea is that such an iterative scheme can also be applied numerically to a discretized formulation, and in this sense our analysis sets the stage for subsequent numerical experiments. We mention also some of the literature on iterative schemes in poroelasticity; in [5,8,20,24] there can be found several iterative procedures for solving Biot's equations, and in [23,27,28] iterative methods for solving Richards' equation were analyzed.

We summarize the main contribution of the article as follows: under a natural hypothesis on the regularity of the convective term, we give a proof of existence and uniqueness of a weak solution to the fully coupled six-field thermo-poroelastic problem within the quasi-static framework.

The article is organized as follows: Section 2 recalls the physical model and the assumptions on the data, introduces the relevant function spaces and introduces the mixed weak formulations. In section 3 we define a linear version of the original mixed variational problem, and proceed to analyze this in the following way; we construct approximate solutions using a Galerkin method, the existence of which is established by the theory of DAEs (Differential Algebraic Equations). Suitable *a priori* estimates are then derived which enables us to pass to the limit, thanks to the weak compactness of the spaces. Section 4 is devoted to analyzing an iterative solution procedure for the original nonlinear problem and to establish the convergence of the algorithm in suitable norms. In Appendix A we propose an alternative to the hypothesis on the temperature gradient, i.e. we show how the required regularity can be obtained by sufficient regularity of the data. For easy reference of the notation used in this article we provide some tables in Appendix B.

2. Presentation of the problem

Let $\Omega \subset \mathbb{R}^d$, for $d \in \{2, 3\}$, be an open and bounded domain, where we denote the boundary by $\Gamma := \partial\Omega$, which is assumed to be Lipschitz continuous. Let a time interval $J = (0, T_f)$ be given with $T_f > 0$ and define $Q := \Omega \times (0, T_f]$ to be the space–time domain. The thermo-poroelastic model problem we consider, as it is exposed in [10], is as follows: given a heat source h, a body force $\mathbf{f}$, and a mass source g, find $(T, \mathbf{u}, p)$ such that

$$\partial_t(a_0 T - b_0 p + \beta \nabla \cdot \mathbf{u}) - \nabla T \cdot (\mathbf{K} \nabla p) - \nabla \cdot (\boldsymbol{\Theta} \nabla T) = h, \quad \text{in } Q, \tag{2.1a}$$

$$-(\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) - \mu \nabla^2 \mathbf{u} + \alpha \nabla p + \beta \nabla T = \mathbf{f}, \quad \text{in } Q, \tag{2.1b}$$

$$\partial_t(c_0 p - b_0 T + \alpha \nabla \cdot \mathbf{u}) - \nabla \cdot (\mathbf{K} \nabla p) = g, \quad \text{in } Q, \tag{2.1c}$$

where a_0 is the effective thermal capacity, b_0 is the thermal dilation coefficient, β is the thermal stress coefficient, $\mathbf{K} = (K_{ij})_{i,j=1}^d$ is the permeability divided by fluid viscosity, $\mathbf{\Theta} = (\Theta_{ij})_{i,j=1}^d$ is the effective thermal conductivity, μ and λ are the Lamé parameters, α is the Biot–Willis constant and c_0 is the specific storage coefficient. The primary variables are the temperature distribution T, displacement $\mathbf{u}$ and fluid pressure p. To close the system, we prescribe homogeneous Dirichlet conditions on the boundary, i.e.,

$$T = 0, \quad \mathbf{u} = 0, \quad \text{and} \quad p = 0, \quad \text{on } \Gamma \times J, \tag{2.1d}$$

and we assume the following initial conditions

$$T(\cdot, 0) = T_0, \quad \mathbf{u}(\cdot, 0) = \mathbf{u}_0, \quad \text{and} \quad p(\cdot, 0) = p_0, \quad \text{in } \Omega \times \{0\}, \tag{2.1e}$$

for some known functions T_0, $\mathbf{u}_0$ and p_0. In practice, we may use nonhomogeneous Dirichlet and Neumann boundary conditions for which the analysis remains valid. Note also that if $\beta = b_0 = 0$, the above system decouples from the energy equation, and the well-known quasi-static Biot equations are recovered (see e.g. [1] where both the two- and four-field formulations are presented).

2.1. Preliminaries

We now define the function spaces that will be used throughout this article, see e.g. [15,38] for more details. For $1 \leq p < \infty$ let $L^p(\Omega) = \{u : \Omega \to \mathbb{R} : \int_\Omega |u|^p \mathrm{d}x < \infty\}$, with the associated norm $\|\cdot\|_p$. In particular, $L^2(\Omega)$ is the Hilbert space of square integrable functions defined on Ω, endowed with the inner product $(\cdot, \cdot)$, and the norm $\|\cdot\| := \|\cdot\|_2$. For $p = \infty$, $L^\infty(\Omega)$ is the space of uniformly bounded measurable functions defined on Ω, i.e. $L^\infty(\Omega) = \{u : \Omega \to \mathbb{R} : \operatorname{ess\,sup}_{x \in \Omega} |u| < \infty\}$, endowed with the norm $\|u\|_\infty = \inf\{C : |u| \leq C \text{ a.e. on } \Omega\}$. We denote by $W^{k,p}(\Omega)$ the Sobolev space of functions in $L^p(\Omega)$, admitting weak derivatives up to order k in the same space. In particular, we denote by $H^1(\Omega) := W^{1,2}(\Omega) = \{u \in L^2(\Omega) : \nabla u \in (L^2(\Omega))^d\}$, and designate by $H_0^1(\Omega)$ its zero-trace subspace. Let $H(\mathrm{div}, \Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$ be the space of vector valued functions, where each component belongs to $L^2(\Omega)$, along with the weak divergence. We endow this space with the norm $\|\mathbf{v}\|_{H(\mathrm{div};\Omega)}^2 := \|\mathbf{v}\|^2 + \|\nabla \cdot \mathbf{v}\|^2$. Let $H_s(\mathrm{div}, \Omega) = \{\boldsymbol{\tau} \in (L^2(\Omega))^{d \times d} : \nabla \cdot \boldsymbol{\tau} \in (L^2(\Omega))^d, \boldsymbol{\tau}_{ij} = \boldsymbol{\tau}_{ji} \text{ for } 1 \leq i, j \leq d\}$ be the space of symmetric tensor valued functions defined on Ω, where each component belongs to $L^2(\Omega)$, and admitting a weak divergence in $(L^2(\Omega))^d$. We denote by $C^1(\Omega)$ the space of continuous functions defined on Ω, admitting continuous partial derivatives. Finally, let X be a Banach space and let $L^p(J; X)$ be the Bochner space of functions in L^p defined on J with values in X. Let $\|\cdot\|_X$ be a norm on X, then for $u \in L^p(J; X)$, $p < \infty$, we have $\|u\|_{L^p(J;X)}^p := \int_0^{T_f} \|u(t)\|_X^p \, \mathrm{d}t$. In particular, we will make use of the spaces $H^1(J; L^2(\Omega)) = \{u(t) : \Omega \to \mathbb{R} : \int_0^{T_f} (\|u(t)\|^2 + \|\partial_t u(t)\|^2) \mathrm{d}t < \infty\}$ and $L^\infty(J; L^2(\Omega)) = \{u(t) : \Omega \to \mathbb{R} : \operatorname{ess\,sup}_{t \in J} \|u(t)\| < \infty\}$. Note that if $\mathbf{u}(t) \in (L^2(\Omega))^d$ is square integrable in time, we shall still write $\mathbf{u} \in L^2(J; L^2(\Omega))$, but this should not cause any confusion as we will always utilize bold fonts for vector (or tensor) valued functions.

We will also frequently apply classical inequalities, i.e. Cauchy–Schwarz (C–S), Young, and Grönwall (see e.g. [17]).

2.2. Assumptions on the data

Before transcribing the mixed variational formulation of the problem (2.1), we make precise the assumptions on the data (further generalizations are possible, bringing more technicalities):

Assumption 1 *(Data).*

A.1 The source terms are such that $g, h \in L^2(J; L^2(\Omega))$, *and* $\mathbf{f} \in H^1(J; L^2(\Omega))$.
A.2 The initial conditions are such that $p_0, T_0 \in H_0^1(\Omega)$, *and* $\mathbf{u}_0 \in (L^2(\Omega))^d$.
A.3 The permeability and heat conductivity tensors are such that $\mathbf{K}, \boldsymbol{\Theta} \in (L^\infty(\Omega))^{d\times d}$. *Furthermore, we assume there exists* $k_M, k_m > 0$ *such that for a.e.* $x \in \Omega$ *there holds*

$$k_m|\zeta|^2 \leq \zeta^T \mathbf{K}^{-1}(x)\zeta \text{ and } |\mathbf{K}^{-1}(x)\zeta| \leq k_M|\zeta|, \ \forall \zeta \in \mathbb{R}^d \setminus \{0\},$$

and there exists $\theta_M, \theta_m > 0$ *such that for a.e.* $x \in \Omega$ *there holds*

$$\theta_m|\zeta|^2 \leq \zeta^T \boldsymbol{\Theta}^{-1}(x)\zeta \text{ and } |\boldsymbol{\Theta}^{-1}(x)\zeta| \leq \theta_M|\zeta|, \ \forall \zeta \in \mathbb{R}^d \setminus \{0\}.$$

A.4 The constants $c_0, b_0, a_0, \alpha, \beta, \mu$, *and* λ, *are strictly positive.*

2.3. Mixed variational formulation

We now give the mixed variational formulation of the problem (2.1), for which we need to introduce the total stress tensor; $\boldsymbol{\sigma}(\mathbf{u}, p, T) := 2\mu\boldsymbol{\varepsilon}(\mathbf{u}) + \lambda\nabla\cdot\mathbf{u}\mathbf{I} - \alpha p\mathbf{I} - \beta T\mathbf{I}$, where $\mathbf{I}$ is the identity tensor and $\boldsymbol{\varepsilon}(\mathbf{u})$ is the linearized strain tensor given by $\boldsymbol{\varepsilon}(\mathbf{u}) := (\nabla\mathbf{u} + \nabla^{\mathrm{T}}\mathbf{u})/2$, the Darcy flux $\mathbf{w} := -\mathbf{K}\nabla p$, and the heat flux $\mathbf{r} := -\boldsymbol{\Theta}\nabla T$. For simplicity, we now restrict our attention to the case $d = 2$, in which case the fourth order compliance tensor, $\mathcal{A}$, is given by

$$\mathcal{A}\boldsymbol{\tau} := \frac{1}{2\mu}\left(\tau - \frac{\lambda}{2(\mu+\lambda)}\operatorname{tr}(\boldsymbol{\tau})\mathbf{I}\right), \quad \boldsymbol{\tau} \in \mathbb{R}^{d\times d}, \tag{2.2}$$

as seen in [37] (see also [20] for the general formula). Note that $\mathcal{A}$ is bounded and symmetric positive definite uniformly with respect to $x \in \Omega$, and defines an L^2-equivalent norm, i.e.

$$\frac{1}{2(\mu+\lambda)}\|\boldsymbol{\tau}\|^2 \leq \|\boldsymbol{\tau}\|_{\mathcal{A}}^2 \leq \frac{1}{2\mu}\|\boldsymbol{\tau}\|^2, \qquad \forall \boldsymbol{\tau} \in \left(L^2(\Omega)\right)^{d\times d}, \tag{2.3}$$

where $\|\boldsymbol{\tau}\|_{\mathcal{A}}^2 = \int_\Omega \mathcal{A}\boldsymbol{\tau} : \boldsymbol{\tau}\mathrm{d}x$. Applying $\mathcal{A}$ to the total stress tensor, it is inferred that

$$\mathcal{A}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\mathbf{u}) - \frac{1}{2(\mu+\lambda)}(\alpha p + \beta T)\mathbf{I}, \tag{2.4}$$

and by taking the trace on both sides, we get the following relationship

$$\nabla\cdot\mathbf{u} = \frac{1}{2(\mu+\lambda)}\operatorname{tr}(\boldsymbol{\sigma}) + \frac{1}{\mu+\lambda}(\alpha p + \beta T). \tag{2.5}$$

We also introduce the following notation

$$c_r := \frac{\alpha^2}{\mu+\lambda}, \qquad b_r := b_0 - \frac{\alpha\beta}{\mu+\lambda}, \qquad a_r := \frac{\beta^2}{\mu+\lambda}. \tag{2.6}$$

The above definitions yields an equivalent mixed form to (2.1):

$$\partial_t(a_0 T - b_0 p + \beta\nabla\cdot\mathbf{u}) + \nabla T\cdot\mathbf{w} + \nabla\cdot\mathbf{r} = h, \quad \text{in } Q, \tag{2.7a}$$

$$\boldsymbol{\Theta}^{-1}\mathbf{r} + \nabla T = 0, \quad \text{in } Q, \tag{2.7b}$$

$$\partial_t(c_0 p - b_0 T + \alpha \nabla \cdot \mathbf{u}) + \nabla \cdot \mathbf{w} = g, \quad \text{in } Q, \tag{2.7c}$$

$$\mathbf{K}^{-1}\mathbf{w} + \nabla p = 0, \quad \text{in } Q, \tag{2.7d}$$

$$\mathcal{A}\boldsymbol{\sigma} - \boldsymbol{\varepsilon}(\mathbf{u}) + \frac{c_r}{2\alpha}\mathbf{I}p + \frac{a_r}{2\beta}\mathbf{I}T = 0, \quad \text{in } Q, \tag{2.7e}$$

$$-\nabla \cdot \boldsymbol{\sigma} = \mathbf{f}, \quad \text{in } Q. \tag{2.7f}$$

We now set

$$\mathcal{T} := L^2(\Omega), \quad \mathcal{R} := H(\mathrm{div}, \Omega), \quad \mathcal{P} := L^2(\Omega), \quad \mathcal{W} := H(\mathrm{div}, \Omega), \quad \mathcal{S} := H_s(\mathrm{div}, \Omega), \quad \mathcal{U} := (L^2(\Omega))^d.$$

The following mixed variational formulation of the problem (2.1) can be obtained by multiplying by adequate test functions and then integrating by parts: find $(T(t), \mathbf{r}(t), p(t), \mathbf{w}(t), \boldsymbol{\sigma}(t), \mathbf{u}(t)) \in \mathcal{T} \times \mathcal{R} \times \mathcal{P} \times \mathcal{W} \times \mathcal{S} \times \mathcal{U}$, such that a.e. for $t \in J$ there holds

$$(a_0 + a_r)(\partial_t T, S) - b_r(\partial_t p, S) + \frac{a_r}{2\beta}(\partial_t \boldsymbol{\sigma}, S\mathbf{I}) + (\boldsymbol{\Theta}^{-1}\mathbf{r} \cdot \mathbf{w}, S) + (\nabla \cdot \mathbf{r}, S) = (h, S), \forall S \in \mathcal{T}, \tag{2.8a}$$

$$(\boldsymbol{\Theta}^{-1}\mathbf{r}, \mathbf{y}) - (T, \nabla \cdot \mathbf{y}) = 0, \quad \forall \mathbf{y} \in \mathcal{R}, \tag{2.8b}$$

$$(c_0 + c_r)(\partial_t p, q) - b_r(\partial_t T, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}, q\mathbf{I}) + (\nabla \cdot \mathbf{w}, q) = (g, q), \ \forall q \in \mathcal{P}, \tag{2.8c}$$

$$(\mathbf{K}^{-1}\mathbf{w}, \mathbf{z}) - (p, \nabla \cdot \mathbf{z}) = 0, \quad \forall \mathbf{z} \in \mathcal{W}, \tag{2.8d}$$

$$(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\mathbf{u}, \nabla \cdot \boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}p, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}T, \boldsymbol{\tau}) = 0, \quad \forall \boldsymbol{\tau} \in \mathcal{S}, \tag{2.8e}$$

$$-(\nabla \cdot \boldsymbol{\sigma}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \ \forall \mathbf{v} \in \mathcal{U}, \tag{2.8f}$$

and such that the initial conditions (2.1e) holds true in the weak sense, i.e.

$$(p(0), q) = (p_0, q) \quad \forall q \in \mathcal{P}, \quad (\mathbf{u}(0), \mathbf{v}) = (\mathbf{u}_0, \mathbf{v}) \quad \forall \mathbf{v} \in \mathcal{U}, \quad \text{and} \quad (T(0), S) = (T_0, S) \quad \forall S \in \mathcal{T}. \tag{2.8g}$$

Remark 2.1. Note that a different variational formulation of the problem (2.7) is possible, using a weakly symmetric space for the stress tensor. This formulation will then involve a new variable acting as a Lagrange multiplier which is enforcing the symmetry of the stress (see e.g. [2,4,20]). For simplicity of presentation we shall keep the formulation (2.8) throughout. The analysis presented next can nevertheless also be extended to the previously mentioned formulation using the same techniques, as done in [1] for the four-field Biot equations.

Remark 2.2. The nonlinear coupling in the above problem makes the analysis difficult. The next section is therefore devoted to analyzing a linearized problem, the results from which will be helpful when analyzing the full nonlinear problem in the last section. We mention also that other nonlinearities can be added, e.g. nonlinear compressibility or nonlinear Lamé parameters.

3. Analysis of the linear problem

In this section we introduce a linear version of the problem (2.8). Precisely, we replace the convective transport term $(\boldsymbol{\Theta}^{-1}\mathbf{r} \cdot \mathbf{w}, S)$ in the energy balance equation (2.8a), by $-(\boldsymbol{\eta} \cdot \mathbf{w}, S)$, for some given $\boldsymbol{\eta} \in L^\infty(\Omega)$. We denote by $\gamma := \|\boldsymbol{\eta}\|_\infty$. We introduce the resulting linear problem which reads: find $(T(t), \mathbf{r}(t), p(t), \mathbf{w}(t), \boldsymbol{\sigma}(t), \mathbf{u}(t)) \in \mathcal{T} \times \mathcal{R} \times \mathcal{P} \times \mathcal{W} \times \mathcal{S} \times \mathcal{U}$, such that for a.e. $t \in J$ there holds

$$(a_0 + a_r)(\partial_t T, S) - b_r(\partial_t p, S) + \frac{a_r}{2\beta}(\partial_t \boldsymbol{\sigma}, S\mathbf{I}) - (\boldsymbol{\eta} \cdot \mathbf{w}, S) + (\nabla \cdot \mathbf{r}, S) = (h, S), \qquad \forall S \in \mathcal{T}, \quad (3.1a)$$

$$(\boldsymbol{\Theta}^{-1}\mathbf{r}, \mathbf{y}) - (T, \nabla \cdot \mathbf{y}) = 0, \qquad \forall \mathbf{y} \in \mathcal{R}, \quad (3.1b)$$

$$(c_0 + c_r)(\partial_t p, q) - b_r(\partial_t T, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}, q\mathbf{I}) + (\nabla \cdot \mathbf{w}, q) = (g, q), \qquad \forall q \in \mathcal{P}, \quad (3.1c)$$

$$(\mathbf{K}^{-1}\mathbf{w}, \mathbf{z}) - (p, \nabla \cdot \mathbf{z}) = 0, \qquad \forall \mathbf{z} \in \mathcal{W}, \quad (3.1d)$$

$$(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\mathbf{u}, \nabla \cdot \boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}p, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}T, \boldsymbol{\tau}) = 0, \qquad \forall \boldsymbol{\tau} \in \mathcal{S}, \quad (3.1e)$$

$$-(\nabla \cdot \boldsymbol{\sigma}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \qquad \forall \mathbf{v} \in \mathcal{U}, \quad (3.1f)$$

and such that initial conditions (2.8g) holds true. The remaining part of this section is devoted to proving the well-posedness of this system. In what follows, we assume the following hypothesis on the effective thermal capacity a_0, the thermal dilation coefficient b_0, the specific storage coefficient c_0 and the Lamé parameters μ, λ;

$$b_0 - \frac{\alpha\beta}{\mu + \lambda} > 0, \qquad c_0 - \frac{c_r}{2} - b_0 - \frac{1}{6(\mu + \lambda)} > 0, \qquad a_0 - \frac{a_r}{2} - b_0 - \frac{1}{6(\mu + \lambda)} > 0. \quad (3.2)$$

These constraints are typically needed in order to ensure a gradient flow structure. Similar constraints were used to analyze the Biot equations in mixed form in [1]. We also refer the reader to [21] for a more detailed discussion about the scaling of Biot's (isothermal) equations. However, compared to these works, our constraints involve also the thermal coefficients. We omit any further discussion on the justification for these constraints, other than they are necessary to prove the results we present. The well-posedness of problem (3.1) is then given in the following result.

Theorem 3.1 *(Well-posedness of the linear problem). Under Assumption 1, the problem* (3.1), (2.8g) *has a unique solution*

$$(T, \mathbf{r}) \in H^1(J; L^2(\Omega)) \times \left(L^2(J; H(\mathrm{div}; \Omega)) \cap L^\infty(J; L^2(\Omega))\right), \quad (3.3a)$$

$$(p, \mathbf{w}) \in H^1(J; L^2(\Omega)) \times \left(L^2(J; H(\mathrm{div}; \Omega)) \cap L^\infty(J; L^2(\Omega))\right), \quad (3.3b)$$

$$(\mathbf{u}, \boldsymbol{\sigma}) \in H^1(J; L^2(\Omega)) \times \left(L^2(J; H_s(\mathrm{div}; \Omega)) \cap H^1(J; L^2(\Omega))\right). \quad (3.3c)$$

Moreover, if $g, h \in H^1(J; L^2(\Omega))$, $\mathbf{f} \in H^2(J; L^2(\Omega))$ *then*

$$(T, \mathbf{r}) \in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H(\mathrm{div}; \Omega)) \cap H^1(J; L^2(\Omega))\right), \quad (3.4a)$$

$$(p, \mathbf{w}) \in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H(\mathrm{div}; \Omega)) \cap H^1(J; L^2(\Omega))\right), \quad (3.4b)$$

$$(\mathbf{u}, \boldsymbol{\sigma}) \in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H_s(\mathrm{div}; \Omega)) \cap W^{1,\infty}(J; L^2(\Omega))\right). \quad (3.4c)$$

The proof will follow from a series of partial results to be done in the sequel. The analysis uses a Galerkin method together with the theory of differential algebraic equations (DAEs), as well as weak compactness arguments (cf. [1,37,26,15]).

3.1. Construction of approximate solutions

In order to employ Galerkin's method we introduce a finite dimensional approximation of the problem (3.1). We need to introduce the following finite dimensional subspaces. Let $(i, j, k, l, m, n) \in \mathbb{N}^6$ be fixed and strictly positive, and let $\mathcal{T}_i := \mathrm{span}\{S_\ell \in \mathcal{T} : \ell = 1, \cdots, i\}$, $\mathcal{R}_j := \mathrm{span}\{\mathbf{y}_\ell \in \mathcal{R} : \ell = 1, \cdots, j\}$,

$\mathcal{P}_k := \text{span}\{q_\ell \in \mathcal{P} : \ell = 1, \cdots, k\}$, $\mathcal{W}_l := \text{span}\{\mathbf{z}_\ell \in \mathcal{W} : \ell = 1, \cdots, l\}$, $\mathcal{S}_m := \text{span}\{\boldsymbol{\tau}_\ell \in \mathcal{S} : \ell = 1, \cdots, m\}$ and $\mathcal{U}_n := \text{span}\{\mathbf{v}_\ell \in \mathcal{U} : \ell = 1, \cdots, n\}$, where the functions $S_\ell, \mathbf{y}_\ell, q_\ell, \mathbf{z}_\ell, \boldsymbol{\tau}_\ell$ and $\mathbf{v}_\ell$, for $\ell \in \mathbb{N}$, constitute Hilbert bases for the spaces $\mathcal{T}, \mathcal{R}, \mathcal{P}, \mathcal{W}, \mathcal{S}$ and $\mathcal{U}$, respectively. Let now $(T_i, \mathbf{r}_j, p_k, \mathbf{w}_l, \boldsymbol{\sigma}_m, \mathbf{u}_n) : [0, T_f]^6 \to \mathcal{T}_i \times \mathcal{R}_j \times \mathcal{P}_k \times \mathcal{W}_l \times \mathcal{S}_m \times \mathcal{U}_n$ be the solution to the following problem:

$$(a_0 + a_r)(\partial_t T_i, S_\ell) - b_r(\partial_t p_k, S_\ell) + \frac{a_r}{2\beta}(\partial_t \boldsymbol{\sigma}_m, S_\ell \mathbf{I})$$

$$-(\boldsymbol{\eta} \cdot \mathbf{w}_l, S_\ell) + (\nabla \cdot \mathbf{r}_j, S_\ell) = (h, S_\ell), \qquad \ell = 1, \cdots, i, \quad (3.5a)$$

$$(\boldsymbol{\Theta}^{-1} \mathbf{r}_j, \mathbf{y}_\ell) - (T_i, \nabla \cdot \mathbf{y}_\ell) = 0, \qquad \ell = 1, \cdots, j, \quad (3.5b)$$

$$(c_0 + c_r)(\partial_t p_k, q_\ell) - b_r(\partial_t T_i, q_\ell) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}_m, q_\ell \mathbf{I}) + (\nabla \cdot \mathbf{w}_l, q_\ell) = (g, q_\ell), \qquad \ell = 1, \cdots, k, \quad (3.5c)$$

$$(\mathbf{K}^{-1} \mathbf{w}_l, \mathbf{z}_\ell) - (p_k, \nabla \cdot \mathbf{z}_\ell) = 0, \qquad \ell = 1, \cdots, l, \quad (3.5d)$$

$$(\mathcal{A}\boldsymbol{\sigma}_m, \boldsymbol{\tau}_\ell) + (\mathbf{u}_n, \nabla \cdot \boldsymbol{\tau}_\ell) + \frac{c_r}{2\alpha}(\mathbf{I} p_k, \boldsymbol{\tau}_\ell) + \frac{a_r}{2\beta}(\mathbf{I} T_i, \boldsymbol{\tau}_\ell) = 0, \qquad \ell = 1, \cdots, m, \quad (3.5e)$$

$$-(\nabla \cdot \boldsymbol{\sigma}_m, \mathbf{v}_\ell) = (\mathbf{f}, \mathbf{v}_\ell), \qquad \ell = 1, \cdots, n. \quad (3.5f)$$

We introduce the coefficient vectors of the solutions: let $\mathbf{T}_i(t) := [T_1(t), \cdots, T_i(t)]^T$ where $T_i(x,t) = \sum_{\ell=1}^{i} T_\ell(t) S_\ell$, $\mathbf{R}_j(t) := [r_1(t), \cdots, r_j(t)]^T$ where $\mathbf{r}_j(x,t) = \sum_{\ell=1}^{j} r_\ell(t) \mathbf{y}_\ell$, $\mathbf{P}_k(t) := [p_1(t), \cdots, p_k(t)]^T$ where $p_k(x,t) = \sum_{\ell=1}^{k} p_\ell(t) q_\ell$, $\mathbf{W}_l(t) := [w_1(t), \cdots, w_l(t)]^T$ where $\mathbf{w}_l(x,t) = \sum_{\ell=1}^{l} w_\ell(t) \mathbf{z}_\ell$, $\boldsymbol{\Sigma}_m(t) := [\sigma_1(t), \cdots, \sigma_m(t)]^T$ where $\boldsymbol{\sigma}_m(x,t) = \sum_{\ell=1}^{m} \sigma_\ell(t) \boldsymbol{\tau}_\ell$ and $\mathbf{U}_n(t) := [u_1(t), \cdots u_n(t)]^T$ where $\mathbf{u}_n(x,t) = \sum_{\ell=1}^{n} u_\ell(t) \mathbf{v}_\ell$.

Thus, we impose the initial conditions by

$$T_\ell(0) = (T_0, S_\ell),\ 1 \le \ell \le i, \quad u_\ell(0) = (\mathbf{u}_0, \mathbf{v}_\ell),\ 1 \le \ell \le n, \quad p_\ell(0) = (p_0, q_\ell),\ 1 \le \ell \le k. \quad (3.5g)$$

We also define the following linear operators: $(\mathbf{A}_{\boldsymbol{\sigma}\boldsymbol{\sigma}})_{\imath\jmath} := (\mathcal{A}\boldsymbol{\tau}_\imath, \boldsymbol{\tau}_\jmath)$, for $1 \le \imath, \jmath \le m$, $(\mathbf{A}_{pp})_{\imath\jmath} := (c_0 + c_r)(q_\imath, q_\jmath)$, for $1 \le \imath, \jmath \le k$, $(\mathbf{A}_{TT})_{\imath\jmath} := (a_0 + a_r)(S_\imath, S_\jmath)$, for $1 \le \imath, \jmath \le i$, $(\mathbf{A}_{\mathbf{ww}})_{\imath\jmath} := (\mathbf{K}^{-1} \mathbf{z}_\imath, \mathbf{z}_\jmath)$, for $1 \le \imath, \jmath \le l$, $(\mathbf{A}_{\mathbf{rr}})_{\imath\jmath} := (\boldsymbol{\Theta}^{-1} \mathbf{y}_\imath, \mathbf{y}_\jmath)$, for $1 \le \imath, \jmath \le j$, $(\mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}})_{\imath\jmath} := (\mathbf{v}_\imath, \nabla \cdot \boldsymbol{\tau}_\jmath)$, for $1 \le \imath \le n, 1 \le \jmath \le m$, $(\mathbf{A}_{p\boldsymbol{\sigma}})_{\imath\jmath} := \frac{c_r}{2\alpha}(\mathbf{I} q_\imath, \boldsymbol{\tau}_\jmath)$, for $1 \le \imath \le k, 1 \le \jmath \le m$, $(\mathbf{A}_{T\boldsymbol{\sigma}})_{\imath\jmath} := \frac{a_r}{2\beta}(\mathbf{I} S_\imath, \boldsymbol{\tau}_\jmath)$, for $1 \le \imath \le i, 1 \le \jmath \le m$, $(\mathbf{A}_{Tp})_{\imath\jmath} := -b_r(S_\imath, q_\jmath)$, for $1 \le \imath \le i, 1 \le \jmath \le k$, $(\mathbf{A}_{\mathbf{w}p})_{\imath\jmath} := (\nabla \cdot \mathbf{z}_\imath, q_\jmath)$, $1 \le \imath \le l, 1 \le \jmath \le k$, $(\mathbf{A}_{\mathbf{r}T})_{\imath\jmath} := (\nabla \cdot \mathbf{y}_\imath, S_\jmath)$, for $1 \le \imath \le l, 1 \le \jmath \le i$, and $(\mathbf{A}_{\mathbf{w}T})_{\imath\jmath} := (\boldsymbol{\eta} \cdot \mathbf{z}_\imath, S_\jmath)$, for $1 \le \imath \le l, 1 \le \jmath \le i$.

Finally, we define the vectors: $(\mathbf{L}_1)_\ell := (\mathbf{f}, \mathbf{v}_\ell)$, for $1 \le \ell \le n$, $(\mathbf{L}_2)_\ell := (g, q_\ell)$, for $1 \le \ell \le k$ and $(\mathbf{L}_3)_\ell := (h, S_\ell)$, for $1 \le \ell \le i$. We rewrite using the above notation the problem (3.5) as a system of ODEs

$$\mathbf{A}_{TT} \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{T}_i + \mathbf{A}_{Tp} \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{P}_k + \mathbf{A}_{T\boldsymbol{\sigma}} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\Sigma}_m - \mathbf{A}_{\mathbf{w}T} \mathbf{W}_l + \mathbf{A}_{\mathbf{r}T}^T \mathbf{R}_j = \mathbf{L}_3, \quad (3.6a)$$

$$\mathbf{A}_{\mathbf{r}} \mathbf{R}_j - \mathbf{A}_{\mathbf{r}T} \mathbf{T}_i = 0, \quad (3.6b)$$

$$\mathbf{A}_{pp} \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{P}_k + \mathbf{A}_{Tp}^T \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{T}_i + \mathbf{A}_{p\boldsymbol{\sigma}} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\Sigma}_m + \mathbf{A}_{\mathbf{w}p}^T \mathbf{W}_l = \mathbf{L}_2, \quad (3.6c)$$

$$\mathbf{A}_{\mathbf{w}} \mathbf{W}_l - \mathbf{A}_{\mathbf{w}p} \mathbf{P}_k = 0, \quad (3.6d)$$

$$\mathbf{A}_{\boldsymbol{\sigma}\boldsymbol{\sigma}} \boldsymbol{\Sigma}_m + \mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}}^T \mathbf{U}_n + \mathbf{A}_{p\boldsymbol{\sigma}}^T \mathbf{P}_k + \mathbf{A}_{T\boldsymbol{\sigma}}^T \mathbf{T}_i = 0, \quad (3.6e)$$

$$-\mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}} \boldsymbol{\Sigma}_m = \mathbf{L}_1. \quad (3.6f)$$

After rearranging, these ODE equations can be written in the form of a DAE system

$$\Phi \frac{\mathrm{d}}{\mathrm{d}t} X(t) + \Psi X(t) = L(t), \quad (3.7)$$

where $X(t) := (\mathbf{P}_k(t), \mathbf{\Sigma}_m(t), \mathbf{T}_i(t), \mathbf{W}_l(t), \mathbf{U}_n(t), \mathbf{R}_j(t))^T$, $L(t) := (\mathbf{L}_2(t), 0, \mathbf{L}_3(t), 0, \mathbf{L}_1(t), 0)^T$ and

$$\Phi := \begin{pmatrix} \mathbf{A}_{pp} & \mathbf{A}_{p\boldsymbol{\sigma}} & \mathbf{A}_{Tp}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{A}_{Tp} & \mathbf{A}_{T\boldsymbol{\sigma}} & \mathbf{A}_{TT} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{3.8}$$

and

$$\Psi := \begin{pmatrix} 0 & 0 & 0 & \mathbf{A}_{\mathbf{w}p}^T & 0 & 0 \\ \mathbf{A}_{p\boldsymbol{\sigma}}^T & \mathbf{A}_{\boldsymbol{\sigma}\boldsymbol{\sigma}} & \mathbf{A}_{T\boldsymbol{\sigma}}^T & 0 & \mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}}^T & 0 \\ 0 & 0 & 0 & -\mathbf{A}_{\mathbf{w}T} & 0 & \mathbf{A}_{\mathbf{r}T}^T \\ -\mathbf{A}_{\mathbf{w}p} & 0 & 0 & \mathbf{A}_{\mathbf{ww}} & 0 & 0 \\ 0 & -\mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\mathbf{A}_{\mathbf{r}T} & 0 & 0 & \mathbf{A}_{\mathbf{rr}} \end{pmatrix}. \tag{3.9}$$

From the theory of DAEs, equation (3.7) together with initial conditions (3.5g) has a solution if the matrix pencil, $s\Phi + \Psi$, is nonsingular for some $s \neq 0$ (see [9]). Note that we can write $s\Phi + \Psi$ as a block 2×2 matrix as follows

$$s\Phi + \Psi = \begin{pmatrix} A & B \\ -C & D \end{pmatrix},$$

where

$$A = \begin{pmatrix} s\mathbf{A}_{pp} & s\mathbf{A}_{p\boldsymbol{\sigma}} & s\mathbf{A}_{Tp}^T \\ \mathbf{A}_{p\boldsymbol{\sigma}}^T & \mathbf{A}_{\boldsymbol{\sigma}\boldsymbol{\sigma}} & \mathbf{A}_{T\boldsymbol{\sigma}}^T \\ s\mathbf{A}_{Tp} & s\mathbf{A}_{T\boldsymbol{\sigma}} & s\mathbf{A}_{TT} \end{pmatrix}, \qquad B = \begin{pmatrix} \mathbf{A}_{\mathbf{w}p}^T & 0 & 0 \\ 0 & \mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}}^T & 0 \\ -\mathbf{A}_{\mathbf{w}T} & 0 & \mathbf{A}_{\mathbf{r}T}^T \end{pmatrix},$$
$$C = \begin{pmatrix} \mathbf{A}_{\mathbf{w}p} & 0 & 0 \\ 0 & \mathbf{A}_{\mathbf{u}\boldsymbol{\sigma}} & 0 \\ & 0 & \mathbf{A}_{\mathbf{r}T} \end{pmatrix}, \qquad D = \begin{pmatrix} \mathbf{A}_{\mathbf{ww}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{Arr} \end{pmatrix}.$$

Let $\mathcal{B} = \mathcal{S}_m \times \mathcal{P}_k \times \mathcal{T}_i$ and $\mathcal{C} = \mathcal{U}_n \times \mathcal{W}_l \times \mathcal{R}_j$, such that the bilinear form associated with $s\Phi + \Psi$ can be decomposed into the bilinear forms associated with each block, i.e. $\phi_A : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$, $\phi_B : \mathcal{C} \times \mathcal{B} \to \mathbb{R}$, $\phi_C : \mathcal{B} \times \mathcal{C} \to \mathbb{R}$, and $\phi_D : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$, where

$$\begin{aligned} \phi_A((\boldsymbol{\sigma}_m, p_k, T_i), (\boldsymbol{\tau}, q, S)) := {} & s(c_0 + c_r)(p_k, q) + \frac{c_r}{2\alpha}(\mathbf{I}p_k, \boldsymbol{\tau}) + s\frac{c_r}{2\alpha}(\boldsymbol{\sigma}_m, q\mathbf{I}) - sb_r(p_k, S) \\ & - sb_r(T_i, q) + (\mathcal{A}\boldsymbol{\sigma}_m, \boldsymbol{\tau}) + s\frac{a_r}{2\beta}(\boldsymbol{\sigma}_m, S\mathbf{I}) \\ & + \frac{a_r}{2\beta}(\mathbf{I}T_i, \boldsymbol{\tau}) + s(a_0 + a_r)(T_i, S), \end{aligned} \tag{3.10a}$$

$$\phi_B((\boldsymbol{\tau}, q, S), (\mathbf{u}_n, \mathbf{w}_l, \mathbf{r}_j)) := (\nabla \cdot \mathbf{w}_l, q) + (\mathbf{u}_n, \nabla \cdot \boldsymbol{\tau}) - (\boldsymbol{\eta} \cdot \mathbf{w}_l, S) + (\nabla \cdot \mathbf{r}_j, S), \tag{3.10b}$$

$$\phi_C((\boldsymbol{\sigma}_m, p_k, T_i), (\mathbf{v}, \mathbf{z}, \mathbf{y})) := (p_k, \nabla \cdot \mathbf{z}) + (\nabla \cdot \boldsymbol{\sigma}_m, \mathbf{v}) + (T_i, \nabla \cdot \mathbf{y}), \tag{3.10c}$$

$$\phi_D((\mathbf{u}_n, \mathbf{w}_l, \mathbf{r}_j), (\mathbf{v}, \mathbf{z}, \mathbf{y})) := (\mathbf{K}^{-1}\mathbf{w}_l, \mathbf{z}) + (\boldsymbol{\Theta}^{-1}\mathbf{r}_j, \mathbf{y}). \tag{3.10d}$$

The following Lemma will imply the invertibility of $s\Phi + \Psi$ for some $s \neq 0$.

Lemma 3.2. *For any tuple $(i, j, k, l, m, n) \geq 1$, there exists an $s \neq 0$ such that the bilinear form associated with $s\Phi + \Psi$ is strictly positive, i.e.*

$$\phi_A + \phi_B - \phi_C + \phi_D > 0,$$

for all nonzero $(\boldsymbol{\tau}, q, S) \in \mathcal{B}$ *and* $(\mathbf{v}, \mathbf{z}, \mathbf{y}) \in \mathcal{C}$.

Proof. Denoting by $\boldsymbol{\tau} = \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix}$, and using the definition of the compliance tensor (2.2), together with the C–S, Young, and triangle inequalities yields

$$\begin{aligned}
&\phi_A((\boldsymbol{\tau}, q, S), (\boldsymbol{\tau}, q, S)) + \phi_B((\mathbf{v}, \mathbf{z}, \mathbf{y}), (\tau, q, S)) - \phi_C((\tau, q, S), (\mathbf{v}, \mathbf{z}, \mathbf{y})) + \phi_D((\mathbf{v}, \mathbf{z}, \mathbf{y}), (\mathbf{v}, \mathbf{z}, \mathbf{y})) \\
&\quad = s(c_0 + c_r) \|q\|^2 + s(a_0 + a_r) \|S\|^2 + (1+s)\frac{c_r}{2\alpha}(\mathbf{I}q, \boldsymbol{\tau}) - 2sb_r(q, S) + (1+s)\frac{a_r}{2\beta}(\boldsymbol{\tau}, S\mathbf{I}) \\
&\qquad + (\mathcal{A}\boldsymbol{\tau}, \boldsymbol{\tau}) - (\boldsymbol{\eta} \cdot \mathbf{z}, S) + (\mathbf{K}^{-1}\mathbf{z}, \mathbf{z}) + (\boldsymbol{\Theta}^{-1}\mathbf{y}, \mathbf{y}) \\
&\quad \geq \left(s(c_0 + c_r - b_r) - (1+s)\frac{c_r}{2\alpha}\frac{\epsilon_1}{2} \right) \|q\|^2 + \left(s(a_0 + a_r - b_r) - (1+s)\frac{a_r}{2\beta}\frac{\epsilon_2}{2} - \frac{\gamma}{2k_m} \right) \|S\|^2 \\
&\qquad + \left(\frac{1}{2(\mu+\lambda)} - (1+s)\frac{c_r}{2\alpha}\frac{1}{2\epsilon_1} - (1+s)\frac{a_r}{2\beta}\frac{1}{2\epsilon_2} \right) \left(\|\boldsymbol{\tau}_{11}\|^2 + \|\boldsymbol{\tau}_{22}\|^2 \right) \\
&\qquad + \theta_m \|\mathbf{y}\|^2 + \frac{k_m}{2} \|\mathbf{z}\|^2 + \frac{1}{\mu} \|\boldsymbol{\tau}_{12}\|^2 .
\end{aligned} \tag{3.11}$$

What remains is to show if there exist parameters ϵ_1, ϵ_2, and s such that the following six constraints are satisfied

$$0 \leq s(c_0 + c_r - b_r) - (1+s)\frac{c_r}{2\alpha}\frac{\epsilon_1}{2}, \tag{3.12a}$$

$$0 \leq s(a_0 + a_r - b_r) - (1+s)\frac{a_r}{2\beta}\frac{\epsilon_2}{2} - \frac{\gamma}{2k_m}, \tag{3.12b}$$

$$0 \leq \frac{1}{2(\mu+\lambda)} - (1+s)\frac{c_r}{2\alpha}\frac{1}{2\epsilon_1} - (1+s)\frac{a_r}{2\beta}\frac{1}{2\epsilon_2}, \tag{3.12c}$$

$$0 < \epsilon_1, \epsilon_2, \quad \text{and} \quad s \neq 0. \tag{3.12d}$$

It is easily verified that the following choices are satisfactory; $s = -2, \epsilon_1 = \dfrac{4\alpha}{c_r(1+s)} s(c_0 + c_r - b_r)$, and $\epsilon_2 = \dfrac{4\beta}{a_r(1+s)} \left(s(a_0 + a_r - b_r) - \dfrac{\gamma}{2k_m} \right)$. We use these choices in (3.11), and letting $\xi := 1 + \dfrac{1}{16(\mu+\lambda)(c_0 + c_r - b_r)} + \dfrac{1}{16(\mu+\lambda)(a_0 + a_r - b_r + \gamma/(2k_m))} > 0$, it is inferred that

$$\begin{aligned}
&\phi_A((\boldsymbol{\tau}, q, S), (\boldsymbol{\tau}, q, S)) + \phi_B((\mathbf{v}, \mathbf{z}, \mathbf{y}), (\tau, q, S)) - \phi_C((\tau, q, S), (\mathbf{v}, \mathbf{z}, \mathbf{y})) + \phi_D((\mathbf{v}, \mathbf{z}, \mathbf{y}), (\mathbf{v}, \mathbf{z}, \mathbf{y})) \\
&\quad \geq \frac{\xi}{2(\mu+\lambda)} \left(\|\boldsymbol{\tau}_{11}\|^2 + \|\boldsymbol{\tau}_{22}\|^2 \right) + \frac{k_m}{2} \|\mathbf{z}\|^2 + \theta_m \|\mathbf{y}\|^2 + \frac{1}{\mu} \|\tau_{12}\|^2 > 0,
\end{aligned} \tag{3.13}$$

for all nonzero $(\boldsymbol{\tau}, q, S) \in \mathcal{B}$, $(\mathbf{v}, \mathbf{z}, \mathbf{y}) \in \mathcal{C}$. Thus, there exists an $s \neq 0$ such that $s\Phi + \Psi$ is nonsingular, and the equation (3.7) has a solution. □

3.2. A priori estimates

In this section, we derive *a priori* estimates for the unknowns which will allow us to pass to the limit in problem (3.5) by weak compactness arguments [38,15]. Throughout this section we denote by $C > 0$ a generic positive constant which may change value from one line to the next, but it will always be independent

of the relevant parameters, i.e. of the tuple (i, j, k, l, m, n). We summarize these estimates in the following theorem.

Theorem 3.3 *(A priori estimates). Under the Assumption 1, there exists a constant $C > 0$, independent of $(i, j, k, l, m, n) \geq 1$, such that*

(i)

$$\begin{aligned}&\|p_k\|^2_{L^\infty(J;L^2(\Omega))} + \|T_i\|^2_{L^\infty(J;L^2(\Omega))} + \|\mathbf{w}_l\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{r}_j\|^2_{L^2(J;L^2(\Omega))} + \|\boldsymbol{\sigma}(0)\|^2_{\mathcal{A}}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right),\end{aligned}$$

(ii)

$$\begin{aligned}&\|\partial_t p_k\|^2_{L^2(J;L^2(\Omega))} + \|\partial_t T_i\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{w}_l\|^2_{L^\infty(J;L^2(\Omega))} + \|\mathbf{r}_j\|^2_{L^\infty(J;L^2(\Omega))}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right),\end{aligned}$$

(iii)

$$\begin{aligned}&\|\boldsymbol{\sigma}_m\|^2_{L^\infty(J;L^2(\Omega))} + \|\partial_t\boldsymbol{\sigma}_m\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{u}_n\|^2_{L^\infty(J;L^2(\Omega))} + \|\partial_t\mathbf{u}_n\|^2_{L^2(J;L^2(\Omega))}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right),\end{aligned}$$

(iv)

$$\begin{aligned}&\|\mathbf{w}_l\|^2_{L^2(J;H(\mathrm{div},\Omega))} + \|\mathbf{r}_j\|^2_{L^2(J;H(\mathrm{div},\Omega))} + \|\boldsymbol{\sigma}_m\|^2_{L^2(J;H_s(\mathrm{div},\Omega))}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right).\end{aligned}$$

Proof. By Thomas' Lemma [33] there exist $\tilde{\boldsymbol{\sigma}} \in H^1(J;\mathcal{S}_m)$ such that $-\nabla\cdot\tilde{\boldsymbol{\sigma}}(\cdot,t) = \mathbf{u}_n(\cdot,t)$ on Ω for $t \in J$, and with $\|\tilde{\boldsymbol{\sigma}}(t)\| \leq C\|\mathbf{u}_n(t)\|$. Thus, we set $\boldsymbol{\tau}_\ell = \tilde{\boldsymbol{\sigma}}(t)$ in (3.5e) and obtain

$$\begin{aligned}\|\mathbf{u}_n\|^2 &= -(\mathbf{u}_n, \nabla\cdot\tilde{\boldsymbol{\sigma}}) = (\mathcal{A}\boldsymbol{\sigma}_m, \tilde{\boldsymbol{\sigma}}) + \frac{c_r}{2\alpha}(\mathbf{I}p_k, \tilde{\boldsymbol{\sigma}}) + \frac{a_r}{2\beta}(\mathbf{I}T_i, \tilde{\boldsymbol{\sigma}})\\ &\leq \left(\frac{1}{2\mu}\|\boldsymbol{\sigma}_m\| + \frac{c_r}{2\alpha}\|p_k\| + \frac{a_r}{2\beta}\|T_i\|\right)\|\tilde{\boldsymbol{\sigma}}\|\\ &\leq \left(\frac{1}{2\mu}\|\boldsymbol{\sigma}_m\| + \frac{c_r}{2\alpha}\|p_k\| + \frac{a_r}{2\beta}\|T_i\|\right)C\|\mathbf{u}_n\|,\end{aligned} \tag{3.14a}$$

which implies

$$\|\mathbf{u}_n\|^2 \leq C\left(\|\boldsymbol{\sigma}_m\|^2 + \|p_k\|^2 + \|T_i\|^2\right). \tag{3.14b}$$

Next, we take $\boldsymbol{\tau}_\ell = \boldsymbol{\sigma}_m$ in (3.5e) and $\mathbf{v}_\ell = \mathbf{u}_n$ in (3.5f), and add the resulting equations together to obtain

$$\|\boldsymbol{\sigma}_m\|^2_{\mathcal{A}} = -\frac{c_r}{2\alpha}(\mathbf{I}p_k, \boldsymbol{\sigma}_m) - \frac{a_r}{2\beta}(\mathbf{I}T_i, \boldsymbol{\sigma}_m) + (\mathbf{f}, \mathbf{u}_n). \tag{3.15a}$$

Applying the C–S and Young inequalities together with the above estimate (3.14b) yields

$$\begin{aligned}\|\boldsymbol{\sigma}_m\|^2_{\mathcal{A}} &\leq \frac{c_r}{2\alpha}\left(\frac{1}{2\epsilon_1}\|p_k\|^2 + \frac{\epsilon_1}{2}\|\boldsymbol{\sigma}_m\|^2\right) + \frac{a_r}{2\beta}\left(\frac{1}{2\epsilon_2}\|T_i\|^2 + \frac{\epsilon_2}{2}\|\boldsymbol{\sigma}_m\|^2\right) + \frac{1}{2\epsilon_3}\|\mathbf{f}\|^2 + \frac{\epsilon_3}{2}\|\mathbf{u}_n\|^2\\ &\leq \left(\frac{\alpha}{2}\epsilon_1 + \frac{\beta}{2}\epsilon_2 + C(\mu+\lambda)\epsilon_3\right)\|\boldsymbol{\sigma}_m\|^2_{\mathcal{A}} + \left(\frac{c_r}{4\alpha\epsilon_1} + C\frac{\epsilon_3}{2}\right)\|p_k\|^2\\ &\quad + \left(\frac{a_r}{4\beta\epsilon_2} + C\frac{\epsilon_3}{2}\right)\|T_i\|^2 + \frac{1}{2\epsilon_3}\|\mathbf{f}\|^2.\end{aligned} \tag{3.15b}$$

Choosing suitable values for the epsilons, i.e. $\epsilon_1 = \dfrac{1}{3\alpha}$, $\epsilon_2 = \dfrac{1}{3\beta}$ and $\epsilon_3 = \dfrac{1}{6C(\mu+\lambda)}$, we obtain

$$\|\boldsymbol{\sigma}_m\|_{\mathcal{A}}^2 \leq \left(\frac{3}{2}c_r + \frac{1}{6(\mu+\lambda)}\right)\|p_k\|^2 + \left(\frac{3}{2}a_r + \frac{1}{6(\mu+\lambda)}\right)\|T_i\|^2 + C\,\|\mathbf{f}\|^2\,. \tag{3.15c}$$

It then follows immediately that

$$\|\mathbf{u}_n\|^2 \leq C\left(\|p_k\|^2 + \|T_i\|^2 + \|\mathbf{f}\|^2\right). \tag{3.16}$$

Take now $\tilde{\boldsymbol{\sigma}} \in L^2(J;\mathcal{S}_m)$ such that $-\nabla\cdot\tilde{\boldsymbol{\sigma}}(\cdot,t) = \partial_t\mathbf{u}_n(\cdot,t)$ on Ω, for $t\in J$, and with $\|\tilde{\boldsymbol{\sigma}}(t)\| \leq C\,\|\partial_t\mathbf{u}_n(t)\|$. Then, by differentiating equation (3.5e) with respect to time, and setting $\boldsymbol{\tau}_\ell = \tilde{\boldsymbol{\sigma}}$, we get in the same way as before

$$\|\partial_t\mathbf{u}_n\|^2 \leq C\left(\|\partial_t\boldsymbol{\sigma}_m\|^2 + \|\partial_t p_k\|^2 + \|\partial_t T_i\|^2\right). \tag{3.17}$$

We continue by differentiating equations (3.5e) and (3.5f) with respect to time, and take $\partial_t\boldsymbol{\sigma}_m$ and $\partial_t\mathbf{u}_n$ as test functions, respectively, and get analogously

$$\|\partial_t\boldsymbol{\sigma}_m\|_{\mathcal{A}}^2 \leq \left(\frac{3}{2}c_r + \frac{1}{6(\mu+\lambda)}\right)\|\partial_t p_k\|^2 + \left(\frac{3}{2}a_r + \frac{1}{6(\mu+\lambda)}\right)\|\partial_t T_i\|^2 + C\,\|\partial_t\mathbf{f}\|^2\,, \tag{3.18}$$

and

$$\|\partial_t\mathbf{u}_n\|^2 \leq C\left(\|\partial_t p_k\|^2 + \|\partial_t T_i\|^2 + \|\partial_t\mathbf{f}\|^2\right). \tag{3.19}$$

Next, we take $\partial_t\boldsymbol{\sigma}_m$, p_k, $\mathbf{w}_l$, T_i and $\mathbf{r}_j$ as a test functions in (3.5e), (3.5c), (3.5d), (3.5a) and (3.5b), respectively. We differentiate then (3.5f) with respect to time, and take $\mathbf{u}_n$ as a test function. Adding together the resulting equations yields

$$\begin{aligned}
&(c_0+c_r)(\partial_t p_k, p_k) + (a_0+a_r)(\partial_t T_i, T_i) + (\mathbf{K}^{-1}\mathbf{w}_l, \mathbf{w}_l) + (\boldsymbol{\Theta}^{-1}\mathbf{r}_j, \mathbf{r}_j)\\
&\quad = (\mathcal{A}\boldsymbol{\sigma}_m, \partial_t\boldsymbol{\sigma}_m) + b_r(\partial_t T_i, p_k) + b_r(\partial_t p_k, T_j) + (\boldsymbol{\eta}\cdot\mathbf{w}_l, T_i)\\
&\qquad - (\partial_t\mathbf{f}, \mathbf{u}_n) + (g, p_k) + (h, T_i).
\end{aligned} \tag{3.20a}$$

Using the properties of $\mathbf{K}$ and $\boldsymbol{\Theta}$, in addition to the C–S and Young inequalities yields

$$\begin{aligned}
&(c_0+c_r-b_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|p_k\|^2 + (a_0+a_r-b_r)\,\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|T_i\|^2 + \left(k_m - \gamma\frac{1}{2\epsilon}\right)\|\mathbf{w}_l\|^2 + \theta_m\|\mathbf{r}_j\|^2\\
&\quad \leq \frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{\sigma}_m\|_{\mathcal{A}}^2 + (\epsilon+1)\|T_i\|^2 + \|\mathbf{u}_n\|^2 + \|p_k\|^2 + \|\partial_t\mathbf{f}\|^2 + \|g\|^2 + \|h\|^2\right).
\end{aligned} \tag{3.20b}$$

Choosing $\epsilon = \dfrac{\gamma}{k_m}$, integrating from 0 to t and substituting the inequalities (3.14b) and (3.15c), we deduce

$$\begin{aligned}
&\left(c_0 - \frac{c_r}{2} - b_r - \frac{1}{6(\mu+\lambda)}\right)\|p_k(t)\|^2 + \left(a_0 - \frac{a_r}{2} - b_r - \frac{1}{6(\mu+\lambda)}\right)\|T_i(t)\|^2\\
&\quad + \int_0^t \left(k_m\|\mathbf{w}_l(\tau)\|^2 + \theta_m\|\mathbf{r}_j(\tau)\|^2\right)\mathrm{d}\tau
\end{aligned}$$

$$\leq C \int_0^t \left(\|p_k(\tau)\|^2 + \|T_i(\tau)\|^2 \right) \mathrm{d}\tau - \|\boldsymbol{\sigma}_m(0)\|_{\mathcal{A}}^2$$
$$+ C \left(\|\mathbf{f}\|_{H^1(J;L^2(\Omega))}^2 + \|g\|_{L^2(J;L^2(\Omega))}^2 + \|h\|_{L^2(J;L^2(\Omega))}^2 + \|p_k(0)\|^2 + \|T_i(0)\|^2 \right). \tag{3.20c}$$

Since from (3.5g) we have

$$\|T_i(0)\|^2 \leq \|T_0\|^2 \text{ and } \|p_k(0)\|^2 \leq \|p_0\|^2, \tag{3.21}$$

we obtain the first estimate (i) using Grönwall's inequality, i.e.

$$\|p_k\|_{L^\infty(J;L^2(\Omega))}^2 + \|T_i\|_{L^\infty(J;L^2(\Omega))}^2 + \|\mathbf{w}_l\|_{L^2(J;L^2(\Omega)}^2 + \|\mathbf{r}_j\|_{L^2(J;L^2(\Omega)}^2 + \|\boldsymbol{\sigma}_m(0)\|_{\mathcal{A}}^2$$
$$\leq C \left(\|\mathbf{f}\|_{H^1(J;L^2(\Omega))}^2 + \|g\|_{L^2(J;L^2(\Omega))}^2 + \|h\|_{L^2(J;L^2(\Omega))}^2 + \|p_0\|^2 + \|T_0\|^2 \right). \tag{3.22}$$

For the second estimate, we differentiate (3.5e), (3.5f), (3.5d) and (3.5b) with respect to time and use $\partial_t \boldsymbol{\sigma}_m, \partial_t \mathbf{u}_n, \mathbf{w}_l$ and $\mathbf{r}_j$ as test functions, respectively. In (3.5c) and (3.5a), we use $\partial_t p_k$ and $\partial_t T_i$ as test functions, respectively. Summing the resulting equations yields

$$(c_0 + c_r) \|\partial_t p_k\|^2 + (a_0 + a_r) \|\partial_t T_i\|^2 + (\mathbf{K}^{-1} \partial_t \mathbf{w}_l, \mathbf{w}_l) + (\boldsymbol{\Theta}^{-1} \partial_t \mathbf{r}_j, \mathbf{r}_j)$$
$$= \|\partial_t \boldsymbol{\sigma}_m\|_{\mathcal{A}}^2 + 2b_r(\partial_t T_i, \partial_t p_k) + (\boldsymbol{\eta} \cdot \mathbf{w}_l, \partial_t T_i) - (\partial_t \mathbf{f}, \partial_t \mathbf{u}_n) + (g, \partial_t p_k) + (h, \partial_t T_i). \tag{3.23a}$$

By applying the C–S and Young inequalities, and substituting the estimates (3.17) and (3.18), we deduce

$$\left(c_0 - \frac{c_r}{2} - b_r - \frac{1}{6(\mu+\lambda)} - \frac{\epsilon_2}{2} \right) \|\partial_t p_k\|^2$$
$$+ \left(a_0 - \frac{a_r}{2} - b_r - \frac{1}{6(\mu+\lambda)} - \frac{\epsilon_4}{2} - \frac{\epsilon_3}{2} \right) \|\partial_t T_i\|^2 + \frac{k_m}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{w}_l\|^2 + \frac{\theta_m}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{r}_j\|^2$$
$$\leq \frac{\epsilon_1}{2} C \left(\|\partial_t p_k\|^2 + \|\partial_t T_i\|^2 + \|\partial_t \mathbf{f}\|^2 \right)$$
$$+ \gamma \frac{1}{2\epsilon_4} \|\mathbf{w}_l\|^2 + \frac{1}{2\epsilon_1} \|\partial_t \mathbf{f}\|^2 + \frac{1}{2\epsilon_2} \|g\|^2 + \frac{1}{2\epsilon_3} \|h\|^2. \tag{3.23b}$$

Choosing suitable values for the epsilons, i.e. $\epsilon_1 = \dfrac{\alpha\beta}{C(\mu+\lambda)}$, $\epsilon_2 = \dfrac{\alpha\beta}{\mu+\lambda}$, $\epsilon_3 = \dfrac{\alpha\beta}{2(\mu+\lambda)}$, and $\epsilon_4 = \dfrac{\alpha\beta}{2(\mu+\lambda)}$, we infer

$$\left(c_0 - \frac{c_r}{2} - b_0 - \frac{1}{6(\mu+\lambda)} \right) \|\partial_t p_k\|^2 + \left(a_0 - \frac{a_r}{2} - b_0 - \frac{1}{6(\mu+\lambda)} \right) \|\partial_t T_i\|^2$$
$$+ \frac{k_m}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{w}_l\|^2 + \frac{\theta_m}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{r}_j\|^2$$
$$\leq C \left(\|\mathbf{w}_l\|^2 + \|\partial_t \mathbf{f}\|^2 + \|g\|^2 + \|h\|^2 \right). \tag{3.23c}$$

Simplifying the above expression, integrating over $(0, t)$ and using the initial conditions yields

$$\|\mathbf{w}_l(t)\|^2 + \|\mathbf{r}_j(t)\|^2 + \int_0^t \left(\|\partial_t p_k(\tau)\|^2 + \|\partial_t T_i(\tau)\|^2 \right) \mathrm{d}\tau$$

$$\leq C\left(\int_0^t \|\mathbf{w}_l(\tau)\|^2\,\mathrm{d}\tau + \|\partial_t\mathbf{f}\|^2_{L^2(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{w}_l(0)\|^2 + \|\mathbf{r}_j(0)\|^2\right). \tag{3.23d}$$

It remains to provide estimates for $\|\mathbf{w}_l(0)\|^2$ and $\|\mathbf{r}_j(0)\|^2$. To this end, take $\mathbf{w}_l$ as a test function in equation (3.5d), and set $t = 0$. This gives

$$(\mathbf{K}^{-1}\mathbf{w}_l(0), \mathbf{w}_l(0)) = (p_k(0), \nabla\cdot\mathbf{w}_l(0)), \tag{3.24a}$$

which holds true for any $k, l \geq 1$. Use now the properties of $\mathbf{K}$ to bound the left-hand side, tend $k \to \infty$ and then integrate by parts in the right-hand side to obtain

$$k_m \|\mathbf{w}_l(0)\|^2 \leq (p_0, \nabla\cdot\mathbf{w}_l(0)) = -(\nabla p_0, \mathbf{w}_l(0)) \leq \|\nabla p_0\|\,\|\mathbf{w}_l(0)\|. \tag{3.24b}$$

Thus, we have

$$\|\mathbf{w}_l(0)\|^2 \leq C\,\|p_0\|^2_{H^1_0(\Omega)}. \tag{3.24c}$$

Similarly, using (3.5b), we obtain

$$\|\mathbf{r}_j(0)\|^2 \leq C\,\|T_0\|^2_{H^1_0(\Omega)}. \tag{3.24d}$$

Taking now (3.24c) and (3.24d) in (3.23d), and applying Grönwall's lemma, we obtain the second estimate (ii), i.e.

$$\begin{aligned}&\|\partial_t p_k\|^2_{L^2(J;L^2(\Omega))} + \|\partial_t T_i\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{w}_l\|^2_{L^\infty(J;L^2(\Omega))} + \|\mathbf{r}_j\|^2_{L^\infty(J;L^2(\Omega))}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right).\end{aligned} \tag{3.25}$$

Now we sum the estimates (3.15c), (3.16), (3.18), and (3.19), and substitute the estimates (3.22) and (3.25), to obtain (iii), i.e.

$$\begin{aligned}&\|\boldsymbol{\sigma}_m\|^2_{L^\infty(J;L^2(\Omega))} + \|\partial_t\boldsymbol{\sigma}_m\|^2_{L^2(J;L^2(\Omega))} + \|\mathbf{u}_n\|^2_{L^\infty(J;L^2(\Omega))} + \|\partial_t\mathbf{u}_n\|^2_{L^2(J;L^2(\Omega))}\\ &\quad\leq C\left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right).\end{aligned} \tag{3.26}$$

It remains to obtain the estimate (iv), for which we need just to bound the divergences. Since $\nabla\cdot\mathbf{r}_j(t) \in L^2(\Omega)$ for $t \in J$, we can write $\nabla\cdot\mathbf{r}_j(t) = \sum_{\ell=1}^{\infty}\xi_\ell(t)S_\ell$, for some functions $\xi_\ell(t) \in \mathbb{R}$. Now, we multiply equation (3.5a) with ξ_ℓ, sum over $\ell = 1, .., i$ and use the C–S and Young inequalities to obtain

$$\begin{aligned}&(\nabla\cdot\mathbf{r}_j, \sum_{\ell=1}^{i}\xi_\ell S_\ell)\\ &\quad= (h, \sum_{\ell=1}^{i}\xi_\ell S_\ell) - (a_0+a_r)(\partial_t T_i, \sum_{\ell=1}^{i}\xi_\ell S_\ell) - \frac{a_r}{2\beta}(\partial_t\boldsymbol{\sigma}_l, \sum_{\ell=1}^{i}\xi_\ell S_\ell) + b_r(\partial_t p_k, \sum_{\ell=1}^{i}\xi_\ell S_\ell) + (\boldsymbol{\eta}\cdot\mathbf{w}_l, \sum_{\ell=1}^{i}\xi_\ell S_\ell)\\ &\quad\leq \frac{1}{2}\left(\left\|\sum_{\ell=1}^{i}\partial_t\xi_\ell q_\ell\right\|^2 + 5\,\|h\|^2 + 5(a_0+a_r)^2\,\|\partial_t T_i\|^2 + \frac{5a_r^2}{4\beta^2}\,\|\partial_t\boldsymbol{\sigma}_m\|^2 + 5b_r^2\,\|\partial_t p_k\|^2 + 5\gamma\,\|\mathbf{w}_l\|^2\right).\end{aligned} \tag{3.27a}$$

Using (3.18), integrating in time and using (3.25) we get

$$\int_0^{T_f} (\nabla \cdot \mathbf{r}_j, \sum_{\ell=1}^{i} \xi_\ell S_\ell) \mathrm{d}t \leq \frac{1}{2} \int_0^{T_f} \left\| \sum_{\ell=1}^{i} \xi_\ell S_\ell \right\|^2 \mathrm{d}t + C \left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right). \tag{3.27b}$$

It remains to tend $i \to \infty$ to obtain

$$\|\nabla \cdot \mathbf{r}_j\|^2_{L^2(J;L^2(\Omega))} \leq C \left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right). \tag{3.27c}$$

Similarly, we obtain the following from equations (3.5c) and (3.5f)

$$\|\nabla \cdot \mathbf{w}_l\|^2_{L^2(J;L^2(\Omega))} \leq C \left(\|\mathbf{f}\|^2_{H^1(J;L^2(\Omega))} + \|g\|^2_{L^2(J;L^2(\Omega))} + \|h\|^2_{L^2(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right), \tag{3.27d}$$

and

$$\|\nabla \cdot \boldsymbol{\sigma}_m\|^2_{L^2(J;L^2(\Omega))} \leq C \|\mathbf{f}\|^2_{L^2(J;L^2(\Omega))}. \tag{3.27e}$$

Combining the estimates (3.27c)–(3.27d) with (i) and (iii), we get the estimate (iv). This ends the proof. □

The following estimates prove that the solution has improved regularity given some additional regularity on the data. We state the result as a lemma:

Lemma 3.4 *(Estimates for improved regularity). Assume that* $\mathbf{f} \in H^2(J;L^2(\Omega))$ *and* $g, h \in H^1(J;L^2(\Omega))$. *Then there exists a constant* $C > 0$ *independent of* (i,j,k,l,m,n) *such that*

(i)

$$\|p_k\|^2_{W^{1,\infty}(J;L^2(\Omega))} + \|T_i\|^2_{W^{1,\infty}(J;L^2(\Omega))} + \|\mathbf{w}_l\|^2_{H^1(J;L^2(\Omega))} + \|\mathbf{r}_j\|^2_{H^1(J;L^2(\Omega))} + \|\partial_t \boldsymbol{\sigma}_m(0)\|^2 \leq C \left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right),$$

(ii)

$$\|\boldsymbol{\sigma}_m\|^2_{W^{1,\infty}(J;L^2(\Omega))} + \|\mathbf{u}_n\|^2_{W^{1,\infty}(J;L^2(\Omega))} \leq C \left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right),$$

(iii)

$$\|\mathbf{w}_l\|^2_{L^\infty(J;H(\mathrm{div},\Omega))} + \|\mathbf{r}_j\|^2_{L^\infty(J;H(\mathrm{div},\Omega))} + \|\boldsymbol{\sigma}_m\|^2_{L^\infty(J;H_s(\mathrm{div},\Omega))} \leq C \left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)} \right).$$

Proof. We begin by differentiating equations (3.5e), (3.5c), (3.5d), (3.5a) and (3.5b) with respect to time, and take $\partial_{tt}\boldsymbol{\sigma}_m$, $\partial_t p_k$, $\partial_t \mathbf{w}_l$, $\partial_t T_i$ and $\partial_t \mathbf{r}_j$ as a test functions respectively. Then, we differentiate (3.5f) twice with respect to time, and take $\partial_t \mathbf{u}_n$ as a test function. Summing the resulting equations yields

$$
\begin{aligned}
&(c_0+c_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t p_k\right\|^2+(a_0+a_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t T_i\right\|^2+(\mathbf{K}^{-1}\partial_t\mathbf{w}_l,\partial_t\mathbf{w}_l)+(\boldsymbol{\Theta}^{-1}\partial_t\mathbf{r}_j,\partial_t\mathbf{r}_j)\\
&\quad=\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t\boldsymbol{\sigma}_m\right\|_{\mathcal{A}}+b_r\frac{\mathrm{d}}{\mathrm{d}t}(\partial_t T_i,\partial_t p_k)+(\boldsymbol{\eta}\cdot\partial_t\mathbf{w}_l,\partial_t T_i)\\
&\qquad-(\partial_{tt}\mathbf{f},\partial_t\mathbf{u}_n)+(\partial_t g,\partial_t p_k)+(\partial_t h,\partial_t T_i).
\end{aligned}
\tag{3.28a}
$$

Using the properties of $\mathbf{K}$ and $\boldsymbol{\Theta}$, in addition to the C–S and Young inequalities, we get

$$
\begin{aligned}
&(c_0+c_r-b_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t p_k\right\|^2+(a_0+a_r-b_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t T_i\right\|^2+\frac{k_m}{2}\left\|\partial_t\mathbf{w}_l\right\|^2+\theta_m\left\|\partial_t\mathbf{r}_j\right\|^2\\
&\quad\leq\frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t\boldsymbol{\sigma}_m\right\|_{\mathcal{A}}^2+\frac{\gamma}{k_m}\left\|\partial_t T_i\right\|^2+\left\|\partial_t p_k\right\|^2+\left\|\partial_t\mathbf{u}_n\right\|^2+\left\|\partial_{tt}\mathbf{f}\right\|^2+\left\|\partial_t g\right\|^2+\left\|\partial_t h\right\|^2\right).
\end{aligned}
\tag{3.28b}
$$

By integrating over $(0,t)$, using the initial conditions and substituting the inequalities (3.18) and (3.19), it is inferred that

$$
\begin{aligned}
&\left(c_0-\frac{c_r}{2}-b_r-\frac{1}{6(\mu+\lambda)}\right)\left\|\partial_t p_k(t)\right\|^2+\left(a_0-\frac{a_r}{2}-b_r-\frac{1}{6(\mu+\lambda)}\right)\left\|\partial_t T_i(t)\right\|^2\\
&\quad+\int_0^t\left(k_m\left\|\partial_t\mathbf{w}_l(\tau)\right\|^2+\theta_m\left\|\partial_t\mathbf{r}_j(\tau)\right\|^2\right)\mathrm{d}\tau+\left\|\partial_t\boldsymbol{\sigma}_m(0)\right\|_{\mathcal{A}}^2\\
&\leq C\int_0^t\left(\left\|\partial_t p_k(\tau)\right\|^2+\left\|\partial_t T_i(\tau)\right\|^2\right)\mathrm{d}\tau\\
&\quad+C\left(\left\|\mathbf{f}\right\|_{H^2(J;L^2(\Omega))}^2+\left\|g\right\|_{H^1(J;L^2(\Omega))}^2+\left\|h\right\|_{H^1(J;L^2(\Omega))}^2+\left\|\partial_t p(0)\right\|^2+\left\|\partial_t T(0)\right\|^2\right).
\end{aligned}
\tag{3.28c}
$$

We proceed to bound $\|\partial_t p_k(0)\|$ and $\|\partial_t T_i(0)\|$. To this end, we discard the terms under the time differential on the left-hand side of (3.23c) and set $t=0$ to obtain

$$
\begin{aligned}
&\left(c_0-\frac{c_r}{2}-b_0-\frac{1}{6(\mu+\lambda)}\right)\left\|\partial_t p_k(0)\right\|^2+\left(a_0-\frac{a_r}{2}-b_0-\frac{1}{6(\mu+\lambda)}\right)\left\|\partial_t T_i(0)\right\|^2\\
&\quad\leq C\left(\left\|\mathbf{w}_l(0)\right\|^2+\left\|\partial_t\mathbf{f}(0)\right\|^2+\left\|g(0)\right\|^2+\left\|h(0)\right\|^2\right).
\end{aligned}
\tag{3.29a}
$$

We use (3.24c) to bound the initial value of the Darcy flux, i.e.,

$$
\begin{aligned}
&\left\|\partial_t p_k(0)\right\|^2+\left\|\partial_t T_i(0)\right\|^2\\
&\quad\leq C\left(\left\|p_0\right\|_{H_0^1(\Omega)}^2+\left\|\mathbf{f}\right\|_{H^2(J;L^2(\Omega))}^2+\left\|g\right\|_{H^1(J;L^2(\Omega))}^2+\left\|h\right\|_{H^1(J;L^2(\Omega))}^2\right).
\end{aligned}
\tag{3.29b}
$$

Now we substitute this in (3.28c), using also (i) from Theorem 3.3 and apply Grönwall's Lemma to obtain

$$
\begin{aligned}
&\left\|\partial_t p_k\right\|_{L^\infty(J;L^2(\Omega))}^2+\left\|\partial_t T_i\right\|_{L^\infty(J;L^2(\Omega))}^2+\left\|\partial_t\mathbf{w}_l\right\|_{L^2(J;L^2(\Omega))}^2+\left\|\partial_t\mathbf{r}_j\right\|_{L^2(J;L^2(\Omega))}^2+\left\|\partial_t\boldsymbol{\sigma}_m(0)\right\|_{\mathcal{A}}^2\\
&\quad\leq C\left(\left\|\mathbf{f}\right\|_{H^2(J;L^2(\Omega))}^2+\left\|g\right\|_{H^1(J;L^2(\Omega))}^2+\left\|h\right\|_{H^1(J;L^2(\Omega))}^2+\left\|p_0\right\|_{H_0^1(\Omega)}^2+\left\|T_0\right\|_{H_0^1(\Omega)}^2\right).
\end{aligned}
\tag{3.30}
$$

Summing with (i) from Theorem 3.3 produces the estimate (i). We continue by summing (3.18) and (3.19), and combine with (3.30) to obtain

$$\|\partial_t \boldsymbol{\sigma}_m\|^2_{L^\infty(J;L^2(\Omega))} + \|\partial_t \mathbf{u}_n\|^2_{L^\infty(J;L^2(\Omega))} \leq C\left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right). \tag{3.31}$$

Summing the above estimate with estimate (iii) from Theorem 3.3 produces the estimate (ii). Going back to the estimate (3.27a), we now substitute in the right-hand side with (3.30) and (3.31), let $i \to \infty$ to obtain

$$\|\nabla \cdot \mathbf{r}_j\|^2_{L^\infty(J;L^2(\Omega))} \leq C\left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right). \tag{3.32}$$

From equations (3.5c) and (3.5f) we obtain using the same technique

$$\|\nabla \cdot \mathbf{w}_l\|^2_{L^\infty(J;L^2(\Omega))} \leq C\left(\|\mathbf{f}\|^2_{H^2(J;L^2(\Omega))} + \|g\|^2_{H^1(J;L^2(\Omega))} + \|h\|^2_{H^1(J;L^2(\Omega))} + \|p_0\|^2_{H^1_0(\Omega)} + \|T_0\|^2_{H^1_0(\Omega)}\right), \tag{3.33}$$

and

$$\|\nabla \cdot \boldsymbol{\sigma}_m\|^2_{L^\infty(J;L^2(\Omega))} \leq C\,\|\mathbf{f}\|^2_{L^\infty(J;L^2(\Omega))}. \tag{3.34}$$

Summing the estimates (3.32)–(3.34) and combining with (ii) and (iii) from Theorem 3.3 produces the estimate (iii). This ends the proof. □

3.3. End of the proof of Theorem 3.1:

The proof of the first part of Theorem 3.1 follows the steps below:

• Lemma 3.3 implies that for the sequences $\{\boldsymbol{\sigma}_m\}_0^\infty$, $\{\mathbf{u}_n\}_0^\infty$, $\{p_k\}_0^\infty$, $\{\mathbf{w}_l\}_0^\infty$, $\{T_i\}_0^\infty$ and $\{\mathbf{r}_j\}_0^\infty$ defined by (3.5): $\{\boldsymbol{\sigma}_m\}_0^\infty$ is bounded in $L^\infty(J;H_s(\mathrm{div},\Omega))\cap H^1(J;L^2(\Omega))$, $\{\mathbf{u}_n\}_0^\infty$ is bounded in $H^1(J;L^2(\Omega))$, $\{p_k\}_0^\infty$ is bounded in $H^1(J;L^2(\Omega))$, $\{\mathbf{w}_l\}_0^\infty$ is bounded in $L^2(J;H(\mathrm{div},\Omega))\cap L^\infty(J;L^2(\Omega))$, $\{T_i\}_0^\infty$ is bounded in $H^1(J;L^2(\Omega))$, and $\{\mathbf{r}_j\}_0^\infty$ is bounded in $L^2(J;H(\mathrm{div},\Omega))\cap L^\infty(J;L^2(\Omega))$.

By the weak compactness properties of the spaces there exist subsequences (denoted the same way as before) and functions $\boldsymbol{\sigma} \in L^\infty(J;H_s(\mathrm{div},\Omega))\cap H^1(J;L^2(\Omega))$, $\mathbf{u} \in H^1(J;L^2(\Omega))$, $p \in H^1(J;L^2(\Omega))$, $\mathbf{w} \in L^2(J;H(\mathrm{div},\Omega))\cap L^\infty(J;L^2(\Omega))$, $T \in H^1(J;L^2(\Omega))$, and $\mathbf{r} \in L^2(J;H(\mathrm{div},\Omega))\cap L^\infty(J;L^2(\Omega))$, such that

- $T_i \rightharpoonup T$ in $H^1(J;L^2(\Omega))$,
- $\mathbf{r}_j \rightharpoonup \mathbf{r}$ in $L^2(J;H(\mathrm{div},\Omega))$,
- $p_k \rightharpoonup p$ in $H^1(J;L^2(\Omega))$,
- $\mathbf{w}_l \rightharpoonup \mathbf{w}$ in $L^2(J;H(\mathrm{div},\Omega))$,
- $\boldsymbol{\sigma}_m \rightharpoonup \boldsymbol{\sigma}$ in $L^2(J;H_s(\mathrm{div},\Omega))$,
- $\partial_t\boldsymbol{\sigma}_m \rightharpoonup \partial_t\boldsymbol{\sigma}$ in $L^2(J;L^2(\Omega))$,
- $\mathbf{u}_n \rightharpoonup \mathbf{u}$ in $H^1(J;L^2(\Omega))$.

In order to pass to the limit in problem (3.5), we fix a tuple $(i,j,k,l,m,n) \geq 1$ and take $(S,\mathbf{y},q,\mathbf{z},\boldsymbol{\tau},\mathbf{v}) \in C^1(J;\mathcal{T}_i\times\mathcal{R}_j\times\mathcal{P}_k\times\mathcal{W}_l\times\mathcal{S}_m\times\mathcal{U}_n)$ as test functions, and then integrate equations (3.5a)–(3.5f) with respect to time to obtain

$$\int_0^{T_f}\{(a_0+a_r)(\partial_t T_i,S) - b_r(\partial_t p_k,S) + \frac{a_r}{2\beta}(\partial_t\boldsymbol{\sigma}_m,S\mathbf{I}) + (\boldsymbol{\eta}\cdot\mathbf{w}_l,S) + (\nabla\cdot\mathbf{r}_j,S)\}\mathrm{d}t$$

$$= \int_0^{T_f} (h, S)\mathrm{d}t, \tag{3.35a}$$

$$\int_0^{T_f} \{(\boldsymbol{\Theta}^{-1}\mathbf{r}_j, \mathbf{y}) - (T_i, \nabla\cdot\mathbf{y})\}\mathrm{d}t = 0. \tag{3.35b}$$

$$\int_0^{T_f} \{(c_0 + c_r)(\partial_t p_k, q) - b_r(\partial_t T_i, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}_m, q\mathbf{I}) + (\nabla\cdot\mathbf{w}_l, q)\}\mathrm{d}t = \int_0^{T_f} (g, q)\mathrm{d}t, \tag{3.35c}$$

$$\int_0^{T_f} \{(\mathbf{K}^{-1}\mathbf{w}_l, \mathbf{z}) - (p_k, \nabla\cdot\mathbf{z})\}\mathrm{d}t = 0, \tag{3.35d}$$

$$\int_0^{T_f} \{(\mathcal{A}\boldsymbol{\sigma}_m, \boldsymbol{\tau}) + (\mathbf{u}_n, \nabla\cdot\boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}p_k, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}T_i, \boldsymbol{\tau})\}\mathrm{d}t = 0, \tag{3.35e}$$

$$-\int_0^{T_f} (\nabla\cdot\boldsymbol{\sigma}_m, \mathbf{v})\mathrm{d}t = \int_0^{T_f} (\mathbf{f}, \mathbf{v})\mathrm{d}t. \tag{3.35f}$$

Passing to the limit yields

$$\int_0^{T_f} \{(a_0 + a_r)(\partial_t T, S) - b_r(\partial_t p, S) + \frac{a_r}{2\beta}(\partial_t \boldsymbol{\sigma}, S\mathbf{I}) + (\boldsymbol{\eta}\cdot\mathbf{w}, S) + (\nabla\cdot\mathbf{r}, S)\}\mathrm{d}t = \int_0^{T_f} (h, S)\mathrm{d}t, \tag{3.36a}$$

$$\int_0^{T_f} \{(\boldsymbol{\Theta}^{-1}\mathbf{r}, \mathbf{y}) - (T, \nabla\cdot\mathbf{y})\}\mathrm{d}t = 0. \tag{3.36b}$$

$$\int_0^{T_f} \{(c_0 + c_r)(\partial_t p, q) - b_r(\partial_t T, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}, q\mathbf{I}) + (\nabla\cdot\mathbf{w}, q)\}\mathrm{d}t = \int_0^{T_f} (g, q)\mathrm{d}t, \tag{3.36c}$$

$$\int_0^{T_f} \{(\mathbf{K}^{-1}\mathbf{w}, \mathbf{z}) - (p, \nabla\cdot\mathbf{z})\}\mathrm{d}t = 0, \tag{3.36d}$$

$$\int_0^{T_f} \{(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\mathbf{u}, \nabla\cdot\boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}p, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}T, \boldsymbol{\tau})\}\mathrm{d}t = 0, \tag{3.36e}$$

$$-\int_0^{T_f} (\nabla\cdot\boldsymbol{\sigma}, \mathbf{v})\mathrm{d}t = \int_0^{T_f} (\mathbf{f}, \mathbf{v})\mathrm{d}t. \tag{3.36f}$$

Finally, by the density of the test function space, $C^1(J; \mathcal{T}_i \times \mathcal{R}_j \times \mathcal{P}_k \times \mathcal{W}_l \times \mathcal{S}_m \times \mathcal{U}_n)$ in $L^2(J; \mathcal{T} \times \mathcal{R} \times \mathcal{P} \times \mathcal{W} \times \mathcal{S} \times \mathcal{U})$ as $(i, j, k, l, m, n) \to \infty$, the equations (3.1) hold true for a.e. $t \in J$. It remains now to show that the initial conditions are satisfied, i.e. $T(0) = T_0$, $\mathbf{u}(0) = \mathbf{u}_0$ and $p(0) = p_0$, in the weak sense. To this end, take $q \in C^1(J; \mathcal{P}_k)$ such that $q(T_f) = 0$ as a test function in (3.35c) and integrate the first term by parts in time

$$\int_0^{T_f} \{-(c_0+c_r)(p_k,\partial_t q) - b_r(\partial_t T_i, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}_m, q\mathbf{I}) + (\nabla\cdot\mathbf{w}_l, q)\}\mathrm{d}t$$
$$= \int_0^{T_f} (g,q)\mathrm{d}t + (c_0+c_r)(p_k(0), q(0)). \tag{3.37}$$

On the other hand, from (3.36c) we obtain

$$\int_0^{T_f} \{-(c_0+c_r)(p,\partial_t q) - b_r(\partial_t T, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}, q\mathbf{I}) + (\nabla\cdot\mathbf{w}, q)\}\mathrm{d}t$$
$$= \int_0^{T_f} (g,q)\mathrm{d}t + (c_0+c_r)(p(0), q(0)). \tag{3.38}$$

Since $q(0)$ was arbitrary, and since $p_n(0)\to p_0$ in $L^2(\Omega)$, we get that $p(0)=p_0$. We obtain in the same way that $\mathbf{u}(0)=\mathbf{u}_0$, and $T(0)=T_0$.

• To finish the proof we show the uniqueness of a weak solution to problem (3.1). To this end, we assume that $(T_1(t),\mathbf{r}_1(t),p_1(t),\mathbf{w}_1(t),\boldsymbol{\sigma}_1(t),\mathbf{u}_1(t))$ and $(T_2(t),\mathbf{r}_2(t),p_2(t),\mathbf{w}_2(t),\boldsymbol{\sigma}_2(t),\mathbf{u}_2(t))$ are two solution tuples in $\mathcal{T}\times\mathcal{R}\times\mathcal{P}\times\mathcal{W}\times\mathcal{S}\times\mathcal{U}$, and let $(e_T(t),\mathbf{e}_\mathbf{r}(t),e_p(t),\mathbf{e}_\mathbf{w}(t),\mathbf{e}_{\boldsymbol{\sigma}}(t),\mathbf{e}_\mathbf{u}(t))$ be the corresponding difference. This then satisfies the following variational problem: find $(e_T(t),\mathbf{e}_\mathbf{r}(t),e_p(t),\mathbf{e}_\mathbf{w}(t),\mathbf{e}_{\boldsymbol{\sigma}}(t),\mathbf{e}_\mathbf{u}(t))\in\mathcal{T}\times\mathcal{R}\times\mathcal{P}\times\mathcal{W}\times\mathcal{S}\times\mathcal{U}$ such that for a.e. $t\in J$ there holds

$$(a_0+a_r)(\partial_t e_T, S) - b_r(\partial_t e_p, S) + \frac{a_r}{2\beta}(\partial_t \mathbf{e}_{\boldsymbol{\sigma}}, S\mathbf{I}) - (\boldsymbol{\eta}\cdot\mathbf{e}_\mathbf{w}, S) + (\nabla\cdot\mathbf{e}_\mathbf{r}, S) = 0, \qquad \forall S\in\mathcal{T}, \tag{3.39a}$$
$$(\boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r}, \mathbf{y}) - (e_T, \nabla\cdot\mathbf{y}) = 0, \qquad \forall \mathbf{y}\in\mathcal{R}, \tag{3.39b}$$
$$(c_0+c_r)(\partial_t e_p, q) - b_r(\partial_t e_T, q) + \frac{c_r}{2\alpha}(\partial_t \mathbf{e}_{\boldsymbol{\sigma}}, q\mathbf{I}) + (\nabla\cdot\mathbf{e}_\mathbf{w}, q) = 0, \qquad \forall q\in\mathcal{P}, \tag{3.39c}$$
$$(\mathbf{K}^{-1}\mathbf{e}_\mathbf{w}, \mathbf{z}) - (e_p, \nabla\cdot\mathbf{z}) = 0, \qquad \forall \mathbf{z}\in\mathcal{W}, \tag{3.39d}$$
$$(\mathcal{A}\mathbf{e}_{\boldsymbol{\sigma}}, \boldsymbol{\tau}) + (\mathbf{e}_\mathbf{u}, \nabla\cdot\boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}e_p, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}e_T, \boldsymbol{\tau}) = 0, \qquad \forall \boldsymbol{\tau}\in\mathcal{S}, \tag{3.39e}$$
$$(\nabla\cdot\mathbf{e}_{\boldsymbol{\sigma}}, \mathbf{v}) = 0, \qquad \forall \mathbf{v}\in\mathcal{U}, \tag{3.39f}$$

together with homogeneous initial conditions. Take now $\boldsymbol{\tau} = \partial_t \mathbf{e}_{\boldsymbol{\sigma}}$ in (3.39e), differentiate (3.39f) with respect to time and set $\mathbf{v}=\mathbf{e}_\mathbf{u}$, $q=e_p$ in (3.39c), $\mathbf{z}=\mathbf{e}_\mathbf{w}$ in (3.39d), $S=e_T$ in (3.39a), and $\mathbf{y}=\mathbf{e}_\mathbf{r}$ in (3.39b), and add the resulting equations together

$$(c_0+c_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(e_p,e_p) + (a_0+a_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(e_T,e_T) + (\mathbf{K}^{-1}\mathbf{e}_\mathbf{w},\mathbf{e}_\mathbf{w}) + (\boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r},\mathbf{e}_\mathbf{r})$$
$$= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{A}\mathbf{e}_{\boldsymbol{\sigma}},\mathbf{e}_{\boldsymbol{\sigma}}) + b_r\frac{\mathrm{d}}{\mathrm{d}t}(e_p,e_T) + (\boldsymbol{\eta}\cdot\mathbf{e}_\mathbf{w}, e_T). \tag{3.40}$$

Integrating the above equation from 0 to t and using the properties of $\mathbf{K}$ and $\boldsymbol{\Theta}$, in addition to the C–S and Young inequalities yields

$$(c_0+c_r)\frac{1}{2}\|e_p(t)\|^2 + (a_0+a_r)\frac{1}{2}\|e_T(t)\|^2 + \int_0^t \left(k_m\|\mathbf{e}_\mathbf{w}(\tau)\|^2 + \theta_m\|\mathbf{e}_\mathbf{r}(\tau)\|^2\right)\mathrm{d}\tau$$

$$\leq \frac{1}{2}\left\|\mathbf{e}_{\boldsymbol{\sigma}}(t)\right\|_{\mathcal{A}}^2+\frac{b_r}{2}\left\|e_p(t)\right\|^2+\frac{b_r}{2}\left\|e_T(t)\right\|^2+\int_0^t\left(\gamma\frac{\epsilon}{2}\left\|\mathbf{e}_{\mathbf{w}}(\tau)\right\|^2+\frac{1}{2\epsilon}\left\|e_T(\tau)\right\|^2\right)\mathrm{d}\tau, \tag{3.41}$$

for some $\epsilon>0$. On the other hand, from (3.39e) and (3.39f) we obtain

$$\begin{aligned}\left\|\mathbf{e}_{\boldsymbol{\sigma}}\right\|_{\mathcal{A}}^2&=-\frac{c_r}{2\alpha}(\mathbf{I}e_p,\mathbf{e}_{\boldsymbol{\sigma}})+\frac{a_r}{2\beta}(\mathbf{I}e_T,\mathbf{e}_{\boldsymbol{\sigma}})\\&\leq\left(\frac{c_r}{2\alpha}\frac{\epsilon_1}{2}+\frac{a_r}{2\beta}\frac{\epsilon_2}{2}\right)2(\mu+\lambda)\left\|\mathbf{e}_{\boldsymbol{\sigma}}\right\|_{\mathcal{A}}^2+\frac{c_r}{2\alpha}\frac{1}{2\epsilon_1}\left\|e_p\right\|^2+\frac{a_r}{2\beta}\frac{1}{2\epsilon_2}\left\|e_T\right\|^2.\end{aligned} \tag{3.42}$$

Choosing $\epsilon_1=\dfrac{1}{2\alpha}$ and $\epsilon_2=\dfrac{1}{2\beta}$, we get

$$\frac{1}{2}\left\|\mathbf{e}_{\boldsymbol{\sigma}}\right\|_{\mathcal{A}}^2\leq\frac{c_r}{2}\left\|e_p\right\|+\frac{a_r}{2}\left\|e_T\right\|. \tag{3.43}$$

Combining now (3.41) and (3.43), and choosing $\epsilon=\dfrac{k_m}{\gamma}$, we get

$$\frac{1}{2}\left((c_0-b_r)\left\|e_p(t)\right\|^2+(a_0-b_r)\left\|e_T(t)\right\|^2\right)+\int_0^t\left(\frac{k_m}{2}\left\|\mathbf{e}_{\mathbf{w}}(\tau)\right\|+\theta_m\left\|\mathbf{e}_{\mathbf{r}}(\tau)\right\|^2\right)\mathrm{d}\tau\leq\frac{\gamma}{2k_m}\int_0^t\left\|e_T(\tau)\right\|^2\mathrm{d}\tau, \tag{3.44}$$

which after application of the Grönwall inequality yields

$$(c_0-b_r)\left\|e_p(t)\right\|^2+(a_0-b_r)\left\|e_T(t)\right\|^2+\int_0^t\left(k_m\left\|\mathbf{e}_{\mathbf{w}}(\tau)\right\|^2+2\theta_m\left\|\mathbf{e}_{\mathbf{r}}(\tau)\right\|^2\right)\mathrm{d}\tau\leq 0. \tag{3.45}$$

Then, using Thomas' Lemma [33] we take $\boldsymbol{\tau}=\tilde{\boldsymbol{\sigma}}(\cdot,t)\in\mathcal{S}$ in (3.39e), such that for $t\in J$, $-\nabla\cdot\tilde{\boldsymbol{\sigma}}(t)=\mathbf{e}_{\mathbf{u}}(t)$ in Ω, with $\left\|\tilde{\boldsymbol{\sigma}}(t)\right\|\leq C\left\|\mathbf{e}_{\mathbf{u}}(t)\right\|$ for some constant $C>0$. Thus, we obtain

$$\begin{aligned}\left\|\mathbf{e}_{\mathbf{u}}\right\|^2&=-(\mathbf{e}_{\mathbf{u}},\nabla\cdot\tilde{\boldsymbol{\sigma}})=(\mathcal{A}\mathbf{e}_{\boldsymbol{\sigma}},\tilde{\boldsymbol{\sigma}})+\frac{c_r}{2\alpha}(\mathbf{I}e_p,\tilde{\boldsymbol{\sigma}})+\frac{a_r}{2\beta}(\mathbf{I}e_T,\tilde{\boldsymbol{\sigma}})\\&\leq\left\|\tilde{\boldsymbol{\sigma}}\right\|\left(\frac{1}{2\mu}\left\|\mathbf{e}_{\boldsymbol{\sigma}}\right\|+\frac{c_r}{2\alpha}\left\|e_p\right\|+\frac{a_r}{2\beta}\left\|e_T\right\|\right)\end{aligned} \tag{3.46}$$

$$\implies\left\|\mathbf{e}_{\mathbf{u}}\right\|\leq C(\left\|\mathbf{e}_{\boldsymbol{\sigma}}\right\|+\left\|e_p\right\|+\left\|e_T\right\|). \tag{3.47}$$

This implies that $e_T(t)=\mathbf{e}_{\mathbf{r}}(t)=e_p(t)=\mathbf{e}_{\mathbf{w}}(t)=\mathbf{e}_{\boldsymbol{\sigma}}(t)=\mathbf{e}_{\mathbf{u}}(t)=0$, in Ω, for a.e. $t\in J$, implying the uniqueness of a weak solution to problem (3.1). Finally, thanks to Lemma 3.4, we can finish the proof of the second part of Theorem 3.1 using similar arguments. □

4. Analysis of the non-linear problem

We now consider the analysis of the mixed variational formulation for the original nonlinear problem (2.8). The analysis uses the results derived previously for the linear case, in addition to the Banach Fixed Point Theorem (see e.g. [11]) in order to obtain a local solution to (2.8) in time. We then proceed to extend this local solution by small increments until a global solution is obtained for any finite final time (see e.g. [19,35] where similar techniques are used). Precisely, an iterative solution procedure is introduced based on linearizing the heat flux term in (2.8a), which is shown to be well-defined, and which converges to the weak solution of

the nonlinear problem in adequate norms. Note that we now must require the iterates to be continuous in time, hence we shall invoke Lemma 3.4. The iterative linearization algorithm we consider is then as follows: let $m \geq 1$, and at the iteration m, we solve for $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m) \in \mathcal{T} \times \mathcal{R} \times \mathcal{P} \times \mathcal{W} \times \mathcal{S} \times \mathcal{U}$ such that for $t \in J$ there holds

$$
\begin{aligned}
(a_0 + a_r)(\partial_t T^m, S) - b_r(\partial_t p^m, S) + \frac{a_r}{2\beta}(\partial_t \boldsymbol{\sigma}^m, S\mathbf{I}) & \\
+(\nabla \cdot \mathbf{r}^m, S) + (\mathbf{w}^m \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}^{m-1}, S) = (h, S), \qquad & \forall S \in \mathcal{T}, \qquad (4.1a)\\
(\boldsymbol{\Theta}^{-1}\mathbf{r}^m, \mathbf{y}) - (T^m, \nabla \cdot \mathbf{y}) = 0, \qquad & \forall \mathbf{y} \in \mathcal{R}, \qquad (4.1b)\\
(c_0 + c_r)(\partial_t p^m, q) - b_r(\partial_t T^m, q) + \frac{c_r}{2\alpha}(\partial_t \boldsymbol{\sigma}^m, q\mathbf{I}) + (\nabla \cdot \mathbf{w}^m, q) = (g, q), \qquad & \forall q \in \mathcal{P}, \qquad (4.1c)\\
(\mathbf{K}^{-1}\mathbf{w}^m, \mathbf{z}) - (p^m, \nabla \cdot \mathbf{z}) = 0, \qquad & \forall \mathbf{z} \in \mathcal{W}, \qquad (4.1d)\\
(\mathcal{A}\boldsymbol{\sigma}^m, \boldsymbol{\tau}) + (\mathbf{u}^m, \nabla \cdot \boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}p^m, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}T^m, \boldsymbol{\tau}) = 0, \qquad & \forall \boldsymbol{\tau} \in \mathcal{S}, \qquad (4.1e)\\
-(\nabla \cdot \boldsymbol{\sigma}^m, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \qquad & \forall \mathbf{v} \in \mathcal{U}, \qquad (4.1f)
\end{aligned}
$$

together with initial conditions, (2.8g), and where the algorithm is initialized by given initial guess $\mathbf{r}^0$. We consider the following hypothesis on the heat flux:

Hypothesis 1 *(The heat flux). We suppose that for all $m \geq 1$, the heat flux is such that $\mathbf{r}^m(t) \in L^\infty(\Omega)$, for $t \in J$.*

The above hypothesis is a natural one, and it is necessary for the solution to the iterative procedure (4.1) to be well-defined for each $m \geq 1$. This hypothesis is satisfied with sufficiently regular data and domain boundary. We provide some formal arguments in Appendix A on the specific requirements such that the solution to the problem (3.1) yields $\mathbf{r} \in C([0, T_f], L^\infty(\Omega))$ (or alternatively $\mathbf{w}, \mathbf{r} \in C([0, T_f]; L^4(\Omega))$), thus making the above hypothesis superfluous. We delegate this discussion to the Appendix in order to avoid overly strict assumptions on the data.

Remark 4.1. Note that if we had approximated the convective term in equation (4.1a) instead as $(\mathbf{w}^{m-1} \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}^m, S)$, Hypothesis 1 would be on the regularity of the Darcy flux $\mathbf{w}$, and the above algorithm would be initialized by some $\mathbf{w}^0$. However the analysis presented next remains true and follows exactly the same lines.

Based on the development of the previous sections, we now state the main result of this article.

Theorem 4.1. *Assume that $\mathbf{f}$ is in $H^2(J; L^2(\Omega))$, g, h in $H^1(J; L^2(\Omega))$, p_0, T_0 in $H_0^1(\Omega)$, and $\mathbf{u}_0$ in $(L^2(\Omega))^d$, then the algorithm (4.1), initialized by any $\mathbf{r}^0 \in C([0, T_f]; L^\infty(\Omega))$, defines a unique sequence of iterates*

$$
\begin{aligned}
(T^m, \mathbf{r}^m) &\in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H(\mathrm{div}; \Omega)) \cap H^1(J; L^2(\Omega))\right), \qquad (4.2a)\\
(p^m, \mathbf{w}^m) &\in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H(\mathrm{div}; \Omega)) \cap H^1(J; L^2(\Omega))\right), \qquad (4.2b)\\
(\mathbf{u}^m, \boldsymbol{\sigma}^m) &\in W^{1,\infty}(J; L^2(\Omega)) \times \left(L^\infty(J; H_s(\mathrm{div}; \Omega)) \cap W^{1,\infty}(J; L^2(\Omega))\right), \qquad (4.2c)
\end{aligned}
$$

that converges to the weak solution $(T, \mathbf{r}, p, \mathbf{w}, \boldsymbol{\sigma}, \mathbf{u})$ of (2.8), admitting the following regularity

$$
\begin{aligned}
(T, \mathbf{r}) &\in H^1(J; L^2(\Omega)) \times \left(L^2(J; H(\mathrm{div}; \Omega)) \cap L^\infty(J; L^2(\Omega))\right), \qquad (4.3a)\\
(p, \mathbf{w}) &\in H^1(J; L^2(\Omega)) \times \left(L^2(J; H(\mathrm{div}; \Omega)) \cap L^\infty(J; L^2(\Omega))\right), \qquad (4.3b)
\end{aligned}
$$

$$(\mathbf{u},\boldsymbol{\sigma}) \in H^1(J;L^2(\Omega)) \times \left(L^2(J;H_s(\mathrm{div};\Omega)) \cap H^1(J;L^2(\Omega))\right). \tag{4.3c}$$

Proof. According to Theorem 3.1 and recalling Hypothesis 1, the iterates $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m)$ are well-defined for all $m \geq 1$, admitting the improved regularity specified in Lemma 3.4. In particular, this guarantees continuity in time for the iterates. Keeping this in mind, we define $\gamma_1 := \sup_{t\in J} \|\mathbf{e}_{\mathbf{w}}^m(t)\|^2$ and $\gamma_2 := \sup_{t\in J} \|\mathbf{e}_{\mathbf{r}}^m(t)\|^2$. It remains to show the convergence of the iterates to the weak solution of (2.8) using suitable norms. To this aim, let $m \geq 2$, and take the difference of equations (4.1) at the iteration step m, with the corresponding equations at iteration step $m-1$ to obtain the following problem: find $(e_T^m, \mathbf{e}_{\mathbf{r}}^m, e_p^m, \mathbf{e}_{\mathbf{w}}^m, \mathbf{e}_{\boldsymbol{\sigma}}^m, \mathbf{e}_{\mathbf{u}}^m) \in \mathcal{T}\times\mathcal{R}\times\mathcal{P}\times\mathcal{W}\times\mathcal{S}\times\mathcal{U}$ such that for $t \in J$ there holds

$$\begin{aligned}
(a_0+a_r)(\partial_t e_T^m, S) - b_r(\partial_t e_p^m, S) + \frac{a_r}{2\beta}(\partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m, S\mathbf{I}) + (\nabla\cdot\mathbf{e}_{\mathbf{r}}^m, S) \\
-(\mathbf{w}^m\cdot\boldsymbol{\Theta}^{-1}\mathbf{e}_{\mathbf{r}}^{m-1}, S) - (\mathbf{e}_{\mathbf{w}}^m\cdot\boldsymbol{\Theta}^{-1}\mathbf{r}^{m-1}, S) = 0, \qquad \forall S\in\mathcal{T}, \qquad &(4.4a)\\
(\boldsymbol{\Theta}^{-1}\mathbf{e}_{\mathbf{r}}^m, \mathbf{y}) - (e_T^m, \nabla\cdot\mathbf{y}) = 0, \qquad \forall \mathbf{y}\in\mathcal{R} \qquad &(4.4b)\\
(c_0+c_r)(\partial_t e_p^m, q) - b_r(\partial_t e_T^m, q) + \frac{c_r}{2\alpha}(\partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m, q\mathbf{I}) + (\nabla\cdot\mathbf{e}_{\mathbf{w}}^m, q) = 0, \qquad \forall q\in\mathcal{P}, \qquad &(4.4c)\\
(\mathbf{K}^{-1}\mathbf{e}_{\mathbf{w}}^m, \mathbf{z}) - (e_p^m, \nabla\cdot\mathbf{z}) = 0, \qquad \forall \mathbf{z}\in\mathcal{W}, \qquad &(4.4d)\\
(\mathcal{A}\mathbf{e}_{\boldsymbol{\sigma}}^m, \boldsymbol{\tau}) + (\mathbf{e}_{\mathbf{u}}^m, \nabla\cdot\boldsymbol{\tau}) + \frac{c_r}{2\alpha}(\mathbf{I}e_p^m, \boldsymbol{\tau}) + \frac{a_r}{2\beta}(\mathbf{I}e_T^m, \boldsymbol{\tau}) = 0, \qquad \forall \boldsymbol{\tau}\in\mathcal{S}, \qquad &(4.4e)\\
-(\nabla\cdot\mathbf{e}_{\boldsymbol{\sigma}}^m, \mathbf{v}) = 0, \qquad \forall \mathbf{v}\in\mathcal{U}, \qquad &(4.4f)
\end{aligned}$$

together with homogeneous initial conditions, i.e.

$$(e_T^m(0), S) = 0, \quad \forall\in\mathcal{T}, \quad (\mathbf{e}_{\mathbf{u}}^m(0), \mathbf{v}) = 0, \quad \forall\mathbf{v}\in\mathcal{U}, \quad \text{and} \quad (e_p^m(0), q) = 0, \quad \forall q\in\mathcal{P}. \tag{4.4g}$$

The solution tuple $(e_T^m, \mathbf{e}_{\mathbf{r}}^m, e_p^m, \mathbf{e}_{\mathbf{w}}^m, \mathbf{e}_{\boldsymbol{\sigma}}^m, \mathbf{e}_{\mathbf{u}}^m)$ denotes the error functions between the solution to (4.1) at the m^{th} and $(m-1)^{\text{th}}$ iterations, i.e. $e_T^m = T^m - T^{m-1}$, and similarly for the other variables. We continue to denote by C a generic positive constant which may change value from one line to the next, but in this section the relevant parameter is m. First, take $\boldsymbol{\tau} = \mathbf{e}_{\boldsymbol{\sigma}}^m$ and $\mathbf{v} = \mathbf{e}_{\mathbf{u}}^m$ in equations (4.4e) and (4.4f), respectively, and sum to obtain

$$\begin{aligned}
\|\mathbf{e}_{\boldsymbol{\sigma}}^m\|_{\mathcal{A}}^2 &= -\frac{c_r}{2\alpha}(\mathbf{I}e_p^m, \mathbf{e}_{\boldsymbol{\sigma}}^m) - \frac{a_r}{2\beta}(\mathbf{I}e_T^m, \mathbf{e}_{\boldsymbol{\sigma}}^m)\\
&\leq \left(\alpha\frac{\epsilon_1}{2} + \beta\frac{\epsilon_2}{2}\right)\|\mathbf{e}_{\boldsymbol{\sigma}}^m\|_{\mathcal{A}}^2 + \frac{c_r}{2\alpha}\frac{1}{2\epsilon_1}\left\|e_p^m\right\|^2 + \frac{a_r}{2\beta}\frac{1}{2\epsilon_1}\left\|e_T^m\right\|^2.
\end{aligned} \tag{4.5a}$$

Setting $\epsilon_1 = \dfrac{1}{2\alpha}$ and $\epsilon_2 = \dfrac{1}{2\beta}$ yields

$$\|\mathbf{e}_{\boldsymbol{\sigma}}^m\|_{\mathcal{A}}^2 \leq c_r\left\|e_p^m\right\|^2 + a_r\left\|e_T^m\right\|^2. \tag{4.5b}$$

Similarly, by differentiating equations (4.4e) and (4.4f) with respect to time and setting $\boldsymbol{\tau} = \partial_t\mathbf{e}_{\boldsymbol{\sigma}}^m$ and $\mathbf{v} = \partial_t\mathbf{e}_{\mathbf{u}}^m$ we obtain

$$\|\partial_t\mathbf{e}_{\boldsymbol{\sigma}}^m\|_{\mathcal{A}}^2 \leq c_r\left\|\partial_t e_p^m\right\|^2 + a_r\left\|\partial_t e_T^m\right\|^2. \tag{4.6}$$

Using Thomas' lemma [33], we take $\boldsymbol{\tau} = \tilde{\boldsymbol{\sigma}}(\cdot,t)$ in equation (4.4e) such that $\mathbf{e}_{\mathbf{u}}^m(\cdot,t) = \nabla\cdot\tilde{\boldsymbol{\sigma}}(\cdot,t)$ with $\|\tilde{\boldsymbol{\sigma}}(t)\| \leq C\|\mathbf{e}_{\mathbf{u}}^m(t)\|$ for $t\in J$, and combine with (4.5b) to obtain

$$\|\mathbf{e}_{\mathbf{u}}^m\|^2 \leq C\left(\left\|e_p^m\right\|^2 + \|e_T^m\|^2\right), \tag{4.7}$$

and similarly using (4.6)

$$\|\partial_t \mathbf{e}_{\mathbf{u}}^m\|^2 \leq C\left(\left\|\partial_t e_p^m\right\|^2 + \|\partial_t e_T^m\|^2\right). \tag{4.8}$$

Now, write $\nabla \cdot \mathbf{e}_{\mathbf{r}}^m(t) = \sum_{\ell=1}^{\infty} \zeta_\ell(t) S_\ell$ for some functions $\zeta_\ell(t) \in \mathbb{R}$, where $\text{span}\{S_\ell : 1 \leq \ell \leq \infty\} = \mathcal{T}$. Then, we take S_ℓ as a test function in equation (4.4a), multiply by ζ_ℓ and sum over $\ell = 1, ..., k$ to obtain

$$\begin{aligned}(\nabla \cdot \mathbf{e}_{\mathbf{r}}^m, \sum_{\ell=1}^{k} \zeta_\ell S_\ell) = b_r(\partial_t e_p^m, \sum_{\ell=1}^{k} \zeta_\ell S_\ell) - (a_0 + a_r)(\partial_t e_T^m, \sum_{\ell=1}^{k} \zeta_\ell S_\ell) - \frac{a_r}{2\beta}(\partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m, \sum_{\ell=1}^{k} \zeta_\ell S_\ell) \\ + (\mathbf{w}^m \cdot \boldsymbol{\Theta}^{-1} \mathbf{e}_{\mathbf{r}}^{m-1}, \sum_{\ell=1}^{k} \zeta_\ell S_\ell) + (\mathbf{e}_{\mathbf{w}}^m \cdot \boldsymbol{\Theta}^{-1} \mathbf{r}^{m-1}, \sum_{\ell=1}^{k} \zeta_\ell S_\ell).\end{aligned} \tag{4.9a}$$

Using the C–S and Young inequalities, tending $k \to \infty$, and using also the estimate (4.6) we get

$$\|\nabla \cdot \mathbf{e}_{\mathbf{r}}^m\|^2 \leq C\left(\left\|\partial_t e_p^m\right\|^2 + \|\partial_t e_T^m\|^2 + \|\mathbf{e}_{\mathbf{w}}^m\|^2 + \left\|\mathbf{e}_{\mathbf{r}}^{m-1}\right\|^2\right). \tag{4.9b}$$

In the same way we get from equation (4.4c) that

$$\|\nabla \cdot \mathbf{e}_{\mathbf{w}}^m\|^2 \leq C\left(\|\partial_t e_T^m\|^2 + \left\|\partial_t e_p^m\right\|^2\right). \tag{4.10}$$

From (4.4f) we also have that

$$\|\nabla \cdot \mathbf{e}_{\boldsymbol{\sigma}}^m\|^2 = 0. \tag{4.11}$$

We continue by setting $S = e_T^m, \mathbf{y} = \mathbf{e}_{\mathbf{r}}^m, q = e_p^m, \mathbf{z} = \mathbf{e}_{\mathbf{w}}^m, \boldsymbol{\tau} = \partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m$ in equations (4.4a)–(4.4e), and differentiate equation (4.4f) with respect to time and set $\mathbf{v} = \mathbf{e}_{\mathbf{u}}^m$. Summing the resulting equations yields

$$\begin{aligned}(c_0 - b_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|e_p^m\right\|^2 + (a_0 - b_r)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_T^m\|^2 + k_m\|\mathbf{e}_{\mathbf{w}}^m\|^2 + \theta_m\|\mathbf{e}_{\mathbf{r}}^m\|^2 \\ \leq \gamma_1\frac{\theta_M}{2}\left\|\mathbf{e}_{\mathbf{r}}^{m-1}\right\|^2 + \gamma_2\theta_M\frac{\epsilon}{2}\|\mathbf{e}_{\mathbf{w}}^m\|^2 + \left(\frac{1}{2} + \frac{1}{2\epsilon}\right)\|e_T^m\|^2,\end{aligned} \tag{4.12a}$$

where we also used the estimate (4.6). Integrating from 0 to t, applying the Grönwall inequality and setting $\epsilon = \dfrac{k_m}{\gamma_2\theta_M}$ yields

$$(c_0 - b_r)\left\|e_p^m(t)\right\|^2 + (a_0 - b_r)\|e_T^m(t)\|^2 + \int_0^t \left(k_m\|\mathbf{e}_{\mathbf{w}}^m(\tau)\|^2 + \theta_m\|\mathbf{e}_{\mathbf{r}}^m(\tau)\|^2\right)\mathrm{d}\tau \leq C\int_0^t \left\|\mathbf{e}_{\mathbf{r}}^{m-1}(\tau)\right\|^2 \mathrm{d}\tau. \tag{4.12b}$$

Take now $S = \partial_t e_T^m$ and $q = \partial_t e_p^m$ in equations (4.4a) and (4.4c), respectively. Then, differentiate equations (4.4e) and (4.4f) with respect to time and let $\boldsymbol{\tau} = \partial_t \boldsymbol{\sigma}^m$ and $\mathbf{v} = \partial_t \mathbf{u}^m$. Finally, we let $\mathbf{y} = \partial_t \mathbf{e}_{\mathbf{r}}^m$ and $\mathbf{z} = \partial_t \mathbf{e}_{\mathbf{w}}^m$ in equations (4.4b) and (4.4d), respectively. Summing yields

$$
\begin{aligned}
&(c_0 + c_r - b_r)\left\|\partial_t e_p^m\right\|^2 + (a_0 + a_r - b_r)\left\|\partial_t e_T^m\right\|^2 + \frac{k_m}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{e}_{\mathbf{w}}^m\right\|^2 + \frac{\theta_m}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{e}_{\mathbf{r}}^m\right\|^2 \\
&\quad \le \left\|\partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m\right\|_{\mathcal{A}}^2 + (\mathbf{w}^m \cdot \boldsymbol{\Theta}^{-1}\mathbf{e}_{\mathbf{r}}^{m-1}, \partial_t e_T^m) + (\mathbf{e}_{\mathbf{w}}^m \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}^{m-1}, \partial_t e_T^m) \\
&\quad \le \left\|\partial_t \mathbf{e}_{\boldsymbol{\sigma}}^m\right\|_{\mathcal{A}}^2 + \left(\frac{\epsilon_1}{2} + \frac{\epsilon_2}{2}\right)\left\|\partial_t e_T^m\right\|^2 + \gamma_1\theta_M\frac{1}{2\epsilon_1}\left\|\mathbf{e}_{\mathbf{w}}^m\right\|^2 + \gamma_2\theta_M\frac{1}{2\epsilon_2}\left\|\mathbf{e}_{\mathbf{r}}^{m-1}\right\|^2,
\end{aligned} \tag{4.13a}
$$

for some $\epsilon_1, \epsilon_2 > 0$. Combining this with the previous estimate (4.6) and setting $\epsilon_1 = \epsilon_2 = \dfrac{\alpha\beta}{\mu + \lambda}$ leads to

$$
\begin{aligned}
&(c_0 - b_0)\left\|\partial_t e_p^m\right\|^2 + (a_0 - b_0)\left\|\partial_t e_T^m\right\|^2 + \frac{k_m}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{e}_{\mathbf{w}}^m\right\|^2 + \frac{\theta_m}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{e}_{\mathbf{r}}^m\right\|^2 \\
&\quad \le \frac{\theta_M}{2}\frac{\mu + \lambda}{\alpha\beta}\left(\gamma_1\left\|\mathbf{e}_{\mathbf{w}}^m\right\|^2 + \gamma_2\left\|\mathbf{e}_{\mathbf{r}}^{m-1}\right\|^2\right).
\end{aligned} \tag{4.13b}
$$

Integrating (4.13b) from 0 to t and applying the Grönwall inequality yields

$$
\begin{aligned}
&(c_0 - b_0)\int_0^t \left\|\partial_t e_p^m(\tau)\right\|^2 \mathrm{d}\tau + (a_0 - b_0)\int_0^t \left\|\partial_t e_T^m(\tau)\right\|^2 \mathrm{d}\tau + \frac{k_m}{2}\left\|\mathbf{e}_{\mathbf{w}}^m(t)\right\|^2 + \frac{\theta_m}{2}\left\|\mathbf{e}_{\mathbf{r}}^m(t)\right\|^2 \\
&\quad \le \frac{\xi\gamma_2}{2}\exp\left(\frac{\xi\gamma_1}{k_m}T_f\right)\int_0^t \left\|\mathbf{e}_{\mathbf{r}}^{m-1}(\tau)\right\|^2 \mathrm{d}\tau \le \frac{\xi\gamma_2}{2}\exp\left(\frac{\xi\gamma_1}{k_m}T_f\right)\int_0^{t_1} \left\|\mathbf{e}_{\mathbf{r}}^{m-1}(\tau)\right\|^2 \mathrm{d}\tau,
\end{aligned} \tag{4.13c}
$$

for $t \le t_1$ where $t_1 > 0$ will be fixed later, and where $\xi := \theta_M\dfrac{\mu + \lambda}{\alpha\beta}$. Integrating in time once more from 0 to t_1 yields

$$
\int_0^{t_1} \left\|\mathbf{e}_{\mathbf{r}}^m(\tau)\right\|^2 \mathrm{d}\tau \le t_1 L\int_0^{t_1} \left\|\mathbf{e}_{\mathbf{r}}^{m-1}(\tau)\right\|^2 \mathrm{d}\tau, \tag{4.13d}
$$

where the constant $L = \dfrac{\xi\gamma_2}{2}\exp\left(\dfrac{\xi\gamma_1}{k_m}T_f\right)$ is such that $0 < L < \infty$ provided $T_f < \infty$, and is independent of m and of the local final time t_1. Thus, for $t_1 = \dfrac{1}{2L}$ the above expression implies that the map $\mathbf{e}_{\mathbf{r}}^{m-1}(t) \mapsto \mathbf{e}_{\mathbf{w}}^m(t)$ is a contraction map for $t \in (0, t_1]$. In particular, this implies that as $m \to \infty$ we have from the Banach Fixed Point Theorem [11] and (4.5b)–(4.8), (4.9b)–(4.11), (4.12b) and (4.13c) the following convergences

- $\mathbf{e}_{\mathbf{w}}^m, \mathbf{e}_{\mathbf{r}}^m \to 0$ in $L^2(0, t_1; H(\mathrm{div}, \Omega)) \cap L^\infty(0, t_1; L^2(\Omega))$,
- $e_p^m, e_T^m \to 0$ in $H^1(0, t_1; L^2(\Omega))$,
- $\mathbf{e}_{\boldsymbol{\sigma}}^m \to 0$ in $H^1(0, t_1; L^2(\Omega)) \cap L^2(0, t_1; H_s(\mathrm{div}, \Omega))$,
- $\mathbf{e}_{\mathbf{u}}^m \to 0$ in $H^1(0, t_1; L^2(\Omega))$.

Therefore, the existence of the solution to problem (2.8) is established for $t \in (0, t_1]$. The question now is how to continue the local solution $(T, \mathbf{r}, p, \mathbf{w}, \boldsymbol{\sigma}, \mathbf{u})$ to the system (2.8) globally in time. To this aim, we let $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m)$ be the solution of (4.1) on the time interval $[t_{k-1}, t_k]$, $k \in \mathbb{N}$, with $t_k - t_{k-1} = \dfrac{1}{2L}$, and starting with the initial data $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m)(\cdot, t_{k-1}) = (T, \mathbf{r}, p, \mathbf{w}, \boldsymbol{\sigma}, \mathbf{u})|_{[t_{k-2}, t_{k-1}]}(\cdot, t_{k-1})$; thanks to the continuity in-time of the convergent solution. The iterates $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m)$ are again well-defined using Theorem 3.1. The iterates also result a contraction, i.e.,

$$\int_{t_{k-1}}^{t_k} \|\mathbf{e}_{\mathbf{r}}^m(\tau)\|^2 \,\mathrm{d}\tau \leq \frac{1}{2} \int_{t_{k-1}}^{t_k} \|\mathbf{e}_{\mathbf{r}}^{m-1}(\tau)\|^2 \,\mathrm{d}\tau, \quad \forall k \geq 2. \tag{4.14}$$

Therefrom, we proceed as on done in the first time interval $[0, t_1]$ to show the convergence of the successive approximations $(T^m, \mathbf{r}^m, p^m, \mathbf{w}^m, \boldsymbol{\sigma}^m, \mathbf{u}^m)|_{[t_{k-1}, t_k]}$, $k \in \mathbb{N}$, to $(T, \mathbf{r}, p, \mathbf{w}, \boldsymbol{\sigma}, \mathbf{u})|_{[t_{k-1}, t_k]}$. This solution is similarly extended to any time $t_\ell \geq t_k$ given by

$$t_\ell = \sum_{k=1}^{\ell} t_k - t_{k-1} = \sum_{k=1}^{\ell} \frac{1}{2L}.$$

Finally, since the series $\sum_{k=1}^{\infty} \frac{1}{2L}$ diverges, the sequence of local solutions is extended to arbitrary finite final time $0 < T_f < \infty$ by incrementing the values ℓ (if T_f is not identically an integer multiple of $\frac{1}{2L}$ take instead $t_k - t_{k-1} = \frac{1}{NL}$ where $N > 1$). This concludes the proof of Theorem 4.1. □

Some remarks on the above proof are in order.

Remark 4.2. We could also define a fully explicit iterative scheme where both the Darcy and heat fluxes in the convective term are given at the previous iteration. If such an explicit scheme was chosen we would have the advantage of a symmetric linearized problem, as the convective terms in the iterative procedure can be viewed as part of the source term on the right hand side.

Remark 4.3. Assume that $\mathbf{f}$ is in $H^1(J; L^2(\Omega))$, g, h in $L^2(J; L^2(\Omega))$, p_0, T_0 in $H_0^1(\Omega)$, and $\mathbf{u}_0$ in $(L^2(\Omega))^d$. Suppose that instead of Hypothesis 1, we have $\mathbf{r}^m$, $\mathbf{w}^m$ in $\in H^1(0, T; L^\infty(\Omega))$. Then, we can reproduce the proof of Theorem 4.1 to prove the convergence of the scheme given by (4.1) to a weak solution of the nonlinear problem (2.8).

5. Conclusions

In this article we have given mixed formulations for the fully coupled quasi-static thermo-poroelastic model. The model is nonlinear, with the nonlinearity appearing on a coupling term. This makes the analysis challenging. A linearization of the model was therefore employed as an intermediate step in analyzing the full nonlinear model. For the linear case, the well-posedness is established using the theory of DAEs, and energy estimates together with a Galerkin method. This result together with derived energy estimates are combined with the Banach Fixed Point Theorem to obtain local solutions in time of the nonlinear problem. Due to the continuity in time of the convergent (local) solutions, we can infer a (global) convergence proof of an iterative procedure approximating the weak solution to the original nonlinear problem. Work underway addresses discretization of this model problem using an appropriate mixed finite element method as well as *a priori* and *a posteriori* error analysis.

Acknowledgments

This work forms part of Norwegian Research Council project 250223. The authors would also like to thank Kundan Kumar and Prof. Ludmil Zikatanov for very helpful discussions concerning the analysis of the nonlinear model.

Appendix A. Alternative to Hypothesis 1

We outline some formal calculations which reveal the assumptions necessary on the data in order to avoid the Hypothesis 1. In particular, we aim to solve the linear Problem 3.1 with sufficiently regular data such that $\mathbf{r} \in C([0,T_f];L^\infty(\Omega))$ (or, alternatively such that $\mathbf{w},\mathbf{r} \in C([0,T_f];L^4(\Omega))$. The following arguments indicate that this is easily done. First, note that from Theorem 3.1 and the Sobolev Embedding Theorem (see e.g. [15]) it follows that the functions $(T(t),\mathbf{r}(t),p(t),\mathbf{w}(t),\boldsymbol{\sigma}(t),\mathbf{u}(t))$ are continuous for $t \in [0,T_f]$, if $g,h \in H^1(J;L^2(\Omega))$ and $\mathbf{f} \in H^2(J;L^2(\Omega))$. Thus, going back to the problem (3.1), we can choose smooth test functions with compact support in Ω and find that $(T,\mathbf{r},p,\mathbf{w},\boldsymbol{\sigma},\mathbf{u})$ solves the following initial boundary value problem

$$a_0\frac{\mathrm{d}T}{\mathrm{d}t}(t) - b_0\frac{\mathrm{d}p}{\mathrm{d}t}(t) + \frac{a_r}{2\beta}\frac{\mathrm{d}\,\mathrm{tr}\,\boldsymbol{\sigma}}{\mathrm{d}t}(t) - \boldsymbol{\eta}\cdot\mathbf{w}(t) + \nabla\cdot\mathbf{r}(t) = h(t), \qquad \text{in } \Omega, \tag{A.1a}$$

$$\boldsymbol{\Theta}^{-1}\mathbf{r}(t) + \nabla T(t) = 0, \qquad \text{in } \Omega, \tag{A.1b}$$

$$c_0\frac{\mathrm{d}p}{\mathrm{d}t}(t) - b_0\frac{\mathrm{d}T}{\mathrm{d}t}(t) + \frac{c_r}{2\alpha}\frac{\mathrm{d}\,\mathrm{tr}\,\boldsymbol{\sigma}}{\mathrm{d}t}(t) + \nabla\cdot\mathbf{w}(t) = g(t), \qquad \text{in } \Omega, \tag{A.1c}$$

$$\mathbf{K}^{-1}\mathbf{w}(t) + \nabla p(t) = 0, \qquad \text{in } \Omega, \tag{A.1d}$$

$$\mathcal{A}\boldsymbol{\sigma}(t) - \boldsymbol{\varepsilon}(\mathbf{u})(t) + \frac{c_r}{2\alpha}\mathbf{I}p(t) + \frac{a_r}{2\beta}\mathbf{I}T(t) = 0, \qquad \text{in } \Omega, \tag{A.1e}$$

$$-\nabla\cdot\boldsymbol{\sigma}(t) = \mathbf{f}(t), \qquad \text{in } \Omega, \tag{A.1f}$$

for a.e. $t \in J$, and with boundary conditions

$$T = 0, \quad \mathbf{u} = 0, \quad p = 0, \quad \text{on } \Gamma\times J, \tag{A.1g}$$

and initial conditions

$$T(0) = T_0, \quad \mathbf{u}(0) = \mathbf{u}_0, \quad \text{and} \quad p(0) = p_0, \quad \text{in } \Omega\times\{0\}. \tag{A.1h}$$

Since $\boldsymbol{\Theta}^{-1}\mathbf{r},\mathbf{K}^{-1}\mathbf{w} \in L^2(J;L^2(\Omega))$ we have from (A.1b) and (A.1d) that $T,p \in L^2(J;H_0^1(\Omega))$. Thus, we can write (A.1a) and (A.1c) in non-mixed form, i.e.

$$a_0\frac{\mathrm{d}T}{\mathrm{d}t}(t) - b_0\frac{\mathrm{d}p}{\mathrm{d}t}(t) + \frac{a_r}{2\beta}\frac{\mathrm{d}\,\mathrm{tr}\,\boldsymbol{\sigma}}{\mathrm{d}t}(t) - \boldsymbol{\eta}\cdot\mathbf{w}(t) - \nabla\cdot(\boldsymbol{\Theta}\nabla T(t)) = h(t), \tag{A.2a}$$

$$c_0\frac{\mathrm{d}p}{\mathrm{d}t}(t) - b_0\frac{\mathrm{d}T}{\mathrm{d}t}(t) + \frac{c_r}{2\alpha}\frac{\mathrm{d}\,\mathrm{tr}\,\boldsymbol{\sigma}}{\mathrm{d}t}(t) - \nabla\cdot(\mathbf{K}\nabla p(t)) = g(t), \tag{A.2b}$$

and use the theory of linear parabolic equations (see [15] p. 349 for details) to get increased regularity for $T(t)$ and $p(t)$, and then use (A.1b) and (A.1d) to infer increased regularity for $\mathbf{r}(t)$ and $\mathbf{w}(t)$. In particular, if the domain boundary Γ is of class C^1, $h,g \in C^1([0,T_f];H^1(\Omega))$, $\mathbf{f} \in C^2([0,T_f];L^2(\Omega))$ and $T_0 \in H_0^1(\Omega)\cap H^2(\Omega)$, then $T \in H^1(J;H^2(\Omega))$ and thus $\mathbf{r} \in H^1(J;H^1(\Omega))$. Due to the special case of the Sobolev embedding theorem for $d = 2$, i.e. $H^1(\Omega) \subset L^\infty(\Omega)$ ([15] p. 270), we get that $\mathbf{r} \in C([0,T_f];L^\infty(\Omega))$. Alternatively, if also $p_0 \in H_0^1(\Omega)\cap H^2(\Omega)$, then we have additionally $\mathbf{w} \in H^1(J;H^1(\Omega))$, and since $H^1(\Omega) \subset L^4(\Omega)$ (independently of spatial dimension), we get $\mathbf{r},\mathbf{w} \in C([0,T_f];L^4(\Omega))$.

Appendix B. Tables

For easy reference we list some of the notations used in this article (Tables 1, 2).

Table 1
Data and parameters.

Data/Parameter	Description
h	heat source
$\mathbf{f}$	body force
g	mass source
T_0	initial temperature distribution
$\mathbf{u}_0$	initial displacement
p_0	initial fluid pressure
a_0	effective thermal capacity
b_0	thermal dilation coefficient
β	thermal stress coefficient
$\mathbf{K}$	matrix permeability divided by fluid viscosity
$\boldsymbol{\Theta}$	effective thermal conductivity
μ, λ	Lamé parameters
α	Biot–Willis constant
c_0	specific storage coefficient

Table 2
Variables.

Variable	Description	Spaces
T	temperature distribution	$\mathcal{T} := L^2(\Omega)$
$\mathbf{u}$	solid displacement	$\mathcal{U} := (L^2(\Omega))^d$
p	fluid pressure	$\mathcal{P} := L^2(\Omega)$
$\boldsymbol{\sigma}$	total stress	$\mathcal{S} := H_s(\mathrm{div}; \Omega)$
$\mathbf{w}$	Darcy flux	$\mathcal{W} := H(\mathrm{div}; \Omega)$
$\mathbf{r}$	heat flux	$\mathcal{R} := H(\mathrm{div}; \Omega)$

References

[1] E. Ahmed, F.A. Radu, J.M. Nordbotten, A posteriori error estimates and adaptivity for fully mixed finite element discretizations for Biot's consolidation model, working paper or preprint, https://hal.inria.fr/hal-01687026, Jan. 2018.
[2] D.N. Arnold, R.S. Falk, R. Winther, Mixed finite element methods for linear elasticity with weakly imposed symmetry, Math. Comp. 76 (2007) 1699–1723, https://doi.org/10.1090/S0025-5718-07-01998-9.
[3] D.N. Arnold, R. Winther, Mixed finite elements for elasticity, Numer. Math. 92 (2002) 401–419, https://doi.org/10.1007/s002110100348.
[4] T. Bærland, J.J. Lee, K.-A. Mardal, R. Winther, Weakly imposed symmetry and robust preconditioners for Biot's consolidation model, Comput. Methods Appl. Math. 17 (2017) 377–396, https://doi.org/10.1515/cmam-2017-0016.
[5] M. Bause, F.A. Radu, U. Köcher, Space–time finite element approximation of the Biot poroelasticity system with iterative coupling, Comput. Methods Appl. Mech. Engrg. 320 (2017) 745–768, https://doi.org/10.1016/j.cma.2017.03.017.
[6] M.A. Biot, General theory of three-dimensional consolidation, J. Appl. Phys. 12 (1941) 155–164.
[7] M.A. Biot, Theory of finite deformations of porous solids, Indiana Univ. Math. J. 21 (1972) 597–620.
[8] J.W. Both, M. Borregales, J.M. Nordbotten, K. Kumar, F.A. Radu, Robust fixed stress splitting for Biot's equations in heterogeneous media, Appl. Math. Lett. 68 (2017) 101–108, https://doi.org/10.1016/j.aml.2016.12.019.
[9] K.E. Brenan, S.L. Campbell, L.R. Petzold, Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations, Classics Appl. Math., vol. 14, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Revised and corrected reprint of the 1989 original.
[10] M.K. Brun, I. Berre, J.M. Nordbotten, F.A. Radu, Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium, Transp. Porous Media (2018), https://doi.org/10.1007/s11242-018-1056-8.
[11] W. Cheney, Analysis for Applied Mathematics, Grad. Texts in Math., vol. 208, Springer-Verlag, New York, 2001.
[12] D. Cioranescu, P. Donato, An Introduction to Homogenization, Oxford Lecture Ser. Math. Appl., vol. 17, The Clarendon Press, Oxford University Press, New York, 1999.
[13] O. Coussy, Poromechanics, John Wiley & Sons, 2004.
[14] E. Detournay, A.H.-D. Cheng, Fundamentals of poroelasticity, in: Analysis and Design Methods, Elsevier, 1995, pp. 113–171.
[15] L.C. Evans, Partial Differential Equations, second ed., Grad. Stud. Math., vol. 19, American Mathematical Society, Providence, RI, 2010.
[16] B. Gatmiri, P. Delage, A formulation of fully coupled thermal–hydraulic–mechanical behaviour of saturated porous media—numerical approach, Int. J. Numer. Anal. Methods Geomech. 21 (1997) 199–225.
[17] G.H. Hardy, J.E. Littlewood, G. Pólya, Inequalities, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1988. Reprint of the 1952 edition.
[18] U. Hornung, Homogenization and Porous Media, vol. 6, Springer Science & Business Media, 2012.
[19] J.K. Hunter, Nonlinear Evolution Equations, University of California, Davis, 1996.
[20] J.J. Lee, Robust error analysis of coupled mixed methods for Biot's consolidation model, J. Sci. Comput. 69 (2016) 610–632, https://doi.org/10.1007/s10915-016-0210-0.

[21] J.J. Lee, K.-A. Mardal, R. Winther, Parameter-robust discretization and preconditioning of Biot's consolidation model, SIAM J. Sci. Comput. 39 (2017) A1–A24, https://doi.org/10.1137/15M1029473.
[22] C.K. Lee, C.C. Mei, Thermal consolidation in porous media by homogenization theory—I. Derivation of macroscale equations, Adv. Water Resour. 20 (1997) 127–144.
[23] F. List, F.A. Radu, A study on iterative methods for solving Richards' equation, Comput. Geosci. 20 (2016) 341–353, https://doi.org/10.1007/s10596-016-9566-3.
[24] A. Mikelić, M.F. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics, Comput. Geosci. 17 (2013) 455–461, https://doi.org/10.1007/s10596-012-9318-y.
[25] J.-C. Nédélec, A new family of mixed finite elements in $\mathbf{R}^3$, Numer. Math. 50 (1986) 57–81, https://doi.org/10.1007/BF01389668.
[26] P.J. Phillips, M.F. Wheeler, A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity, Comput. Geosci. 12 (2008) 417–435, https://doi.org/10.1007/s10596-008-9082-1.
[27] I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure, J. Comput. Appl. Math. 168 (2004) 365–373, https://doi.org/10.1016/j.cam.2003.04.008.
[28] F.A. Radu, J.M. Nordbotten, I.S. Pop, K. Kumar, A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media, J. Comput. Appl. Math. 289 (2015) 134–141, https://doi.org/10.1016/j.cam.2015.02.051.
[29] P.-A. Raviart, J.M. Thomas, A mixed finite element method for 2nd order elliptic problems, Lecture Notes in Math. 606 (1977) 292–315.
[30] R.E. Showalter, Diffusion in poro-elastic media, J. Math. Anal. Appl. 251 (2000) 310–340, https://doi.org/10.1006/jmaa.2000.7048.
[31] A.P. Suvorov, A.P.S. Selvadurai, Macroscopic constitutive equations of thermo-poroviscoelasticity derived using eigenstrains, J. Mech. Phys. Solids 58 (2010) 1461–1473, https://doi.org/10.1016/j.jmps.2010.07.016.
[32] K. Terzaghi, Theoretical Soil Mechanics, Chapman and Hall, Limited John Wiler and Sons, Inc., New York, 1944.
[33] J.M. Thomas, Méthode des éléments finis équilibre, in: Journées "Éléments Finis", Rennes, 1975, Univ. Rennes, Rennes, 1975, p. 25.
[34] C.J. Van Duijn, A. Mikelic, M. Wheeler, T. Wick, Thermoporoelasticity via homogenization I. Modeling and formal two-scale expansions, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01650194, Nov. 2017.
[35] C.J. van Duijn, I.S. Pop, Crystal dissolution and precipitation in porous media: pore scale analysis, J. Reine Angew. Math. 577 (2004) 171–211, https://doi.org/10.1515/crll.2004.2004.577.171.
[36] H.F. Wang, Theory of Linear Poroelasticity with Applications to Geomechanics and Hydrogeology, Princeton University Press, 2017.
[37] S.-Y. Yi, Convergence analysis of a new mixed finite element method for Biot's consolidation model, Numer. Methods Partial Differential Equations 30 (2014) 1189–1210, https://doi.org/10.1002/num.21865.
[38] K. Yosida, Functional Analysis, Classics Math., Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

Paper C

Monolithic and splitting solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport

M. K. Brun, E. Ahmed, I. Berre, J. M. Nordbotten, F. A. Radu

In review

MONOLITHIC AND SPLITTING SOLUTION SCHEMES FOR FULLY COUPLED QUASI-STATIC THERMO-POROELASTICITY WITH NONLINEAR CONVECTIVE TRANSPORT*

MATS K. BRUN†, ELYES AHMES†, INGA BERRE†‡, JAN M. NORDBOTTEN†§, AND FLORIN A. RADU†

Abstract. This paper concerns monolithic and splitting-based iterative procedures for the coupled nonlinear thermo-poroelasticity model problem. The thermo-poroelastic model problem we consider is formulated as a three-field system of PDE's, consisting of an energy balance equation, a mass balance equation and a momentum balance equation, where the primary variables are temperature, fluid pressure, and elastic displacement. Due to the presence of a nonlinear convective transport term in the energy balance equation, it is convenient to have access to both the pressure and temperature gradients. Hence, we introduce these as two additional variables and extend the original three-field model to a five-field model. For the numerical solution of this five-field formulation, we compare six approaches that differ by how we treat the coupling/decoupling between the flow and/from heat and/from the mechanics, suitable for varying coupling strength between the three physical processes. The approaches have in common a simultaneous application of the L-scheme, which works both to stabilize iterative splitting as well as to linearize nonlinear problems, and can be seen as a generalization of the Undrained and Fixed-Stress Split algorithms. More precisely, the derived procedures transform a nonlinear and fully coupled problem into a set of simpler subproblems to be solved sequentially in an iterative fashion. We provide a convergence proof for the derived algorithms, and demonstrate their performance through several numerical examples investigating different strengths of the coupling between the different processes.

Key words. Quasi-static thermo-poroelasticity, nonlinear convective transport, porous media, monolithic scheme, fixed-stress splitting iterative coupling, L-scheme linearization, contraction mapping, mixed finite elements.

AMS subject classifications. 65M02, 65Z02, 74F02

1. Introduction.

1.1. Problem statement.

The field of *poroelasticity* aims to describes the interaction between viscous fluid flow and elastic solid deformation within a porous material, and was pioneered through the works of K. Terzhagi [43] and M. A. Biot [5, 6]. In the fully-saturated, quasi-static regime, the mathematical modeling of such processes constitutes a coupled two-field linear model where the primary variables are the fluid pressure and the elastic displacement of the solid. This is known as the quasi-static Biot's model.

In many important applications, such as geothermal energy extraction, nuclear waste disposal and carbon storage, temperature also plays a vital role and must therefore be included in the aforementioned model. Thus, we consider here a *thermo-poroelastic* system which can be seen as a generalization of the Biot system to the non-isothermal case; i.e., the coupled processes are heat, flow, and geomechanics. Since it is the cornerstone of many complex models, we focus on the following nonlinear and coupled quasi-static thermo-poroelastic equations as described in [10, 30, 46]: Find

*Submitted to the editors 04.03.19.

Funding: This work forms part of Norwegian Research Council project 250223

†Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020 Bergen, Norway (mats.brun@uib.no, elyes.ahmed@uib.no, inga.berre@uib.no, jan.nordbotten@uib.no, florin.radu@uib.no).

‡NORCE Norwegian Research Centre AS, Bergen, Norway.

§Department of Civil and Environmental Engineering, Princeton University, Princeton, N. J., USA.

the temperature T, the pressure p, and the displacement $\mathbf{u}$ such that

$$\partial_t \psi(p,\mathbf{u},T) + c_f(\mathbf{K}\nabla p)\cdot\nabla T - \nabla\cdot(\boldsymbol{\Theta}\nabla T) = z, \quad \text{in } \Omega\times(0,t_f), \tag{1a}$$

$$-\nabla\cdot\boldsymbol{\theta}(\mathbf{u}) + \alpha\nabla p + \beta\nabla T = \mathbf{f}, \quad \text{in } \Omega\times(0,t_f), \tag{1b}$$

$$\partial_t \varphi(p,T,\mathbf{u}) - \nabla\cdot(\mathbf{K}\nabla p) = g, \quad \text{in } \Omega\times(0,t_f), \tag{1c}$$

$$T(\cdot,0) = T_0, \quad \mathbf{u}(\cdot,0) = \mathbf{u}_0, \quad p(\cdot,0) = p_0, \quad \text{in } \Omega, \tag{1d}$$

$$T = 0, \quad \mathbf{u} = 0, \quad p = 0, \quad \text{on } \partial\Omega\times(0,t_f). \tag{1e}$$

In the above model, Ω is a bounded (connected and open) domain in $\mathbb{R}^d$, $d = 2$ or 3, and $t_f > 0$ is the final time. The function z is the heat source, g is the mass source, and $\mathbf{f}$ is the body force. The functionals ψ and φ denote the heat content and fluid content, respectively; i.e., $\psi(p,\mathbf{u},T) := a_0T - b_0p + \beta\nabla\cdot\mathbf{u}$, and $\varphi(p,\mathbf{u},T) := c_0p - b_0T + \alpha\nabla\cdot\mathbf{u}$, where c_0 is the constrained-specific storage coefficient, a_0 is the effective thermal capacity, b_0 is the thermal dilation coefficient, α is the Biot–Willis constant, and β is the thermal stress coefficient. The parameter c_f is the volumetric heat capacity of the fluid, $\mathbf{K} = (K_{ij})_{i,j=1}^d$ is the permeability divided by fluid viscosity, and $\boldsymbol{\Theta} = (\Theta_{ij})_{i,j=1}^d$ is the effective thermal conductivity. The function $\boldsymbol{\theta}$ denotes the effective stress tensor, i.e., $\boldsymbol{\theta}(\mathbf{u}) := 2\mu\boldsymbol{\varepsilon}(\mathbf{u}) + \lambda\nabla\cdot\mathbf{u}\mathbf{I}$, where $\boldsymbol{\varepsilon}(\mathbf{u}) := (\nabla\mathbf{u} + \nabla\mathbf{u}^\top)/2$ the symmetric part of $\nabla\mathbf{u}$, and $\mathbf{I}$ is the identity tensor. Finally, T_0 is the initial temperature, $\mathbf{u}_0$ is the initial displacement and p_0 is the initial pressure.

Note that the above model introduces a nonlinearity in a coupling term, which is the convective transport term in the energy balance equation (1a). The presence of this nonlinear coupling term strongly complicates the problem compared to the isothermal case (i.e., the linear Biot's model). Note that if $b_0 = \beta = 0$, the flow and mechanics decouples from the heat, and Biot's model is recovered. For the derivation of the constitutive equations of thermo-poroelasticity we refer to the works [23, 42, 46], and particularly to [10, 30, 46] where the above model was derived within the framework of the two-scale asymptotic expansion method (see, *e.g.,* [24] for a review of this technique).

REMARK 1.1 (Conservative energy). *The energy balance equation* (1a) *can equivalently be written in conservative form [16], whence the second term takes the form* $\nabla\cdot(c_f(\mathbf{K}\nabla p)T)$. *All results presented in the following remain valid for both formulations.*

1.2. Weak solution and well-posedness of the continuous problem. The common structure of mathematical models which are based on (systems of) scalar conservation laws of the form (1a) and where nonlinear gradient terms appear, suggests introducing the heat flux, $\mathbf{r} := -\boldsymbol{\Theta}\nabla T$, or the Darcy flux, $\mathbf{w} := -\mathbf{K}\nabla p$, as an additional variable. Thus, either the term $c_f(\mathbf{K}\nabla p)\cdot\nabla T$ becomes $[-c_f(\mathbf{w}\cdot\nabla T)]$ or $\big[-c_f((\mathbf{K}\otimes\boldsymbol{\Theta}^{-1})\mathbf{r}\cdot\nabla p)\big]$, *e.g.* [41, 40]. Precisely, it is well known that such terms, dealing non-linearly with the coupled convection, can be quite difficult to approximate correctly in their actual forms. This altogether leads to challenging numerical issues. Furthermore, the choice to introduce the heat flux or the Darcy flux as a new variable depends strongly on which process (flow or heat) that dominates, and may result in a different treatment of the convective term. Here, to avoid some of these complexities, we adopt from [9] the mixed form for both the heat and flow subproblems (1a) and (1b), taking in mind that Mixed Finite Element (also Finite Volume) literature has developed techniques to handle convective terms [17, 14]. Throughout the paper, we assume that the following assumptions hold true:

(**A1**) $\mathbf{K} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is assumed to be constant in time, symmetric, definite and positive; there exist $k_m > 0$ and k_M such that $k_m|\zeta|^2 \le \zeta^\top \mathbf{K}(x)\zeta$ and $|\mathbf{K}(x)\zeta| \le k_M|\zeta|,\ \forall \zeta \in \mathbb{R}^d \setminus \{0\}$.

(**A2**) $\boldsymbol{\Theta} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is assumed to be constant in time, symmetric, definite and positive; there exist $\theta_m > 0$ and θ_M such that $\theta_m|\zeta|^2 \le \zeta^\top \boldsymbol{\Theta}(x)\zeta$ and $|\boldsymbol{\Theta}(x)\zeta| \le \theta_M|\zeta|,\ \forall \zeta \in \mathbb{R}^d \setminus \{0\}$.

(**A3**) The coefficients a_0, b_0, c_0, c_f, α and β are strictly positive constants.

(**A4**) The coefficients a_0, b_0 and c_0 are such that $c_0 - b_0 > 0$ and $a_0 - b_0 > 0$.

(**A5**) The source terms are such that $z, g \in L^2(0,t_f;L^2(\Omega))$ and $\mathbf{f} \in H^1(0,t_f;L^2(\Omega))$. We further assume that z, g and $\mathbf{f}$ are piecewise constant in time with respect to the temporal mesh of Section 2.

(**A6**) The initial data are such that $p_0, T_0 \in H_0^1(\Omega)$ and $\mathbf{u}_0 \in (L^2(\Omega))^d$.

Before transcribing the mixed variational formulation of the problem, we introduce some notations:

$$\mathcal{T} := L^2(\Omega),\quad \mathcal{R} := H(\mathrm{div},\Omega),\quad \mathcal{P} := L^2(\Omega),\quad \mathcal{W} := H(\mathrm{div},\Omega),\quad \mathcal{U} := (L^2(\Omega))^d,$$

where we denote by $(\cdot,\cdot)$ the standard $L^2(\Omega)$ inner product, and by $\|\cdot\|$ the induced $L^2(\Omega)$ norm. Due to (**A1**) and (**A2**), the tensors $\mathbf{K}$ and $\boldsymbol{\Theta}$ (and their inverses) define $L^2(\Omega)$-equivalent norms, which we denote by $\|\mathbf{v}\|_{\mathbf{K}} := (\mathbf{K}\mathbf{v},\mathbf{v})^{1/2}$ (and $\|\mathbf{v}\|_{\mathbf{K}^{-1}} := (\mathbf{K}^{-1}\mathbf{v},\mathbf{v})^{1/2}$), and similarly with $\boldsymbol{\Theta}$. With this, we define the variational formulation of (1a)–(1d) as follows:

DEFINITION 1.1 (The continuous formulation [9]). *Assuming* (**A1**)–(**A6**) *holds true, the fully coupled mixed-primal formulation of* (1) *reads:*
Find $(T(t),\mathbf{r}(t),p(t),\mathbf{w}(t),\mathbf{u}(t)) \in \mathcal{T}\times\mathcal{R}\times\mathcal{P}\times\mathcal{W}\times\mathcal{U}$, *such that for a.e.* $t \in (0,t_f)$

$$(\partial_t\psi(p,T,\mathbf{u}),S) + c_f(\mathbf{w}\cdot\boldsymbol{\Theta}^{-1}\mathbf{r},S) + (\nabla\cdot\mathbf{r},S) = (z,S), \quad \forall S\in\mathcal{T}, \tag{2a}$$

$$(\boldsymbol{\Theta}^{-1}\mathbf{r},\mathbf{y}) - (T,\nabla\cdot\mathbf{y}) = 0, \quad \forall \mathbf{y}\in\mathcal{R}, \tag{2b}$$

$$(\partial_t\varphi(p,T,\mathbf{u}),q) + (\nabla\cdot\mathbf{w},q) = (g,q), \quad \forall q\in\mathcal{P}, \tag{2c}$$

$$(\mathbf{K}^{-1}\mathbf{w},\mathbf{z}) - (p,\nabla\cdot\mathbf{z}) = 0, \quad \forall \mathbf{z}\in\mathcal{W}, \tag{2d}$$

$$(\boldsymbol{\theta}(\mathbf{u}),\boldsymbol{\varepsilon}(\mathbf{v})) - (\beta T + \alpha p,\nabla\cdot\mathbf{v}) = (\mathbf{f},\mathbf{v}), \quad \forall \mathbf{v}\in\mathcal{U}, \tag{2e}$$

together with the initial conditions (1e).

The above variational problem was analyzed in [9]. There, it was shown that under the assumption that the heat flux (or Darcy flux) is such that $\mathbf{r}(t) \in (L^\infty(\Omega))^d$, for $t \in (0,t_f)$, the problem (2) has a unique weak solution. Moreover, it was shown that with additional regularity on the data, i.e., $\mathbf{f} \in H^2\left(0,t_f;(L^2(\Omega))^d\right)$, $h, g \in H^1(0,t_f;L^2(\Omega))$, and $T_0, p_0 \in H_0^1(\Omega)\cap H^2(\Omega)$, the fluxes are bounded functions.

1.3. Goal and positioning of the paper. The simulation of thermo-poroelasticity problems is challenging due to the coexistence of different physics, necessitating a coupled set of equations. For these types of problems, there are typically three different approaches employed in modeling fluid flow coupled with reservoir geomechanics. They are known as the fully implicit, the explicit (loosely or weakly) coupling, and the *splitting-iterative* approaches. The main problem for the applicability of the fully implicit approach, which solves simultaneously the above *three-processes* (flow, heat and mechanics) problem, is that it results in a very large system of equations to be solved at each time step. Moreover, it does not facilitate the (re-)use of existing codes dedicated to the various subproblems. On the other hand, the fully

coupled approach has excellent stability properties [4, 18]. An alternative is weakly coupled approaches, which results in smaller systems and a lower computational cost compared to the fully implicit (monolithic) approach. On the other hand, accuracy may be sacrificed, and the sequential approach is only conditionally stable [20, 34]. Herein, we adopt an iterative coupling approach, which provides a compromise between the implicit and explicit: At each iteration it has the cost of the sequential approach, yet it converges to the fully coupled implicit approach. We implement the idea of iterative coupling by resolving iteratively the two/three subsystems (depending on the choice of splitting procedure) and by exchanging the values of the shared state variables in an iterative fashion using a general framework of linearly stabilized schemes [8, 32].

We argue that adopting an iterative method for the *nonlinear and fully coupled three-processes* problem, can be considered almost essential for efficient simulation, since the fully coupled approach leads to a prohibitivily large system (particularly if MFE methods are adopted [1, 12, 21, 47]), incorporating different equations varied in type and with nonlinearities. The advantage of the iterative approaches considered in this paper is that, at each iteration, smaller, easier-to-solve systems cooperate iteratively through algorithms [21, 11]. Another advantage that distinguishes our approaches is the possibility to *reuse existing codes* for different numerical schemes and coupling techniques specialized to each component of the problem (see e.g. [2, 35]). For classical linear poroelasticity, the iterative coupling procedures mentioned in the above has been studied extensively [3, 8, 13, 25, 27, 28, 32, 33, 45]. In particular, two such algorithms have received considerable attention: The "Undrained Split"(constant fluid mass during structure deformation), and the "Fixed Stress Split"(constant volumetric mean total stress during solution of flow problem). In [27] these were first shown to be unconditionally stable. In [32, 33] contraction estimates and rates of convergence were derived.

The Undrained Split/Fixed Stress Split algorithms have been generalized in the context of the so-called *L-schemes*. In the context of coupled problems, these schemes involves adding an artificial stabilization term to one or more of the subproblems with a parameter $L > 0$. Here, the quantity held constant during solving of one of the subproblems needs not have any physical interpretation. In this sense, the L-scheme generalizes the Undrained Split/Fixed Stress Split algorithms and, due to the removal of physical constraints on the stabilization terms, allows for further optimization. The L-scheme can also be employed as a linearization procedure for nonlinear problems, with the parameter $L > 0$ mimicking the Jacobian from Newton iteration. However, in order to determine the parameter $L > 0$ for any given problem, derived convergence estimates are necessary. The L-scheme has been shown to perform robustly for Richards equation [31, 38] and for both linear and nonlinear coupled flow and geomechanics [7, 8]. In this paper, we will utilize the L-scheme framework both as a decoupling strategy and as a linearization method.

Although the literature on iterative coupling procedures for (isothermal) poroelastic problems is quite extensive, thermo-poroelastic problems have not received the same amount of attention. Sequential iterative methods for linear thermo-poroelasticity was considered in [26]. Iterative splitting schemes for separate poroelasticity and thermoelasticity problems were considered in [29]. Compared to problems of (two-field) coupled flow and mechanics (which can be solved either sequentially or monolithically) we now have additional options in partial decoupling, i.e., solving two of the subproblems together decoupled from the third. Combinatorially, this yields six combinations of iterative procedures, ranging from monolithic to fully decoupled.

In this work, we propose and analyze all six iterative algorithms for nonlinear thermo-poroelasticity based on these six combinations of coupling/decoupling. In particular, we employ variations of the L-scheme in all six algorithms, with artificial stabilization terms added to both the flow and heat sub-problems. By proving a contraction of all schemes, we obtain explicit expressions for the linearization parameters L that guarantees the stability and convergence of all schemes. The main *advantage* of the L-scheme is that it treats simultaneously the coupling and the non-linearity effects. Thus, no inner iterative approaches are required, see *e.g.* [39] where L-scheme type approaches are developed to treat iteratively a combined domain decomposition and nonlinearity problem. In most cases, the convergence is linear in the required energy norms. Furthermore, the necessary constraint on the time step is not severe.

The reason we propose six algorithms is the following: The coupling strength of the heat, flow and mechanics may vary depending on the physics at hand. Moreover, the practitioner may have access to existing software of various capabilities. Precisely, to develop robust and efficient solution procedures for the *three-processes problem* at hand, one should *in principle* take into account which process (*the mechanics* and/or *flow* and/or *heat flow*) dominate the full problem. In practice, one must also take into account implementation time and available frameworks. Thus, to be agnostic towards the dominating processes and other real-world constraints, we derive a complete framework for this model problem. The six variations of iterative coupling/decoupling algorithms for thermo-poroelasticity, cover all possibilities of varying coupling strength between the three physical processes involved. Note that developed algorithms are applicable on any numerical schemes used to obtain the solutions of the different processes [37, 48]. For the convergence analysis, we derive energy-type estimates, from which we infer the convergence of the iterate solutions as well as obtaining strict lower bounds on the stabilization parameters, and an upper bound on the time step. However, a "cut-off" operator $\mathcal{M}$ is introduced in the mixed setting in order to make the iterative schemes converge. Several numerical tests validate our proposed algorithms. In particular, we show that by using the derived stabilization estimates, the proposed algorithms perform robustly with respect to both mesh refinement and a wide range of different problem parameters.

The article is organized as follows: In Section 2 we present the fully discretization of the thermo-poroelasticity model, and in Section 3 we present all six iterative algorithms. In Section 4, convergence analysis based on contraction estimates are derived, from which the well-posedness of the discrete scheme is inferred in addition to the bounds on the stabilization parameters and time step. In Section 5 we provide several numerical experiments, and finally in Section 6 some concluding remarks.

2. Discrete setting. Let $\mathcal{X}_h$ be a simplicial mesh of Ω, matching in the sense that for two distinct elements of $\mathcal{X}_h$ their intersection is either an empty set or their common vertex or edge. Let h_K denote the diameter of $K \in \mathcal{X}_h$ and let h be the largest diameter of all such triangles, i.e., $h := \max_{K\in\mathcal{X}_h} h_K$. For the time partition, we let $\{t^n : n = 0, 1, \cdots, N\}$ be the discrete time steps, where $0 := t^0 < t^1 < \cdots < t^N = t_f$, and let $\tau^n = t^n - t^{n-1}$, $n \geq 1$, be the difference between consecutive discrete times. In other words, we have $t^n := \sum_{\ell=1}^{n} \tau^\ell$, $1 \leq n \leq N$, and therefrom $t_f = \sum_{n=1}^{N} \tau^n$.

For the discrete spaces, we let $\mathcal{T}_h, \mathcal{R}_h, \mathcal{P}_h, \mathcal{W}_h$ and $\mathcal{U}_h$ be suitable finite element spaces corresponding to the infinite dimensional spaces of subsection 1.2, where we assume that

$$\operatorname{div} \mathcal{R}_h = \mathcal{T}_h \quad \text{and} \quad \operatorname{div} \mathcal{W}_h = \mathcal{P}_h. \tag{3}$$

For the time discretization we employ a backward Euler scheme. For the sake of simplicity, we take the source terms $\mathbf{f}$, g and z to be piecewise constant in time. We denote by $(T_h^n, \mathbf{r}_h^n, p_h^n, \mathbf{w}_h^n, \mathbf{u}_h^n)$ the discrete counterpart of the solution tuple to problem (2) at time t^n.

DEFINITION 2.1 (The coupled *mixed*×*mixed* and *Galerkin* finite element scheme). *The discrete formulation of the problem* (2) *reads: Given* $\psi(p_h^0, T_h^0, \mathbf{u}_h^0)$ *and* $\varphi(p_h^0, T_h^0, \mathbf{u}_h^0)$*, then, for* $n = 1, \cdots, N$*, find* $(T_h^n, \mathbf{r}_h^n, p_h^n, \mathbf{w}_h^n, \mathbf{u}_h^n) \in \mathcal{T}_h \times \mathcal{R}_h \times \mathcal{P}_h \times \mathcal{W}_h \times \mathcal{U}_h$ *such that*

$$
\begin{aligned}
&(\psi(p_h^n, T_h^n, \mathbf{u}_h^n), S_h) + \tau^n c_f(\mathbf{w}_h^{n,M} \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}_h^{n,M}, S_h) + \tau^n(\nabla \cdot \mathbf{r}_h^n, S_h) \\
&\qquad = \tau^n(z^n, S_h) + (\psi(p_h^{n-1}, T_h^{n-1}, \mathbf{u}_h^{n-1}), S_h), && \forall S_h \in \mathcal{T}_h, \qquad \text{(4a)}\\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}_h^n, \mathbf{y}_h) - (T_h^n, \nabla \cdot \mathbf{y}_h) = 0, && \forall \mathbf{y}_h \in \mathcal{R}_h, \qquad \text{(4b)}\\
&\psi(p_h^n, T_h^n, \mathbf{u}_h^n), q_h) + \tau^n(\nabla \cdot \mathbf{w}_h^n, q_h) \\
&\qquad = \tau^n(g^n, q_h) + (\psi(p_h^{n-1}, T_h^{n-1}, \mathbf{u}_h^{n-1}), q_h), && \forall q_h \in \mathcal{P}_h, \qquad \text{(4c)}\\
&(\mathbf{K}^{-1}\mathbf{w}_h^n, \mathbf{z}_h) - (p_h^n, \nabla \cdot \mathbf{z}_h) = 0, && \forall \mathbf{z}_h \in \mathcal{W}_h, \qquad \text{(4d)}\\
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}_h^n), \boldsymbol{\varepsilon}(\mathbf{v}_h)) + \lambda(\nabla \cdot \mathbf{u}_h^n, \nabla \cdot \mathbf{v}_h) \\
&\qquad - (\beta T_h^n + \alpha p_h^n, \nabla \cdot \mathbf{v}_h) = (\mathbf{f}^n, \mathbf{v}_h), && \forall \mathbf{v}_h \in \mathcal{U}_h. \qquad \text{(4e)}
\end{aligned}
$$

where the functions $(\mathbf{w}_h^{n,M}, \mathbf{r}_h^{n,M})$ *are defined as*

$$
\mathbf{w}_h^{n,M} := \min(|\mathbf{w}_h^n|, M)\frac{\mathbf{w}_h^n}{|\mathbf{w}_h^n|}, \quad \text{and} \quad \mathbf{r}_h^{n,M} := \min(|\mathbf{r}_h^n|, M)\frac{\mathbf{r}_h^n}{|\mathbf{r}_h^n|}, \qquad (5)
$$

where M *is a fixed positive real number and* $|\mathbf{v}| := \sqrt{\sum_{i=1}^d (\mathbf{v})_i^2}$.

In the above scheme, we used $(\mathbf{w}_h^{n,M} \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}_h^{n,M}, S_h)$ for the approximation of the convective coupling term instead of the original $(\mathbf{w}_h^n \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}_h^n, S_h)$. The reason for this approximation will be clarified later. The equations (4a)-(4b) form the discrete mixed scheme of the *heat subproblem*, (4c)-(4d) form the discrete mixed scheme for the *flow subproblem*, and (4e) is the discrete form of the *mechanics subproblem* with the Galerkin finite element method. Together, these subproblems make up the nonlinear and fully coupled discrete version of the *thermo-poroelastic problem* to be solved iteratively in the next section.

REMARK 2.1 (Convective coupling term). *The convective coupling term* $(\mathbf{w}_h^n \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}_h^n, S_h)$ *can also be approximated by* $(\mathbf{w}_h^{n,M} \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}_h^{n,R}, S_h)$*, where two different constants* M *and* R *are used in the definitions* (5)*. In that case, the underlying iterative methods of Section 3 as well as the convergence analysis of Section 4 remains true with minor modifications in the proofs. For simplicity, we let* $M = R$.

3. The *L*-type iterative schemes. We now present six iterative (splitting) algorithms for the discrete thermo-poroelastic problem (4). These algorithms involve either decoupling all the subproblems and solving each separately at every iteration (three-step algorithm), or decoupling only one subproblem from the other two which are then solved together (two-step algorithm), or solving a linearized problem monolithically at every iteration (one-step algorithm). We use the letters **H** (Heat), **F** (Flow), and **M** (Mechanics), to abbreviate the algorithms, *e.g.* a two-step algorithm where the heat and flow subproblems are solved together decoupled from the mechanics subproblem is referred to as (**HF-M**), and similarly for other combinations

of coupling/decoupling of the subproblems. Throughout the rest of the article we will mostly refer to the discrete problems, and therefore omit the h-subscript on the variables and test functions for cleaner notation. We shall also denote the time step simply by τ, keeping in mind it may depend on n.

At the time step $n \geq 1$, let $(T^{n-1}, \mathbf{r}^{n-1}, p^{n-1}, \mathbf{w}^{n-1}, \mathbf{u}^{n-1})$ be given. We then approximate the solution at the actual time step $n \in \{1, \cdots, N\}$, using the sequence $(T^{n,k}, \mathbf{r}^{n,k}, p^{n,k}, \mathbf{w}^{n,k}, \mathbf{u}^{n,k})$ for $k \geq 0$, defined in an iterative fashion, and where the iterate $(T^{n,0}, \mathbf{r}^{n,0}, p^{n,0}, \mathbf{w}^{n,0}, \mathbf{u}^{n,0})$ is an initial guess. All the algorithms involve adding the stabilization terms $L_T(T^{n,k} - T^{n,k-1}, S)$ and $L_p(p^{n,k} - p^{n,k-1}, q)$ to the left hand sides of equations (4a) and (4c), respectively, where $L_T, L_p > 0$ are the stabilization parameters (to be chosen later). Furthemore, to make the notation easier, we introduce the parametrized fluid and heat content functionals: For a given $L_T, L_p > 0$, we define

$$\psi_{L_T}(p, \mathbf{u}, T) := (a_0 + L_T)T - b_0 p + \beta \nabla \cdot \mathbf{u}, \tag{6a}$$

$$\varphi_{L_p}(p, \mathbf{u}, T) := (c_0 + L_p)p - b_0 T + \alpha \nabla \cdot \mathbf{u}. \tag{6b}$$

For the analysis of the coupled mixed formulation (4a)–(4e) and the corresponding iterative approach introduced in this section, we need to introduce the cut-off operator $\mathcal{M}$ as described in *e.g.* [41, 40] as

$$\mathcal{M}(\mathbf{z})(x) := \begin{cases} \mathbf{z}(x), & |\mathbf{z}(x)| \leq M, \\ M\mathbf{z}(x)/|\mathbf{z}(x)|, & |\mathbf{z}(x)| > M, \end{cases} \tag{7}$$

where M is a (large) positive constant. The notation $(\mathbf{w}_h^{n,M}, \mathbf{r}_h^{n,M})$ used in Definition 2.1 is then equivalent to $(\mathcal{M}(\mathbf{w}_h^n), \mathcal{M}(\mathbf{r}_h^n))$. Note that the use of $(\mathbf{w}_h^{n,M}, \mathbf{r}_h^{n,M})$ instead of $(\mathbf{w}_h^n, \mathbf{r}_h^n)$ has little or no practical implications, but is necessary in order to facilitate the convergence analysis; obviously, if the exact fluxes are bounded, i.e., $\mathbf{w}^n, \mathbf{r}^n \in (L^\infty(\Omega))^d$, then if we picked M large enough, we have practically $\mathcal{M}(\mathbf{w}^n)(x) = \mathbf{w}^n(x)$ and $\mathcal{M}(\mathbf{r}^n)(x) = \mathbf{r}^n(x)$. We are now able to present our six iterative algorithms:

3.1. The monolithic scheme (HFM). At the each iteration $k \geq 1$ of the L-type monolothic scheme, we solve the linearized thermo-poroelastic problem: Given $(T^{n,k-1}, p^{n,k-1}, \mathbf{w}^{n,k-1}, \mathbf{u}^{n,k-1})$, find $(T^{n,k}, \mathbf{r}^{n,k}, p^{n,k}, \mathbf{w}^{n,k}, \mathbf{u}^{n,k})$ such that

$$\begin{aligned} &(\psi_{L_T}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k}), S) \\ &\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1}) \cdot \boldsymbol{\Theta}^{-1} \mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla \cdot \mathbf{r}^{n,k}, S) \\ &\quad = \tau(z^n, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S) \\ &\quad + L_T(T^{n,k-1}, S), \qquad \forall S \in \mathcal{T}_h, \end{aligned} \tag{8a}$$

$$(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla \cdot \mathbf{y}) = 0, \qquad \forall \mathbf{y} \in \mathcal{R}_h, \tag{8b}$$

$$\begin{aligned} &(\varphi_{L_p}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k}), q) + \tau(\nabla \cdot \mathbf{w}^{n,k}, q) \\ &\quad = \tau(g^n, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q) \\ &\quad + L_p(p^{n,k-1}, q), \qquad \forall q \in \mathcal{P}_h, \end{aligned} \tag{8c}$$

$$(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla \cdot \mathbf{z}) = 0, \qquad \forall \mathbf{z} \in \mathcal{W}_h, \tag{8d}$$

$$\begin{aligned} &2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v})) \\ &\quad + \lambda(\nabla \cdot \mathbf{u}^{n,k}, \nabla \cdot \mathbf{v}) \\ &\quad = (\mathbf{f}^n, \mathbf{v}) + (\beta T^{n,k} + \alpha p^{n,k}, \nabla \cdot \mathbf{v}), \qquad \forall \mathbf{v} \in \mathcal{U}_h. \end{aligned} \tag{8e}$$

This algorithm in continued until a fixed tolerance is reached. Clearly, in the above algorithm, the L-scheme acts only as a linearization procedure, where we approximate the convective transport term by $\mathcal{M}(\mathbf{w}^{n,k-1})\cdot\boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k})$. Note that, one can also approximate this term by $\mathcal{M}(\mathbf{w}^{n,k})\cdot\boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k-1})$, and the analysis presented next remains true and follows exactly the same lines. The complexity in this algorithm is that it requires solving a large system generated by (8), which combines equations varied in type, and this is at each iteration $k \geq 1$. This encourages the development of efficient techniques for the resolution of these coupled systems.

3.2. The partially decoupled schemes. In the second set of iterative schemes, we only decouple the flow (**F**), mechanics (**M**) or heat (**H**) from the remaining two processes, which are being solved monolithically. Thus, we transform the monolithic solver (**HFM**) into a *two-level* iterative approach in which two simpler subproblems are solved sequentially. We note that for the partially and fully decoupled schemes, we do not consider a cyclical permutation of the order in which the subproblems are solved to yield a different algorithm. In practice, however, such a permutation might yield a slightly different algorithm. The partially decoupled setting delivers the following three iterative approaches:

3.2.1. (HF-M): Coupled heat and flow. Decoupling the mechanics calculation from the coupled flow and heat flow calculation, the first *two-level* iterative scheme reads as follows: At the iteration $k \geq 1$, do:

• **Step 1**: Given $(T^{n,k-1}, p^{n,k-1}, \mathbf{w}^{n,k-1}, \mathbf{u}^{n,k-1})$, find $(T^{n,k}, \mathbf{r}^{n,k}, p^{n,k}, \mathbf{w}^{n,k})$ such that

$$
\begin{aligned}
&(\psi_{L_T}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k-1}), S)\\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1})\cdot\boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla\cdot\mathbf{r}^{n,k}, S)\\
&\qquad = \tau(z^n, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S)\\
&\qquad\quad + L_T(T^{n,k-1}, S), && \forall S \in \mathcal{T}_h, && \text{(9a)}\\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla\cdot\mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h, && \text{(9b)}\\
&(\varphi_{L_p}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k-1}), q) + \tau(\nabla\cdot\mathbf{w}^{n,k}, q)\\
&\qquad = \tau(g^n, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q)\\
&\qquad\quad + L_p(p^{n,k-1}, q), && \forall q \in \mathcal{P}_h, && \text{(9c)}\\
&(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla\cdot\mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h. && \text{(9d)}
\end{aligned}
$$

• **Step 2**: Given $(p^{n,k}, T^{n,k})$, find the displacement $\mathbf{u}^{n,k}$ such that

$$
\begin{aligned}
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v}))\\
&\quad + \lambda(\nabla\cdot\mathbf{u}^{n,k}, \nabla\cdot\mathbf{v})\\
&\qquad = (\mathbf{f}^n, \mathbf{v}) + (\beta T^{n,k} + \alpha p^{n,k}, \nabla\cdot\mathbf{v}), && \forall \mathbf{v} \in \mathcal{U}_h. && \text{(9e)}
\end{aligned}
$$

3.2.2. (HM-F): Coupled heat and mechanics. The second scheme in this subsection is obtained by decoupling the flow calculation from the remaining coupled thermo-elasticity calculation. This iterative scheme reads: At the iteration $k \geq 1$, do:

• **Step 1**: Given $(T^{n,k-1}, p^{n,k-1}, \mathbf{w}^{n,k-1}, \mathbf{u}^{n,k-1})$, find $(T^{n,k}, \mathbf{r}^{n,k}, \mathbf{u}^{n,k})$ such

that

$$
\begin{aligned}
&(\psi_{L_T}(T^{n,k}, p^{n,k-1}, \mathbf{u}^{n,k}), S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1}) \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla \cdot \mathbf{r}^{n,k}, S) \\
&\quad = \tau(z^n, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S) \\
&\quad + L_T(T^{n,k-1}, S), && \forall S \in \mathcal{T}_h, \qquad \text{(10a)} \\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h, \qquad \text{(10b)} \\
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v})) \\
&\quad + \lambda(\nabla \cdot \mathbf{u}^{n,k}, \nabla \cdot \mathbf{v}) \\
&\quad - \beta(T^{n,k}, \nabla \cdot \mathbf{v}) = (\mathbf{f}^n, \mathbf{v}) + \alpha(p^{n,k-1}, \nabla \cdot \mathbf{v}), && \forall \mathbf{v} \in \mathcal{U}_h. \qquad \text{(10c)}
\end{aligned}
$$

- **Step 2**: Given $(T^{n,k}, \mathbf{u}^{n,k}, p^{n,k-1})$, find $(p^{n,k}, \mathbf{w}^{n,k})$ such that

$$
\begin{aligned}
&(c_0 + L_p)(p^{n,k}, q) + \tau(\nabla \cdot \mathbf{w}^{n,k}, q) \\
&\quad = \tau(g^n, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q) \\
&\quad + L_p(p^{n,k-1}, q) + b_0(T^{n,k}, q) - \alpha(\nabla \cdot \mathbf{u}^{n,k}, q), && \forall q \in \mathcal{P}_h, \qquad \text{(10d)} \\
&(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h. \qquad \text{(10e)}
\end{aligned}
$$

3.2.3. (FM-H): Coupled flow and mechanics. The last *two-level* scheme is obtained by decoupling the poro-elasticity (solved monolithically) calculation from the heat flow. Note that a similar scheme was proposed in [19] for two-phase flow. This iterative scheme reads: At the iteration $k \geq 1$, do:

- **Step 1**: Given $(p^{n,k-1}, \mathbf{u}^{n,k-1}, T^{n,k-1})$, find $(p^{n,k}, \mathbf{w}^{n,k}, \mathbf{u}^{n,k})$ such that

$$
\begin{aligned}
&(\varphi_{L_p}(T^{n,k-1}, p^{n,k}, \mathbf{u}^{n,k}), q) + \tau(\nabla \cdot \mathbf{w}^{n,k}, q) \\
&\quad = \tau(g^n, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q) \\
&\quad + L_p(p^{n,k-1}, q), && \forall q \in \mathcal{P}_h, \qquad \text{(11a)} \\
&(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h, \qquad \text{(11b)} \\
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v})) \\
&\quad + \lambda(\nabla \cdot \mathbf{u}^{n,k}, \nabla \cdot \mathbf{v}) - \alpha(p^{n,k}, \nabla \cdot \mathbf{v}) \\
&\quad = (\mathbf{f}^n, \mathbf{v}) + \beta(T^{n,k-1}, \nabla \cdot \mathbf{v}), && \forall \mathbf{v} \in \mathcal{U}_h. \qquad \text{(11c)}
\end{aligned}
$$

- **Step 2**: Given $(p^{n,k}, \mathbf{w}^{n,k}, \mathbf{u}^{n,k}, T^{n,k-1})$, find $(T^{n,k}, \mathbf{r}^{n,k})$ such that

$$
\begin{aligned}
&(a_0 + L_T)(T^{n,k}, S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k}) \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla \cdot \mathbf{r}^{n,k}, S) \\
&\quad = \tau(z^n, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S) \\
&\quad + L_T(T^{n,k-1}, S) + b_0(p^{n,k}, S) - \beta(\nabla \cdot \mathbf{u}^{n,k}, S), \forall S \in \mathcal{T}_h, \qquad \text{(11d)} \\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h. \qquad \text{(11e)}
\end{aligned}
$$

3.3. The fully decoupled schemes. In this set of iterative coupling schemes, we simply split the three processes, providing three sub-problems to be solved sequentially. Fixing the mechanics calculation in the third level, two approaches are then derived in which either the problem of flow or the heat is solved first followed

by solving the other system and then the mechanics using the already calculated information, leading to recover the original solution. These schemes enjoy the solving of much simpler subsystems through the algorithm, as well as the facility to reuse existing codes for each component of the problem.

3.3.1. (H-F-M): Decoupled heat - flow - mechanics. At each iteration all three subproblems are decoupled, and are solved in the order heat $\rightarrow$ flow $\rightarrow$ mechanics. This iterative scheme reads: At the iteration $k \geq 1$, do:

- **Step 1**: Given $(p^{n,k-1}, \mathbf{w}^{n,k-1}, T^{n,k-1}, \mathbf{u}^{n,k-1})$ find $(T^{n,k}, \mathbf{r}^{n,k})$ such that

$$
\begin{aligned}
&(\psi_{L_T}(T^{n,k}, p^{n,k-1}, \mathbf{u}^{n,k-1}), S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k}) \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla \cdot \mathbf{r}^{n,k}, S) \\
&\qquad = \tau(z^n, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S) \\
&\qquad\quad + L_T(T^{n,k-1}, S), && \forall S \in \mathcal{T}_h, \qquad (12a) \\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h. \qquad (12b)
\end{aligned}
$$

- **Step 2**: Given $(p^{n,k-1}, T^{n,k}, \mathbf{u}^{n,k-1})$ find $(p^{n,k}, \mathbf{w}^{n,k})$ such that

$$
\begin{aligned}
&(\varphi_{L_p}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k}), q) + \tau(\nabla \cdot \mathbf{w}^{n,k}, q) \\
&\qquad = \tau(g, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q) \\
&\qquad\quad + L_p(p^{n,k-1}, q) + b_0(T^{n,k}, q) - \alpha(\nabla \cdot \mathbf{u}^{n,k-1}, q), && \forall q \in \mathcal{P}_h, \qquad (12c) \\
&(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h. \qquad (12d)
\end{aligned}
$$

- **Step 3**: Given $(p^{n,k}, T^{n,k})$ find $\mathbf{u}^{n,k}$ such that

$$
\begin{aligned}
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{u}^{n,k}, \nabla \cdot \mathbf{v}) \\
&\qquad = (\mathbf{f}, \mathbf{v}) + (\beta T^{n,k} + \alpha p^{n,k}, \nabla \cdot \mathbf{v}), && \forall \mathbf{v} \in \mathcal{U}_h. \qquad (12e)
\end{aligned}
$$

3.3.2. (F-H-M): Decoupled flow - heat - mechanics. At each iteration all three subproblems are decoupled, and are solved in the order flow $\rightarrow$ heat $\rightarrow$ mechanics. This iterative scheme reads: At iteration $k \geq 1$, do:

- **Step 1**: Given $(p^{n,k-1}, T^{n,k-1}, \mathbf{u}^{n,k-1})$ find $(p^{n,k}, \mathbf{w}^{n,k})$ such that

$$
\begin{aligned}
&(\varphi_{L_p}(T^{n,k-1}, p^{n,k}, \mathbf{u}^{n,k-1}), q) + \tau(\nabla \cdot \mathbf{w}^{n,k}, q) \\
&\qquad = \tau(g, q) + (\varphi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), q) \\
&\qquad\quad + L_p(p^{n,k-1}, q), && \forall q \in \mathcal{P}_h, \qquad (13a) \\
&(\mathbf{K}^{-1}\mathbf{w}^{n,k}, \mathbf{z}) - (p^{n,k}, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h. \qquad (13b)
\end{aligned}
$$

- **Step 2**: Given $(p^{n,k}, \mathbf{w}^{n,k}, T^{n,k-1}, \mathbf{u}^{n,k-1})$, find $(T^{n,k}, \mathbf{r}^{n,k})$ such that

$$
\begin{aligned}
&(\psi_{L_T}(T^{n,k}, p^{n,k}, \mathbf{u}^{n,k-1}), S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k}) \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) + \tau(\nabla \cdot \mathbf{r}^{n,k}, S) \\
&\qquad = \tau(h, S) + (\psi(T^{n-1}, p^{n-1}, \mathbf{u}^{n-1}), S) \\
&\qquad\quad + L_T(T^{n,k-1}, S) + b_0(p^{n,k}, S) - \beta(\nabla \mathbf{u}^{n,k-1}, S), && \forall S \in \mathcal{T}_h, \qquad (13c) \\
&(\boldsymbol{\Theta}^{-1}\mathbf{r}^{n,k}, \mathbf{y}) - (T^{n,k}, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h. \qquad (13d)
\end{aligned}
$$

- **Step 3**: Given $(p^{n,k}, T^{n,k})$, find $\mathbf{u}^{n,k}$ such that

$$
\begin{aligned}
&2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^{n,k}), \boldsymbol{\varepsilon}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{u}^{n,k}, \nabla \cdot \mathbf{v}) \\
&\qquad = (\mathbf{f}^n, \mathbf{v}) + (\beta T^{n,k} + \alpha p^{n,k}, \nabla \cdot \mathbf{v}), && \forall \mathbf{v} \in \mathcal{U}_h. \qquad (13e)
\end{aligned}
$$

4. Convergence analysis. The starting point for our analysis is the existence and uniqueness of a solution to (4). To this aim, we will make use of the following Lemma (cf. [41]), stating the Lipschitz property of the cut-off operator $\mathcal{M}$:

LEMMA 4.1 (Property of $\mathcal{M}$). *The "cut-off" operator $\mathcal{M}$ defined as in equation (7) is uniformly Lipschitz continuous, i.e.*

$$\|\mathcal{M}(\mathbf{z}_1) - \mathcal{M}(\mathbf{z}_2)\|_{(L^\infty(\Omega))^d} \leq \|\mathbf{z}_1 - \mathbf{z}_2\|_{(L^\infty(\Omega))^d}. \tag{14}$$

Thus, we have

$$\left\|\mathcal{M}(\mathbf{w}^n) - \mathcal{M}(\mathbf{w}^{n,k})\right\|_{(L^\infty(\Omega))^d} \leq \left\|\mathbf{w}^n - \mathbf{w}^{n,k}\right\|_{(L^\infty(\Omega))^d}, \tag{15a}$$

and

$$\|\mathcal{M}(\mathbf{w}^n)\|_{(L^\infty(\Omega))^d} \leq M. \tag{15b}$$

The proof of the next Theorem is based on showing that the scheme (8) is a contraction, and then by applying the Banach fixed-point theorem [15], to deduce convergence of the scheme. In what follows we will frequently use the following polarization and binomial identities,

$$4(u,v) = \|u+v\|^2 - \|u-v\|^2, \quad \text{and} \quad 2(u-v,u) = \|u\|^2 + \|u-v\|^2 - \|v\|^2. \tag{16}$$

Finally, we define the difference functions between the solutions at the iteration k and $k-1$ of problem (8), respectively as

$$\begin{aligned}(e_T^k, \mathbf{e}_\mathbf{r}^k, e_p^k, \mathbf{e}_\mathbf{w}^k, \mathbf{e}_\mathbf{u}^k)\\ := (T^{n,k} - T^{n,k-1}, \mathbf{r}^{n,k} - \mathbf{r}^{n,k-1},\\ p^{n,k} - p^{n,k-1}, \mathbf{w}^{n,k} - \mathbf{w}^{n,k-1}, \mathbf{u}^{n,k} - \mathbf{u}^{n,k-1}).\end{aligned} \tag{17}$$

With this, we state the first of our main results:

THEOREM 4.2 (Convergence of the monolithic L-scheme **HFM**). *Assuming* (**A1**)–(**A6**) *holds true, and the time step is small enough, i.e.*

$$\tau < \frac{2(a_0 - b_0)}{c_f^2 M^2 \left(\frac{k_M}{\theta_m} + 1\right) - \frac{\theta_m}{4c_{\Omega,d}}}, \tag{18}$$

then, the monolithic L-scheme **HFM** *(Algorithm 3.1) defines a contraction satisfying*

$$\begin{aligned}&\left(a_0 - b_0 + \frac{L_T}{2} + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m} + 1\right)\right)\left\|e_T^k\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_\mathbf{r}^k\right\|_{\mathbf{\Theta}^{-1}}^2\\ &\quad + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 + \tau\left\|\mathbf{e}_\mathbf{w}^k\right\|_{\mathbf{K}^{-1}}^2\\ &\quad + 2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k)\right\|^2 + \lambda\left\|\nabla\cdot\mathbf{e}_\mathbf{u}^k\right\|^2\\ &\quad \leq \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_\mathbf{w}^{k-1}\right\|_{\mathbf{K}^{-1}}^2.\end{aligned} \tag{19}$$

Therefrom, the limit is the unique solution of the problem (4).

REMARK 4.1 (Bound on time step). *Note that $a_0 - b_0 > 0$ due to the Assumption (A4), and*

$$c_f^2 M^2 \left(\frac{k_M}{\theta_m} + 1\right) - \frac{\theta_m}{4c_{\Omega,d}} > 0, \tag{20}$$

by the choice of M sufficiently large (thus, the right hand side of (18) is a positive number). Note also that if a priori bounds on the fluxes are available, and these are small enough such that M can be chosen to yield equality in (20), then there would be no constraint on the time step.

Proof. We begin by deriving the error equations satisfied by $(e_T^k, \mathbf{e}_\mathbf{r}^k, e_p^k, \mathbf{e}_\mathbf{w}^k, \mathbf{e}_\mathbf{u}^k)$, i.e. subtract the equations (8) for k from the ones for $k-1$, and obtain

$$\begin{aligned}
&(\psi_{L_T}(e_T^k, e_p^k, \mathbf{e}_\mathbf{u}^k), S) + \tau(\nabla \cdot \mathbf{e}_\mathbf{r}^n, S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1}) \cdot \boldsymbol{\Theta}^{-1}[\mathcal{M}(\mathbf{r}^{n,k}) - \mathcal{M}(\mathbf{r}^{n,k-1})], S) \\
&\quad + \tau c_f([\mathcal{M}(\mathbf{w}^{n,k-1}) - \mathcal{M}(\mathbf{w}^{n,i-2})] \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n,k}), S) \\
&\quad = L_T(e_T^{k-1}, S), && \forall S \in \mathcal{T}_h, \qquad (21a)\\
&(\boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r}^k, \mathbf{y}) - (e_T^k, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h, \qquad (21b)\\
&(\varphi_{L_p}(e_T^k, e_p^k, \mathbf{e}_\mathbf{u}^k), q) + \tau(\nabla \cdot \mathbf{e}_\mathbf{w}^k, q) = L_p(e_p^{k-1}, q), && \forall q \in \mathcal{P}_h, \qquad (21c)\\
&(\mathbf{K}^{-1}\mathbf{e}_\mathbf{w}^k, \mathbf{z}) - (e_p^k, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h, \qquad (21d)\\
&2\mu(\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k), \boldsymbol{\varepsilon}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{e}_\mathbf{u}^k, \nabla \cdot \mathbf{v}) \\
&\quad - (\beta e_T^k + \alpha e_p^k, \nabla \cdot \mathbf{v}) = 0, && \forall \mathbf{v} \in \mathcal{U}_h. \qquad (21e)
\end{aligned}$$

We choose now $S = e_T^k$, $\mathbf{y} = \tau \mathbf{e}_\mathbf{r}^k$, $q = e_p^k$, $\mathbf{z} = \tau \mathbf{e}_\mathbf{w}^k$, and $\mathbf{v} = \mathbf{e}_\mathbf{u}^k$ as test functions in equations (21a)– (21e), respectively. Then, summing the resulting equations and using the identity (16) together with applying Cauchy-Schwarz and Young inequalities and some algebraic manipulations, we get, for any $\epsilon_1, \epsilon_2 > 0$,

$$\begin{aligned}
&\left(a_0 - b_0 + \frac{L_T}{2}\right)\left\|e_T^k\right\|^2 + \tau\left\|\mathbf{e}_\mathbf{r}^k\right\|_{\boldsymbol{\Theta}^{-1}}^2 + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 \\
&\quad + \tau\left\|\mathbf{e}_\mathbf{w}^k\right\|_{\mathbf{K}^{-1}}^2 + 2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k)\right\|^2 + \lambda\left\|\nabla \cdot \mathbf{e}_\mathbf{u}^k\right\|^2 \\
&\quad \le \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 \\
&\qquad + \tau c_f\left\|\mathcal{M}(\mathbf{w}^{k-1}) \cdot \boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r}^k\right\|\left\|e_T^k\right\| + \tau c_f\left\|\mathbf{e}_\mathbf{w}^{k-1} \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{k-1})\right\|\left\|e_T^k\right\|, \\
&\quad \le \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \tau c_f M\left(\frac{\epsilon_1}{2} + \frac{\epsilon_2}{2}\right)\left\|e_T^k\right\|^2 \\
&\qquad + \tau c_f M \frac{1}{2\epsilon_1}\left\|\mathbf{e}_\mathbf{r}^k\right\|_{\boldsymbol{\Theta}^{-1}}^2 + \tau c_f M \frac{k_M}{\theta_m}\frac{1}{2\epsilon_2}\left\|\mathbf{e}_\mathbf{w}^{k-1}\right\|_{\mathbf{K}^{-1}}^2. \qquad (22)
\end{aligned}$$

From equation (21b), and by Thomas' lemma [44], there exists $\hat{\mathbf{y}} \in \mathcal{R}_h$ and a constant $c_{\Omega,d} > 0$ depending only on the domain and spatial dimension such that $\nabla \cdot \hat{\mathbf{y}} = e_T^k$ with $\|\hat{\mathbf{y}}\| \le c_{\Omega,d}\left\|e_T^k\right\|$. Thus, taking $\hat{\mathbf{y}}$ as a test function in (30d) we deduce

$$\begin{aligned}
\left\|e_T^k\right\|^2 = (e_T^k, \nabla \cdot \hat{\mathbf{y}}) &= (\boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r}^k, \hat{\mathbf{y}}) \\
&\le \left\|\mathbf{e}_\mathbf{r}^k\right\|_{\boldsymbol{\Theta}^{-1}} \cdot \frac{1}{\sqrt{\theta_m}}\|\hat{\mathbf{y}}\| \\
&\le \left\|\mathbf{e}_\mathbf{r}^k\right\|_{\boldsymbol{\Theta}^{-1}} \cdot \frac{c_{\Omega,d}}{\sqrt{\theta_m}}\left\|e_T^k\right\|, \qquad (23)
\end{aligned}$$

which leads to

$$\frac{\theta_m}{c_{\Omega,d}}\left\|e_T^k\right\|^2 \leq \left\|\mathbf{e}_{\mathbf{r}}^k\right\|^2_{\boldsymbol{\Theta}^{-1}}. \tag{24}$$

Replacing (24) in (22) while choosing $\epsilon_1 = c_f M$ and $\epsilon_2 = c_f M k_M/\theta_m$, we obtain

$$\begin{aligned}\left(a_0 - b_0 + \frac{L_T}{2} + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m}+1\right)\right)\left\|e_T^k\right\|^2 + \frac{\tau}{4}\left\|\mathbf{e}_{\mathbf{r}}^k\right\|^2_{\boldsymbol{\Theta}^{-1}}\\ + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{w}}^k\right\|^2_{\mathbf{K}^{-1}}\\ + 2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^k)\right\|^2 + \lambda\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\|^2\\ \leq \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_{\mathbf{w}}^{k-1}\right\|^2_{\mathbf{K}^{-1}}.\end{aligned} \tag{25}$$

The contraction of the residuals follows if the time step τ satisfies (18). This proves the convergence of the monolithic L-scheme. The limit is then the unique solution of (4). ∎

The well-posedness of the discrete variational problem (4) is established by the Theorem 4.2, where the solution at time t^n, $n \leq 0$, is denoted by $(T^n, \mathbf{r}^n, p^n, \mathbf{w}^n, \mathbf{u}^n)$. Thus, we can now prove the convergence of the decoupled schemes to this solution. We begin with analyzing the partially decoupled schemes, introduced in Subsection 3.2. To this end, we let the difference functions defined in (17) now be the differences between the solutions at the iteration k of problem (9), and the solutions to (4), i.e.

$$(e_T^k, \mathbf{e}_{\mathbf{r}}^k, e_p^k, \mathbf{e}_{\mathbf{w}}^k, \mathbf{e}_{\mathbf{u}}^k) := (T^{n,k} - T^n, \mathbf{r}^{n,k} - \mathbf{r}^n, p^{n,k} - p^n, \mathbf{w}^{n,k} - \mathbf{w}^n, \mathbf{u}^{n,k} - \mathbf{u}^n). \tag{26}$$

The second of our main results is given through

THEOREM 4.3 (Convergence of the partially decoupled schemes). *Assuming* (**A1**)–(**A6**) *holds true, the stabilization parameters are such that*

$$L_p \geq \frac{4\alpha^2}{3(\frac{2\mu}{d}+\lambda)} \quad \text{and} \quad L_T \geq \frac{4\beta^2}{3(\frac{2\mu}{d}+\lambda)}, \tag{27}$$

and the time step satisfies (18)*, then the partially decoupled L-scheme* **HF-M** *(Algorithm 3.2.1) is a contraction given by*

$$\begin{aligned}\left(a_0 - b_0 + \frac{L_T}{2} + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m}+1\right)\right)\left\|e_T^k\right\|^2\\ + \frac{\tau}{4}\left\|\mathbf{e}_{\mathbf{r}}^k\right\|^2_{\boldsymbol{\Theta}^{-1}} + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{w}}^k\right\|^2_{\mathbf{K}^{-1}}\\ \leq \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_{\mathbf{w}}^{k-1}\right\|^2_{\mathbf{K}^{-1}}.\end{aligned} \tag{28}$$

Furthemore, there holds,

$$\frac{\mu}{2}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^k)\right\|^2 + \frac{\lambda}{4}\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\|^2 \leq \frac{2\alpha^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_p^k\right\|^2 + \frac{2\beta^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_T^k\right\|^2. \tag{29}$$

Proof. We start by taking the difference of equations (9a) – (9e) at iteration k with the corresponding equations solved by $(T^n, \mathbf{r}^n, p^n, \mathbf{w}^n, \mathbf{u}^n)$. This leads to the following set of difference equations

$$
\begin{aligned}
&(\psi_{L_T}(e_T^k, e_p^k, \mathbf{e}_\mathbf{u}^{k-1}), S) + \tau(\nabla \cdot \mathbf{e}_\mathbf{r}^k, S) \\
&\quad + \tau c_f([\mathcal{M}(\mathbf{w}^{n,k-1}) - \mathcal{M}(\mathbf{w}^n)] \cdot \boldsymbol{\Theta}^{-1}\mathbf{r}^n, S) \\
&\quad + \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1}) \cdot \boldsymbol{\Theta}^{-1}[\mathcal{M}(\mathbf{r}^{n,k}) - \mathcal{M}(\mathbf{r}^n)], S) \\
&\quad = L_T(e_T^{k-1}, S), && \forall S \in \mathcal{T}_h, \qquad (30a)\\
&(\boldsymbol{\Theta}^{-1}\mathbf{e}_\mathbf{r}^k, \mathbf{y}) - (e_T^k, \nabla \cdot \mathbf{y}) = 0, && \forall \mathbf{y} \in \mathcal{R}_h \qquad (30b)\\
&(\varphi_{L_p}(e_T^k, e_p^k, \mathbf{e}_\mathbf{u}^{k-1}), q) + \tau(\nabla \cdot \mathbf{e}_\mathbf{w}^k, q) = L_p(e_p^{k-1}, q), && \forall q \in \mathcal{P}_h, \qquad (30c)\\
&(\mathbf{K}^{-1}\mathbf{e}_\mathbf{w}^k, \mathbf{z}) - (e_p^k, \nabla \cdot \mathbf{z}) = 0, && \forall \mathbf{z} \in \mathcal{W}_h, \qquad (30d)\\
&2\mu(\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k), \boldsymbol{\varepsilon}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{e}_\mathbf{u}^k, \nabla \cdot \mathbf{v}) \\
&\quad - (\alpha e_p^k + \beta e_T^k, \nabla \cdot \mathbf{v}) = 0, && \forall \mathbf{v} \in \mathcal{U}_h. \qquad (30e)
\end{aligned}
$$

The aim now is to show a contraction of successive error functions, thereby implying convergence of the sequences $(T^{n,k}, \mathbf{r}^{n,k}, p^{n,k}, \mathbf{w}^{n,k}, \mathbf{u}^{n,k})$ as $k \to \infty$ for $n \geq 1$, by the Banach Fixed Point Theorem [15]. Taking as test functions $q = e_p^k, \mathbf{z} = \tau\mathbf{e}_\mathbf{w}^k, S = e_T^k, \mathbf{y} = \tau\mathbf{e}_\mathbf{r}^k$, and $\mathbf{v} = \mathbf{e}_\mathbf{u}^{k-1}$ in (30a) – (30e), respectively, and adding the resulting equations together, we obtain

$$
\begin{aligned}
&\left(a_0 + \frac{L_T}{2}\right)\left\|e_T^k\right\|^2 + \frac{L_T}{2}\left\|e_T^k - e_T^{k-1}\right\|^2 + \tau\left\|\mathbf{e}_\mathbf{r}^k\right\|_{\boldsymbol{\Theta}^{-1}}^2 \\
&\quad + \left(c_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 + \frac{L_p}{2}\left\|e_p^k - e_p^{k-1}\right\|^2 + \tau\left\|\mathbf{e}_\mathbf{w}^k\right\|_{\mathbf{K}^{-1}}^2 \\
&\quad + 2\mu\frac{1}{4}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k + \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 + \lambda\frac{1}{4}\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k + \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 \\
&\quad = \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + 2b_0(e_T^k, e_p^k) \\
&\quad\quad + 2\mu\frac{1}{4}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 + \lambda\frac{1}{4}\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 \\
&\quad\quad - \tau c_f([\mathcal{M}(\mathbf{w}^{n,k-1}) - \mathcal{M}(\mathbf{w}^n)] \cdot \boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^n), e_T^k) \\
&\quad\quad - \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1}) \cdot \boldsymbol{\Theta}^{-1}[\mathcal{M}(\mathbf{r}^{n,k}) - \mathcal{M}(\mathbf{r}^n)], e_T^k), \qquad (31)
\end{aligned}
$$

where we used the identities (16). On the other hand, by taking the difference of eq. (30e) at iteration k and $k-1$, testing with $\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1}$, and using the Cauchy-Schwarz inequality we get

$$
\begin{aligned}
&2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 + \lambda\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 \\
&\quad = \alpha(e_p^k - e_p^{k-1}, \nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})) + \beta(e_T^k - e_T^{k-1}, \nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})) \\
&\quad \leq \left(\alpha\left\|e_p^k - e_p^{k-1}\right\| + \beta\left\|e_T^k - e_T^{k-1}\right\|\right)\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|. \qquad (32)
\end{aligned}
$$

Let now $\xi \in (0,1)$ and rewrite the above estimate as

$$
\begin{aligned}
&2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 + \lambda\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|^2 \\
&\quad \leq \left(\alpha\left\|e_p^k - e_p^{k-1}\right\| + \beta\left\|e_T^k - e_T^{k-1}\right\|\right)\left(\xi\sqrt{d}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|\right. \\
&\quad\quad \left. + (1-\xi)\left\|\nabla \cdot (\mathbf{e}_\mathbf{u}^k - \mathbf{e}_\mathbf{u}^{k-1})\right\|\right). \qquad (33)
\end{aligned}
$$

We now follow [8] and choose $\xi = \dfrac{2\mu}{2\mu + d\lambda}$, which together with the Young inequality yields

$$\begin{aligned}
&\frac{\mu}{2}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^{k} - \mathbf{e}_{\mathbf{u}}^{k-1})\right\|^2 + \frac{\lambda}{4}\left\|\nabla\cdot(\mathbf{e}_{\mathbf{u}}^{k} - \mathbf{e}_{\mathbf{u}}^{k-1})\right\|^2 \\
&\qquad \leq \frac{2\alpha^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_p^{k} - e_p^{k-1}\right\|^2 + \frac{2\beta^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_T^{k} - e_T^{k-1}\right\|^2. \qquad (34)
\end{aligned}$$

Combining this with eq. (31) leads to

$$\begin{aligned}
&(a_0 + \frac{L_T}{2})\left\|e_T^{k}\right\|^2 + \left(\frac{L_T}{2} - \frac{2\beta^2}{3(\frac{2\mu}{d}+\lambda)}\right)\left\|e_T^{k} - e_T^{k-1}\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{r}}^{k}\right\|^2_{\boldsymbol{\Theta}^{-1}} \\
&\quad + (c_0 + \frac{L_p}{2})\left\|e_p^{k}\right\|^2 + \left(\frac{L_p}{2} - \frac{2\alpha^2}{3(\frac{2\mu}{d}+\lambda)}\right)\left\|e_p^{k} - e_p^{k-1}\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{w}}^{k}\right\|^2_{\mathbf{K}^{-1}} \\
&\quad + \frac{\mu}{2}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^{k} + \mathbf{e}_{\mathbf{u}}^{k-1})\right\|^2 + \frac{\lambda}{4}\left\|\nabla\cdot(\mathbf{e}_{\mathbf{u}}^{k} + \mathbf{e}_{\mathbf{u}}^{k-1})\right\|^2 \\
&\quad \leq \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + 2b_0(e_T^{k}, e_p^{k}) \\
&\qquad - \tau c_f([\mathcal{M}(\mathbf{w}^{n,k-1}) - \mathcal{M}(\mathbf{w}^{n})]\cdot\boldsymbol{\Theta}^{-1}\mathcal{M}(\mathbf{r}^{n}), e_T^{k}) \\
&\qquad - \tau c_f(\mathcal{M}(\mathbf{w}^{n,k-1})\cdot\boldsymbol{\Theta}^{-1}[\mathcal{M}(\mathbf{r}^{n,k}) - \mathcal{M}(\mathbf{r}^{n})], e_T^{k}). \qquad (35)
\end{aligned}$$

We thus need to impose some constraints on the stabilization parameters, i.e. $L_p \geq \dfrac{4\alpha^2}{3(\frac{2\mu}{d}+\lambda)}$ and $L_T \geq \dfrac{4\beta^2}{3(\frac{2\mu}{d}+\lambda)}$. With this, we can discard some positive terms on the left hand side of (35), and use the Cauchy-Schwarz and Young inequalities, together with the Lipschitz property of $\mathcal{M}$ to obtain

$$\begin{aligned}
&\left(a_0 - b_0 + \frac{L_T}{2} - \tau c_f M(\frac{\epsilon_1}{2} + \frac{\epsilon_2}{2})\right)\left\|e_T^{k}\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{r}}^{k}\right\|^2_{\boldsymbol{\Theta}^{-1}} \\
&\quad + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^{k}\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{w}}^{k}\right\|^2_{\mathbf{K}^{-1}} \\
&\quad \leq \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 \\
&\qquad + \tau c_f M\frac{k_M}{\theta_m}\frac{1}{2\epsilon_1}\left\|\mathbf{e}_{\mathbf{w}}^{k-1}\right\|^2_{\mathbf{K}^{-1}} + \tau c_f M\frac{1}{2\epsilon_2}\left\|\mathbf{e}_{\mathbf{r}}^{k}\right\|^2_{\boldsymbol{\Theta}^{-1}}, \qquad (36)
\end{aligned}$$

for some $\epsilon_1, \epsilon_2 > 0$, and where k_M and θ_m are given by (**A1**) – (**A2**). From (30d), we obtain in the same way as in (24)

$$\frac{\theta_m}{c_{\Omega,d}}\left\|e_T^{k}\right\|^2 \leq \left\|\mathbf{e}_{\mathbf{r}}^{k}\right\|^2_{\boldsymbol{\Theta}^{-1}}. \qquad (37)$$

Replacing (37) in (36) while choosing $\epsilon_1 = c_f M k_M/\theta_m$ and $\epsilon_2 = c_f M$, we get

$$\begin{aligned}
&\left(a_0 - b_0 + \frac{L_T}{2} + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m} + 1\right)\right)\left\|e_T^{k}\right\|^2 + \frac{\tau}{4}\left\|\mathbf{e}_{\mathbf{r}}^{k}\right\|^2_{\boldsymbol{\Theta}^{-1}} \\
&\quad + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^{k}\right\|^2 + \tau\left\|\mathbf{e}_{\mathbf{w}}^{k}\right\|^2_{\mathbf{K}^{-1}} \\
&\quad \leq \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2 + \frac{L_T}{2}\left\|e_T^{k-1}\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_{\mathbf{w}}^{k-1}\right\|^2_{\mathbf{K}^{-1}}. \qquad (38)
\end{aligned}$$

Thus, if the time step τ satisfies (18), we can write (38) as

$$F^k \leq \frac{1}{1+\delta} F^{k-1}, \tag{39}$$

where

$$F^k := \frac{L_p}{2}\left\|e_p^k\right\|^2 + \frac{L_T}{2}\left\|e_T^k\right\| + \frac{\tau}{4}\left\|\mathbf{e}_{\mathbf{w}}^k\right\|_{\mathbf{K}^{-1}}^2, \tag{40}$$

and

$$\delta := \min\left\{\frac{2}{L_p}(c_0 - b_0), \frac{2}{L_T}\left(a_0 - b_0 + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m}+1\right)\right), \frac{1}{2}\right\} > 0. \tag{41}$$

Going back to eq. (30e), we choose $\mathbf{v} = \mathbf{e}_{\mathbf{u}}^k$ as test function which leads to

$$\begin{aligned} 2\mu\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^k)\right\|^2 + \lambda\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\|^2 &= \alpha(e_p^k, \nabla\cdot\mathbf{e}_{\mathbf{u}}^k) + \beta(e_T^k, \nabla\cdot\mathbf{e}_{\mathbf{u}}^k) \\ &\leq \left(\alpha\left\|e_p^k\right\| + \beta\left\|e_T^k\right\|\right)\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\| \\ &\leq \left(\alpha\left\|e_p^k\right\| + \beta\left\|e_T^k\right\|\right)\left(\xi\sqrt{d}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^k)\right\| + (1-\xi)\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\|\right), \end{aligned} \tag{42}$$

for some $\xi \in (0,1)$. Following the same steps which led to (34), and choosing as before $\xi = \dfrac{2\mu}{2\mu + d\lambda}$, we get by the Young inequality

$$\frac{\mu}{2}\left\|\boldsymbol{\varepsilon}(\mathbf{e}_{\mathbf{u}}^k)\right\|^2 + \frac{\lambda}{4}\left\|\nabla\cdot\mathbf{e}_{\mathbf{u}}^k\right\|^2 \leq \frac{2\alpha^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_p^k\right\|^2 + \frac{2\beta^2}{3(\frac{2\mu}{d}+\lambda)}\left\|e_T^k\right\|^2. \tag{43}$$

This shows a contraction of the residuals and therefore completes the proof. ∎

REMARK 4.2 (The other partially decoupled schemes). *For the partially decoupled schemes* **HM-F** *and* **FM-H** *(Algorithms 3.2.2 and 3.2.3 respectively) the contractions can obtained similarly to the scheme* **HF-M** *with minor changes in the coefficients.*

Before we state the last of our main results, we let the difference functions defined in (26) now be the difference between the solutions at the iteration k of problem (13) and the solutions to (4). The last of our main results then reads:

COROLLARY 4.4 (Convergence of the fully decoupled algorithms). *Under the assumptions of Theorem 4.3, the fully decoupled L-scheme* **F-H-M** *(Algorithm 3.3.2) defines a contraction*

$$\begin{aligned} &\left(a_0 - \frac{b_0}{2} + \frac{L_T}{2} + \frac{\tau\theta_m}{4c_{\Omega,d}} - \frac{\tau c_f^2 M^2}{2}\left(\frac{k_M}{\theta_m}+1\right)\right)\left\|e_T^k\right\|^2 \\ &\quad + \left(c_0 - b_0 + \frac{L_p}{2}\right)\left\|e_p^k\right\|^2 + \frac{\tau}{2}\left\|\mathbf{e}_{\mathbf{w}}^k\right\|_{\mathbf{K}^{-1}}^2 + \frac{\tau}{4}\left\|\mathbf{e}_{\mathbf{r}}^k\right\|_{\boldsymbol{\Theta}^{-1}}^2 \\ &\quad \leq \left(\frac{L_T}{2} + \frac{b_0}{2}\right)\left\|e_T^{k-1}\right\|^2 + \frac{L_p}{2}\left\|e_p^{k-1}\right\|^2. \end{aligned} \tag{44}$$

Furthermore, the estimate (29) holds true.

Proof. We follow the same lines as in the proof of Theorem 4.3, and take the difference of equations (13a) – (13d) with the same equations solved by $(T^n, \mathbf{r}^n, p^n, \mathbf{w}^n, \mathbf{u}^n)$, and obtain the difference equations for the fully decoupled scheme **F-H-M**. We then promptly obtain estimate (44), from which the contraction is inferred by choosing the stabilization parameters and the time step. That of the second estimate follows in exactly the same way. ∎

REMARK 4.3 (The fully decoupled scheme **H-F-M**). *The contraction 44 holds true for Algorithm 3.3.1 by exchanging in there the coefficients in the right-hand side, i.e.,* $\frac{L_p}{2}$ *becomes* $\frac{L_p}{2} + \frac{b_0}{2}$ *and* $\frac{L_T}{2} + \frac{b_0}{2}$ *becomes* $\frac{L_p}{2}$.

REMARK 4.4 (Other schemes). *The results os Section 4 are valid also for other choices of temporal discretizations, as well as different (i.e., non-mixed) formulations for the heat and flow problems. Different spatial discretizations can even be chosen for each of the three subproblems.*

5. Numerical experiments. In the following we present three numerical test cases using the algorithms from Section 3. The first is a constructed problem, posed on the unit square domain, with prescribed solutions for the temperature, pressure and displacements. Here, we consider five different parameter regimes, exhausting all possibilities of weak/strong coupling between the subproblems, and compare the number of iterations needed for convergence with decreasing mesh sizes for both stabilized and non-stabilized algorithms. Since analytical solutions are available, we present also discretization errors.

Next, we present two implementations of Mandel's problem, which is originally a benchmark problem in linear poroelasticity, extended here to nonlinear thermo-poroelasticity. For the original Mandel problem, analytical solutions for the pressure and displacement field are known. Due to the similarity of the thermo-poroelastic equations we consider with the linear Biot's equations, and due to the lack of benchmark problems for thermo-poroelasticity, we choose to use this problem for our second and third numerical test cases. Even though the analytical solutions are no longer valid when including temperature, we have sufficiently weak temperature effects in the first implementation of Mandel's problem that the computed pressure and displacement field matches the (isothermal) analytical solutions. The second implementation of Mandel's problem includes a heat source, which has a significant effect on both the pressure and displacement. Regarding the spatial discretization, we choose the following finite element spaces:

$$\mathcal{R}_h, \mathcal{W}_h := \{\psi \in H(\mathrm{div};\Omega) : \forall K \in \mathcal{X}_h,\ \psi|_K \in \mathbb{RT}_0(K)\}, \tag{45a}$$

$$\mathcal{T}_h, \mathcal{P}_h := \{\varphi \in L^2(\Omega) : \forall K \in \mathcal{X}_h,\ \varphi|_K \in \mathbb{P}_0(K)\}, \tag{45b}$$

$$\mathcal{U}_h := \{\eta \in (H^1(\Omega))^d : \forall K \in \mathcal{X}_h,\ \eta|_K \in [\mathbb{P}_1(K)]^d\}, \tag{45c}$$

where $\mathbb{RT}_0(K)$ denotes the lowest-order Raviart–Thomas finite-dimensional subspace associated with the element $K \in \mathcal{X}_h$, and $\mathbb{P}_l(K)$ is the space of polynomials on $K \in \mathcal{X}_h$ of total degree less than or equal to l. Thus, the spaces $(\mathcal{T}_h, \mathcal{R}_h)$ and $(\mathcal{P}_h, \mathcal{W}_h)$ are the lowest order Raviart-Thomas mixed finite element spaces for the mixed flow and heat flow subproblems, respectively. Note that both spaces satisfy the condition (3), see *e.g.* [22] for more details on (mixed) finite elements. The vector valued space $\mathcal{U}_h$ is the first order Lagrange finite element space for the mechanics problem. We employ the following stopping criterion for the iterative algorithms, given in terms of the relative

and absolute tolerances, aTOL and rTOL, i.e.

$$\begin{aligned} &\left\|(T^k, \mathbf{r}^k, p^k, \mathbf{w}^k, \mathbf{u}^k) - (T^{k-1}, \mathbf{r}^{k-1}, p^{k-1}, \mathbf{w}^{k-1}, \mathbf{u}^{k-1})\right\| \\ &\leq \text{aTOL} + \text{rTOL}\left\|(T^k, \mathbf{r}^k, p^k, \mathbf{w}^k, \mathbf{u}^k)\right\|, \end{aligned} \tag{46}$$

where we set aTOL = rTOL $= 1e-6$ for all the computations. For the solution of the linear subproblems, we make use of a direct sparse linear solver from the Python library SciPy [36], i.e., `scipy.sparse.linalg.spsolve`. The present approaches can also be combined with iterative solvers adapted to the various subproblems. All numerical tests are implemented in a finite element code written in Python, the complete source code is accessible at https://github.com/matkbrun/FEM.

5.1. Test case 1: Example with manufactured solution. As a first test case, we let the domain be a regular triangularization of the unit square, i.e., $\Omega = [0,1] \times [0,1] \subset \mathbb{R}^2$, and prescribe the following smooth solutions for the temperature, pressure and displacement

$$T(x,t) = tx_1(1-x_1)x_2(1-x_2), \tag{47a}$$

$$p(x,t) = tx_1(1-x_1)x_2(1-x_2), \tag{47b}$$

$$\mathbf{u}(x,t) = tx_1(1-x_1)x_2(1-x_2)[1,1]^\top, \tag{47c}$$

where $x := (x_1, x_2) \in \mathbb{R}^2$, $t \geq 0$. The flux fields are then computed by

$$\mathbf{r} = -\boldsymbol{\Theta}\nabla T, \quad \text{and} \quad \mathbf{w} = -\mathbf{K}\nabla p, \tag{47d}$$

while right hand sides, i.e., z, g and $\mathbf{f}$, can be calculated explicitly using equations (1a)–(1c). We prescribe homogenous initial conditions and homogenous Dirichlet boundary conditions, for the temperature, pressure and displacement. All computations are done on a fixed time step, i.e., $\tau = 1.0$, and continued until criterion (46) is satisfied.

For the analysis and comparison of our algorithms, we consider dimensionless equations, i.e. all parameters are set to $1.0e-1$, except for the three coupling coefficients $\{\alpha, \beta, b_0\}$, which we vary in order to *weaken/strengthen* the coupling between the three subproblems. In particular, we consider five different parameter regimes, **PR1** − **PR5**, specified in Table 1:

	PR1	**PR2**	**PR3**	**PR4**	**PR5**
α	1.0	0.1	0.1	1.0	0.1
β	1.0	0.1	1.0	0.1	0.1
b_0	1.0	1.0	0.1	0.1	0.1

Table 1: Smooth solution: Parameter regimes for varying strong/weak coupling between subproblems.

We also set $a_0 = c_0 = 2b_0$, thus satisfying (**A4**). Table 2 shows number of iterations needed for convergence using the six algorithms from Subsections 3.1, 3.2 and 3.3, for a single time step with decreasing mesh sizes, and stabilization according to equality in (27).

We see that for parameter regimes 1, 3 and 4 we have higher iterations numbers than for parameter regimes 2 and 5, for all six algorithms. This is because $L_T \sim \beta^2$ and

	PR1	**PR2**	**PR3**	**PR4**	**PR5**	**PR1**	**PR2**	**PR3**	**PR4**	**PR5**
h	**HFM**					**HF-M**				
1/4	7	3	8	8	3	31	4	11	11	4
1/8	7	3	7	7	3	35	4	13	13	4
1/16	6	3	7	7	3	40	4	13	13	4
1/32	6	3	7	7	3	41	4	13	13	4
1/64	6	3	7	7	3	41	4	13	13	4
h	**HM-F**					**FM-H**				
1/4	9	6	8	11	4	9	6	11	8	4
1/8	9	6	7	11	4	9	6	11	7	4
1/16	9	6	7	11	4	9	6	11	7	4
1/32	9	6	7	11	4	9	6	11	7	4
1/64	9	6	7	11	4	9	6	11	7	4
h	**H-F-M**					**F-H-M**				
1/4	20	6	11	11	4	20	6	11	11	4
1/8	22	6	12	12	4	22	6	12	12	4
1/16	24	6	13	13	4	24	6	13	13	4
1/32	24	6	13	13	4	24	6	13	13	4
1/64	24	6	13	13	4	24	6	13	13	4

Table 2: Smooth solution: Number of iteration with decreasing mesh sizes for parameter regimes **PR1** − **PR5**. Stabilization from theory.

$L_p \sim \alpha^2$, and larger stabilization results in higher iteration numbers. Furthermore, as expected, the strongly coupled parameter regime (**PR1**) yields the highest iteration numbers, in particular for the algorithms **HF-M**, **H-F-M** and **F-H-M**. Apart from this, the algorithms are performing robustly both with respect to different coupling regimes. All algorithms are performing robustly with respect to decreasing mesh sizes. For comparison we also provide in Table 3, the results without stabilization, i.e., $L_T = L_p = 0$.

We see here that the fully monolithic algorithm (**HFM**) has low iteration counts for all parameter regimes since this is only a linearization scheme, and does not require stabilization (cf. Theorem 4.2). For the two-level (Section 3.2) and three-level (Section 3.3) algorithms, which involves some splitting as well as linearization, we see that iteration counts for different parameter regimes corresponds to the various coupling/decoupling of the subproblems present in the algorithms (splitting of subproblems which are strongly coupled yields high iteration numbers, compared to solving the strongly coupled subproblems together). This is in contrast to employing stabilization, which greatly improves the robustness of the algorithms with respect to variations in parameters. For the strongly coupled parameter regime (**PR1**), we even have no convergence for algorithm **HF-M**, when no stabilization is applied.

Furthermore, in order to check the robustness of the proposed schemes with respect to the nonlinearity, we adjust the coefficient of the nonlinear term, c_f, in order to make this term dominate. Table 4 shows number of iterations needed for convergence when $c_f = 1.0e1$, for both the strongly coupled parameter regime (**PR1**) and the weakly coupled parameter regime (**PR5**). We also compare the results when no stabilization is applied. Note that we here only use a single mesh with $h = 1/16$.

For the weakly coupled parameter regime (**PR5**), there is no difference in it-

	PR1	**PR2**	**PR3**	**PR4**	**PR5**	**PR1**	**PR2**	**PR3**	**PR4**	**PR5**
h	**HFM**					**HF-M**				
1/4	3	3	3	3	3	-	4	16	16	4
1/8	3	3	3	3	3	-	4	19	19	4
1/16	3	3	3	3	3	-	4	20	20	4
1/32	3	3	3	3	3	-	4	20	20	4
1/64	3	3	3	3	3	-	4	20	21	4
h	**HM-F**					**FM-H**				
1/4	11	6	4	22	4	11	6	21	4	4
1/8	11	6	4	23	4	11	6	23	4	4
1/16	12	6	4	24	4	11	6	24	4	4
1/32	12	6	4	24	4	12	6	24	4	4
1/64	12	6	4	25	4	12	6	24	4	4
h	**H-F-M**					**F-H-M**				
1/4	34	6	17	16	4	34	6	16	17	4
1/8	38	5	19	19	4	38	5	19	19	4
1/16	44	5	20	20	4	44	5	20	20	4
1/32	46	5	20	20	4	46	5	20	21	4
1/64	46	5	21	20	4	46	5	20	21	4

Table 3: Smooth solution: Number of iterations with decreasing mesh sizes for parameter regimes **PR1** – **PR5**. $L_T = L_p = 0$.

Parameters	**PR1**	**PR5**	**PR1**	**PR5**
#	**HFM**		**HF-M**	
Non-stabilized	4	4	-	5
Stabilized	7	4	41	5
#	**HM-F**		**FM-H**	
Non-stabilized	11	4	10	4
Stabilized	9	4	8	4
#	**H-F-M**		**F-H-M**	
Non-stabilized	48	5	36	4
Stabilized	25	5	22	4

Table 4: Smooth solution: Number of iterations with strong nonlinear effects, i.e. $c_f = 10$, and mesh size $h = 1/16$.

eration numbers between the stabilized and non-stabilized algorithms, even with a dominating nonlinearity. For the strongly coupled parameter regime (**PR1**), however, the stabilized algorithms has a significantly lower iteration count. This might be due to the fact that the nonlinearity appears as a coupling term.

Since analytical solutions are available for this problem, we provide also the discretization errors, denoted by $(e_{h,T}, e_{h,\mathbf{r}}, e_{h,p}, e_{h,\mathbf{w}}, e_{h,\mathbf{u}})$, measured in the L^2-norm. Due to almost no variation in discretization errors between the six algorithms and between the different parameter regimes (less than 5%), we provide in Table 5 the

discretization errors using algorithm **H-F-M** applied on the weakly coupled parameter regime (**PR5**). We also include the convergence rates, defined by $r_T := e_{h_j,T}/e_{h_{j+1},T}$, and similarly for the other variables.

h	$e_{h,T}$	r_T	$e_{h,\mathbf{r}}$	$r_{\mathbf{r}}$	$e_{h,p}$	r_p	$e_{h,\mathbf{w}}$	$r_{\mathbf{w}}$	$e_{h,\mathbf{u}}$	$r_{\mathbf{u}}$
1/4	8.5e-3	-	3.5e-3	-	8.5e-3	-	3.5e-3	-	5.6e-3	-
1/8	4.4e-3	1.93	1.8e-3	1.94	4.4e-3	1.93	1.8e-3	1.94	1.4e-3	4.0
1/16	2.2e-3	2.0	9.3e-4	1.94	2.2e-3	2.0	9.3e-4	1.94	3.6e-4	3.89
1/32	1.1e-3	2.0	4.7e-4	1.98	1.1e-3	2.0	4.7e-4	1.98	9.1e-5	3.96
1/64	5.5e-4	2.0	2.3e-4	2.04	5.5e-4	2.0	2.3e-4	2.04	2.3e-5	3.96

Table 5: Smooth solution: Discretization errors using algorithm **H-F-M** applied on the weakly coupled parameter regime (**PR5**), and with $c_f = 0.1$. Stabilization from theory. Convergence rate is of first order for all variables, except for that of the displacement which is of second order. We note that these rates are optimal.

5.2. Test case 2: Mandel's problem. We refer to [16] for a detailed description of Mandel's problem. Formulas for the analytical pressure and displacements can be found in [37]. We provide here only a brief description. Mandel's problem is posed on a rectangular domain representing a poroelastic slab of extent $2a$ in the horizontal direction, $2b$ in the vertical direction, and infinitely long in the third direction. The poroelastic slab is contained between two rigid plates, where at the initial time a downward force of magnitude $2F$ is applied to the top plate, with an equal but opposite force applied to the bottom plate. The top and bottom boundary is treated as impermeable, while zero pressure (and temperature) is prescribed at the right and left boundary. Due to the nature of Mandel's problem, the pressure, temperature and horizontal component of the displacement varies only in the horizontal direction, while the vertical component of the displacement varies only in the vertical direction. From symmetry considerations, it suffices to consider only the top right quarter rectangle, i.e. the computational domain is $[0, a] \times [0, b]$ (see Figure 1).

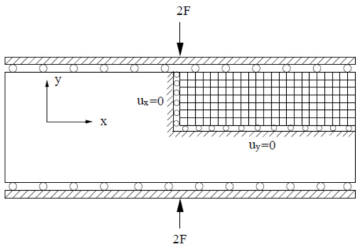


Fig. 1: Setting of Mandel's problem quarter domain (figure from [32]).

We perform now all computations with realistic choices of physical parameters.

In particular, we take mechanics and flow parameters identical to [32], and heat parameters identical to [26]. However, in [26] the flow-heat coupling coefficient b_0 is taken to be identically zero, so in order to preserve this coupling we instead choose a suitably small number (that satisfies (**A4**)). All parameters are listed in Table 6.

Symbol	Quantity	Value	Unit
E	Bulk modulus	5.94e9	Pa
ν	Poisson's ratio	0.2	-
c_0	Storage coefficient	6.06e-11	Pa^{-1}
α	Biot's coefficient	1.0	-
μ_f	Fluid viscosity	1.0e-3	$\mathrm{Pa\,s}$
$\hat{\mathbf{K}}$	Permeability	9.87e-14 $\mathbf{I}$	m^2
$\boldsymbol{\Theta}$	Effective thermal conductivity	1.7 $\mathbf{I}$	$\mathrm{W\,m^{-1}\,K^{-1}}$
b_0	Thermal dilation coefficient	3.03e-11	K^{-1}
β	Thermal stress coefficient	9.9e6	$\mathrm{Pa\,K^{-1}}$
a_0	Effective heat capacity	0.92e3	$\mathrm{J\,kg^{-1}\,K^{-1}}$
T_{ref}	Reference temperature	298.15	K
c_f	Volumetric heat capacity fluid	4.18e6	$\mathrm{J\,m^{-3}\,K^{-1}}$
τ	Time step	10	s

Table 6: Mandel's problem: Physical parameters, taken from [32, 26].

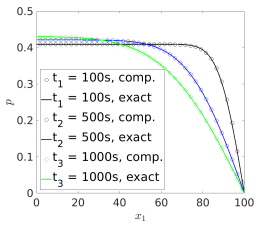
In terms of our previous notation, we now have $\mathbf{K} = \mu_f^{-1}\hat{\mathbf{K}}$, and $\mu = \dfrac{E}{2(1+\nu)}$ and $\lambda = \dfrac{E\nu}{(1+\nu)(1+2\nu)}$. Note also that we will now employ the dimensional version of the heat equation (1a), which reads (in primal form)

$$\partial_t \left(a_0 \frac{T}{T_{\mathrm{ref}}} - b_0 p + \beta \nabla \cdot \mathbf{u} \right) + c_f (\mathbf{K}\nabla p) \cdot \nabla \frac{T}{T_{\mathrm{ref}}} - \nabla \cdot \left(\boldsymbol{\Theta} \nabla \frac{T}{T_{\mathrm{ref}}} \right) = z. \tag{48}$$

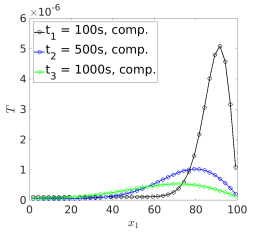
The magnitude of the compressive force is $F = 2 \times 10^8\,\mathrm{Pa\,m}$, and the physical dimensions of the quarter rectangle is given by $a = 100\,\mathrm{m}$ and $b = 10\,\mathrm{m}$, of which we make a regular triangularization. We impose the compressive force as a Dirichlet boundary condition on the top boundary ($x_2 = b$) for the vertical component of the displacement. We denote by n_1 and n_2 the number of subdivisions of the domain in the x_1 and x_2 directions, respectively. For the first implementation of Mandel's problem we prescribe homogenous boundary conditions and zero source term and initial condition for the heat problem. Figure 2 shows the solution profiles for the pressure, temperature and displacements for selected time steps, with the analytical (isothermal) solutions for the pressure and displacement included for comparison.

The computed solutions for pressure and displacement matches the analytical ones, even though the analytical solutions are only valid for the linear isothermal problem. This is because the induced temperature effect in the system is small enough that the heat decouples from the flow and mechanics. For the second implementation of Mandel's problem we prescribe a constant source term for the heat problem, i.e., $z = 2 \times 10^{-4}\,\mathrm{W\,m^{-3}\,K^{-1}}$ and zero initial condition. Figure 3 shows the solution profiles for the pressure, temperature and displacements at selected time steps.

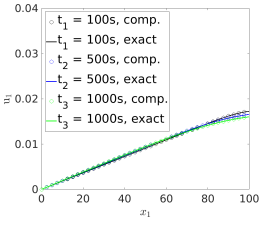
The temperature source now interacts with the other processes and thus has an effect on the pressure and horizontal component of the displacement. Furthermore,



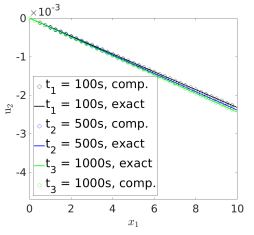
(a) Pressure profile.



(b) Temperature profile.



(c) Displacement profile, 1st component.

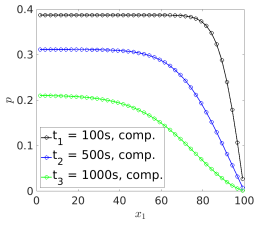


(d) Displacement profile, 2nd component.

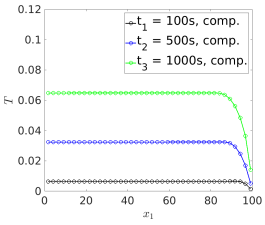
Fig. 2: Mandel's problem: Solution profiles for Mandel's problem at $t \in \{100\,\mathrm{s}, 500\,\mathrm{s}, 1000\,\mathrm{s}\}$, with $z = 0\,\mathrm{W\,m^{-3}\,K^{-1}}$, and $n_1 = n_2 = 40$.

the temperature change in the system is now increasing with increasing time. Table 7 shows the number of iterations for Mandel's problem using the derived algorithms.

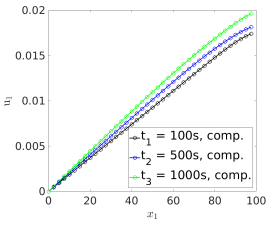
6. Conclusions. Based on developments on iterative splitting schemes from linear poroelasticity, we have proposed six novel iterative procedures for nonlinear thermo-poroelasticity. These algorithms are using stabilization and linearization techniques similar to [8, 31], which is known in the literature as the 'L-scheme'. The thermo-poroelastic problem we consider can be viewed as a coupling of three physical processes (or subproblems): Flow, geomechanics and heat. Solving this system either monolithically (all three subproblems simultaneously), partially decoupled (two subproblems simultaneously), or fully decoupled (each subproblem separately), yields six possible combinations of coupling/decoupling, which we have used to design our six algorithms. All of these involve a linearization of the convective term and added stabilization terms to both the flow and heat subproblems. In this sense, our use of the L-scheme is both as a stabilization for iterative splitting and a linearization of nonlinear problems.



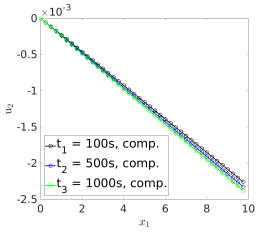
(a) Pressure profile.



(b) Temperature profile.



(c) Displacement profile, 1st component.



(d) Displacement profile, 2nd component.

Fig. 3: Mandel's problem: Solution profiles at $t \in \{100\,\mathrm{s}, 500\,\mathrm{s}, 1000\,\mathrm{s}\}$, with $z = 2 \times 10^{-4}\,\mathrm{W\,m^{-3}\,K^{-1}}$, and $n_1 = n_2 = 40$.

Heat source	$z = 0$	z = 2e-4	$z = 0$	z = 2e-4	$z = 0$	z = 2e-4
$n_1 = n_2$	**HFM**		**HF–M**		**HM–F**	
10	18	18	14	14	14	14
20	18	18	13	12	13	12
40	18	18	13	12	13	12
$n_1 = n_2$	**FM–H**		**H–F–M**		**F–H–M**	
10	18	18	14	13	14	14
20	18	18	13	13	13	12
40	18	18	13	13	13	12

Table 7: Mandel's problem: Number of iterations with decreasing mesh sizes for Mandel's problem. Stabilization from theory.

For any given situation the coupling strength between the three subproblems may vary. A-priori, the expectation is that solving together subproblems that are strongly coupled yields better efficiency properties than does splitting. On the other hand, if the coupling between two or more subproblems is weak, a splitting procedure might be beneficial. For this reason, and due to the fact that splitting the three-way coupled multi-physics problem into smaller subproblems allows for combining existing codes that separately can handle any of the three processes involved (or two of them combined), six different algorithms are presented. These six algorithms covers all possibilities of strong/weak coupling between the three subproblems. Using the well-posedness of the continuous problem, we obtained lower bounds on the stabilization parameters, and proved the convergence of our proposed algorithms under a constraint on the time step. In practice, however, we find that this bound is not tight; as long as the fluxes are not becoming unbounded (*e.g.* due to a singularity), a 'reasonable' time step can safely be chosen.

Our algorithms are tested in detail with several numerical examples. In particular, we find that all six algorithms are performing robustly with respect to both mesh refinement and different parameter regimes (i.e. strong/weak coupling between the subproblems and strong/weak nonlinear effects), using the stabilization revealed by our analysis. We also find that using no stabilization results in the algorithms being more sensitive to the parameter regimes, i.e. splitting subproblems that are strongly coupled yields high iteration numbers compared to solving these subproblems together. This phenomena is also observed in the stabilized algorithms, but to a significantly lesser extent. In particular, our conclusion is that with no stabilization, each of the algorithms is suitable only for a certain parameter regime (i.e. one that corresponds to the coupling/decoupling structure present in the algorithm), in contrast to the stabilized algorithms, which can handle a much wider range of different parameter regimes.

Acknowledgment. The research is supported by the Norwegian Research Council Toppforsk project 250223 (The TheMSES project: https://themses.w.uib.no).

References.

[1] E. Ahmed, F. Adrian Radu, and J. M. Nordbotten, *Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model*, research report, Department of Mathematics, University of Bergen, Nov. 2018, https://hal.inria.fr/hal-01687026.

[2] E. Ahmed, J. Jaffré, and J. E. Roberts, *A reduced fracture model for two-phase flow with different rock types*, Math. Comput. Simulation, 137 (2017), pp. 49–70, https://doi.org/10.1016/j.matcom.2016.10.005, https://doi.org/10.1016/j.matcom.2016.10.005.

[3] E. Ahmed, J. M. Nordbotten, and F. A. Radu, *Adaptive asynchronous time-stepping, stopping criteria, and a posteriori error estimates for fixed-stress iterative schemes for coupled poromechanics problems*, arXiv preprint arXiv:1901.01206, (2019).

[4] E. Ahmed, F. A. Radu, and J. M. Nordbotten, *Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 264–294, https://doi.org/10.1016/j.cma.2018.12.016, https://doi.org/10.1016/j.cma.2018.12.016.

[5] M. A. Biot, *General theory of three-dimensional consolidation*, Journal of Applied Physics, 12 (1941), pp. 155–164.

[6] M. A. Biot, *Theory of finite deformations of porous solids*, Indiana University Mathematics Journal, 21 (1972), pp. 597–620.
[7] M. Borregales, F. A. Radu, K. Kumar, and J. M. Nordbotten, *Robust iterative schemes for non-linear poromechanics*, Comput. Geosci., 22 (2018), pp. 1021–1038, https://doi.org/10.1007/s10596-018-9736-6, https://doi.org/10.1007/s10596-018-9736-6.
[8] J. W. Both, M. Borregales, J. M. Nordbotten, K. Kumar, and F. A. Radu, *Robust fixed stress splitting for Biot's equations in heterogeneous media*, Appl. Math. Lett., 68 (2017), pp. 101–108, https://doi.org/10.1016/j.aml.2016.12.019, https://doi.org/10.1016/j.aml.2016.12.019.
[9] M. K. Brun, E. Ahmed, J. M. Nordbotten, and F. A. Radu, *Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport*, Journal of Mathematical Analysis and Applications, 471 (2019), pp. 239–266.
[10] M. K. Brun, I. Berre, J. M. Nordbotten, and F. A. Radu, *Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium*, Transport in Porous Media, (2018), https://doi.org/10.1007/s11242-018-1056-8.
[11] M. Bukač, I. Yotov, and P. Zunino, *An operator splitting approach for the interaction between a fluid and a multilayered poroelastic structure*, Numer. Methods Partial Differential Equations, 31 (2015), pp. 1054–1100, https://doi.org/10.1002/num.21936, https://doi.org/10.1002/num.21936.
[12] N. Castelletto, J. White, and H. Tchelepi, *A unified framework for fully-implicit and sequential-implicit schemes for coupled poroelasticity*, in ECMOR XIV-14th European Conference on the Mathematics of Oil Recovery, 2014.
[13] N. Castelletto, J. White, and H. Tchelepi, *Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics*, International Journal for Numerical and Analytical Methods in Geomechanics, 39 (2015), pp. 1593–1618.
[14] C. Chainais-Hillairet and J. Droniou, *Convergence analysis of a mixed finite volume scheme for an elliptic-parabolic system modeling miscible fluid flows in porous media*, SIAM J. Numer. Anal., 45 (2007), pp. 2228–2258, https://doi.org/10.1137/060657236, https://doi.org/10.1137/060657236.
[15] W. Cheney, *Analysis for applied mathematics*, vol. 208 of Graduate Texts in Mathematics, Springer-Verlag, New York, 2001, https://doi.org/10.1007/978-1-4757-3559-8, https://doi.org/10.1007/978-1-4757-3559-8.
[16] O. Coussy, *Poromechanics*, John Wiley & Sons, 2004.
[17] B. a. da Veiga, J. Droniou, and G. Manzini, *A unified approach for handling convection terms in finite volumes and mimetic discretization methods for elliptic problems*, IMA J. Numer. Anal., 31 (2011), pp. 1357–1401, https://doi.org/10.1093/imanum/drq018, https://doi.org/10.1093/imanum/drq018.
[18] C. N. Dawson, H. Klíe, M. F. Wheeler, and C. S. Woodward, *A parallel, implicit, cell-centered method for two-phase flow with a preconditioned Newton-Krylov solver*, Comput. Geosci., 1 (1997), pp. 215–249 (1998), https://doi.org/10.1023/A:1011521413158, https://doi.org/10.1023/A:1011521413158.
[19] F. Doster, J. M. Nordbotten, et al., *Full pressure coupling for geomechanical multi-phase multi-component flow simulations*, in SPE Reservoir Simulation Symposium, Society of Petroleum Engineers, 2015.
[20] M. A. Fernández, J.-F. Gerbeau, and C. Grandmont, *A projection semi-implicit scheme for the coupling of an elastic structure with an incompressible*

fluid, Internat. J. Numer. Methods Engrg., 69 (2007), pp. 794–821, https://doi.org/10.1002/nme.1792, https://doi.org/10.1002/nme.1792.

[21] J. G. Garcia, L. W. Teufel, et al., *Numerical simulation of fully coupled fluid-flow/geomechanical deformation in hydraulically fractured reservoirs*, in SPE Production Operations Symposium, Society of Petroleum Engineers, 2005.

[22] G. N. Gatica, *A simple introduction to the mixed finite element method*, Theory and Applications. Springer Briefs in Mathematics. Springer, London, (2014).

[23] B. Gatmiri and P. Delage, *A formulation of fully coupled thermal–hydraulic–mechanical behaviour of saturated porous media—numerical approach*, International Journal for Numerical and Analytical Methods in Geomechanics, 21 (1997), pp. 199–225.

[24] U. Hornung, *Homogenization and porous media*, vol. 6, Springer Science & Business Media, 2012.

[25] O. Iliev, A. Kolesov, and P. Vabishchevich, *Numerical solution of plate poroelasticity problems*, Transport in Porous Media, 115 (2016), pp. 563–580.

[26] J. Kim, *Unconditionally stable sequential schemes for thermoporomechanics: Undrained-adiabatic and extended fixed-stress splits*, dim, 1 (2015), p. 2.

[27] J. Kim, H. A. Tchelepi, R. Juanes, et al., *Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics*, in SPE reservoir simulation symposium, Society of Petroleum Engineers, 2009.

[28] A. E. Kolesov and P. N. Vabishchevich, *Splitting schemes with respect to physical processes for double-porosity poroelasticity problems*, Russian Journal of Numerical Analysis and Mathematical Modelling, 32 (2017), pp. 99–113.

[29] A. E. Kolesov, P. N. Vabishchevich, and M. V. Vasilyeva, *Splitting schemes for poroelasticity and thermoelasticity problems*, Computers & Mathematics with Applications, 67 (2014), pp. 2185–2198.

[30] C. K. Lee and C. C. Mei, *Thermal consolidation in porous media by homogenization theory—i. derivation of macroscale equations*, Advances in water resources, 20 (1997), pp. 127–144.

[31] F. List and F. A. Radu, *A study on iterative methods for solving Richards' equation*, Comput. Geosci., 20 (2016), pp. 341–353, https://doi.org/10.1007/s10596-016-9566-3, https://doi.org/10.1007/s10596-016-9566-3.

[32] A. Mikelić, B. Wang, and M. F. Wheeler, *Numerical convergence study of iterative coupling for coupled flow and geomechanics*, Comput. Geosci., 18 (2014), pp. 325–341, https://doi.org/10.1007/s10596-013-9393-8, https://doi.org/10.1007/s10596-013-9393-8.

[33] A. Mikelić and M. F. Wheeler, *Convergence of iterative coupling for coupled flow and geomechanics*, Comput. Geosci., 17 (2013), pp. 455–461, https://doi.org/10.1007/s10596-012-9318-y, https://doi.org/10.1007/s10596-012-9318-y.

[34] S. E. Minkoff, C. Stone, S. Bryant, M. Peszynska, and M. F. Wheeler, *Coupled fluid flow and geomechanical deformation modeling*, Journal of Petroleum Science and Engineering, 38 (2003), pp. 37 – 56, https://doi.org/https://doi.org/10.1016/S0920-4105(03)00021-4, http://www.sciencedirect.com/science/article/pii/S0920410503000214.

[35] D. Néron and D. Dureisseix, *A computational strategy for thermo-poroelastic structures with a time-space interface coupling*, Internat. J. Numer. Methods Engrg., 75 (2008), pp. 1053–1084, https://doi.org/10.1002/nme.2283, https://doi.org/10.1002/nme.2283.

[36] T. E. Oliphant, *Python for scientific computing*, Computing in Science & Engineering, 9 (2007), pp. 10–20.

[37] P. J. Phillips and M. F. Wheeler, *A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity*, Comput. Geosci., 12 (2008), pp. 417–435, https://doi.org/10.1007/s10596-008-9082-1, https://doi.org/10.1007/s10596-008-9082-1.

[38] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373, https://doi.org/10.1016/j.cam.2003.04.008, https://doi.org/10.1016/j.cam.2003.04.008.

[39] D. Seus, K. Mitra, I. S. Pop, F. A. Radu, and C. Rohde, *A linear domain decomposition method for partially saturated flow in porous media*, Comput. Methods Appl. Mech. Engrg., 333 (2018), pp. 331–355, https://doi.org/10.1016/j.cma.2018.01.029, https://doi.org/10.1016/j.cma.2018.01.029.

[40] S. Sun, B. Rivière, and M. F. Wheeler, *A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media*, in Recent progress in computational and applied PDEs (Zhangjiajie, 2001), Kluwer/Plenum, New York, 2002, pp. 323–351.

[41] S. Sun and M. F. Wheeler, *Discontinuous Galerkin methods for coupled flow and reactive transport problems*, Appl. Numer. Math., 52 (2005), pp. 273–298, https://doi.org/10.1016/j.apnum.2004.08.035, https://doi.org/10.1016/j.apnum.2004.08.035.

[42] A. P. Suvorov and A. P. S. Selvadurai, *Macroscopic constitutive equations of thermo-poroviscoelasticity derived using eigenstrains*, J. Mech. Phys. Solids, 58 (2010), pp. 1461–1473, https://doi.org/10.1016/j.jmps.2010.07.016, https://doi.org/10.1016/j.jmps.2010.07.016.

[43] K. Terzaghi, *Theoretical soil mechanics*, Chapman And Hali, Limited John Wiler And Sons, Inc; New York, 1944.

[44] J. M. Thomas, *Méthode des éléments finis équilibre*, in Journées "Éléments Finis" (Rennes, 1975), Univ. Rennes, Rennes, 1975, p. 25.

[45] D. Tran, L. Nghiem, L. Buchanan, et al., *An overview of iterative coupling between geomechanical deformation and reservoir flow*, in SPE International Thermal Operations and Heavy Oil Symposium, Society of Petroleum Engineers, 2005.

[46] C. J. Van Duijn, A. Mikelic, M. Wheeler, and T. Wick, *Thermoporoelasticity via homogenization i. modeling and formal two-scale expansions*. Nov. 2017, https://hal.archives-ouvertes.fr/hal-01650194.

[47] J. A. White, N. Castelletto, and H. A. Tchelepi, *Block-partitioned solvers for coupled poromechanics: a unified framework*, Comput. Methods Appl. Mech. Engrg., 303 (2016), pp. 55–74, https://doi.org/10.1016/j.cma.2016.01.008, https://doi.org/10.1016/j.cma.2016.01.008.

[48] S.-Y. Yi, *Convergence analysis of a new mixed finite element method for Biot's consolidation model*, Numer. Methods Partial Differential Equations, 30 (2014), pp. 1189–1210, https://doi.org/10.1002/num.21865, https://doi.org/10.1002/num.21865.

Paper D

An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters

M. K. Brun, T. Wick, I. Berre, J. M. Nordbotten, F. A. Radu

In review

An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters

Mats Kirkesæther Brun[†] Thomas Wick[‡] Inga Berre[†§] Jan Martin Nordbotten[†¶]
Florin Adrian Radu[†]

March 20, 2019

Abstract

This paper concerns the analysis and implementation of a novel iterative staggered scheme for quasi-static brittle fracture propagation models, where the fracture evolution is tracked by a phase field variable. The model we consider is a two-field variational inequality system, with the phase field function and the elastic displacements of the solid material as independent variables. Using a penalization strategy, this variational inequality system is transformed into a variational equality system, which is the formulation we take as the starting point for our algorithmic developments. The proposed scheme involves a partitioning of this model into two subproblems; *phase field* and *mechanics*, with added stabilization terms to both subproblems for improved efficiency and robustness. We analyze the convergence of the proposed scheme using a fixed point argument, and find that under a natural condition, the elastic mechanical energy remains bounded, and, if the diffusive zone around crack surfaces is sufficiently thick, monotonic convergence is achieved. Finally, the proposed scheme is validated numerically with several bench-mark problems.

Key words: phase field; quasi-static; brittle fracture; fracture propagation; L-scheme; fixed stress; iterative algorithm; linearization; convergence analysis; fixed point; finite element;

1 Introduction

Fracture propagation is currently an important topic with many applications in various engineering fields. Specifically, phase-field descriptions are intensively investigated. The theory of brittle fracture mechanics goes back to the works of A. Griffith [23], wherein a criterion for crack propagation is formulated. Despite a foundational treatment on the subject of brittle fracture, Griffith's theory fails to predict crack initiation. This deficiency can however be overcome by a variational approach, which was first proposed in [10, 20]. Using such a variational approach, discontinuities in the displacement field u across the lower-dimensional crack surface are approximated by an auxiliary phase-field function φ. The latter can be viewed as an indicator function, which introduces a diffusive transition zone between the broken and the unbroken material. The enforcement of irreversibility of crack growth finally yields a variational inequality system, of which we seek the solution $\{u, \varphi\}$.

In this work, we concentrate on improvements of the nonlinear solution algorithm, which is still a large bottleneck of phase-field fracture evolution problems. Specifically, high iteration numbers when the crack initiates or is further growing are reported in many works [21, 29, 44, 45]. However, in most studies iteration numbers are omitted. Both staggered (splitting) schemes and monolithic schemes are frequently employed. Important developments include alternating minimization/staggered schemes [9, 11, 12, 29, 30], quasi-monolithic scheme with a partial linearization [25], and fully monolithic schemes [21, 44, 45].

The goal of this work is to propose a *linearized staggered scheme with stabilizing parameters*. In particular, the proposed scheme is based on recent developments on *iterative splitting schemes* coming from poroelasticity [13, 26, 32, 33]. Iterative splitting schemes are widely applied to problems of coupled flow and mechanics, where at each iteration step either of the subproblems (i.e., flow or mechanics) is solved first, keeping some physical quantity constant (e.g., *fixed stress* or *fixed strain*), followed by solving the next subproblem with updated solution information. This procedure is then repeated until an accepted tolerance is reached. Further extensions of this technique involves tuning some artificial stabilization terms according to a derived contraction estimate in energy

[†]Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020 Bergen, Norway. mats.brun@uib.no, inga.berre@uib.no, jan.nordbotten@uib.no, florin.radu@uib.no

[‡]Leibniz University Hannover, Institute of Applied Mathematics, AG Wissenschaftliches Rechnen, Welfengarten 1, 30167 Hannover. thomas.wick@ifam.uni-hannover.de

[§]NORCE Norwegian Research Centre AS, Bergen, Norway.

[¶]Department of Civil and Environmental Engineering, Princeton University, Princeton, N. J., USA.

norms. Here, the quantity held constant during solving of the subproblems need not represent any physical quantity present in the model. This is the central idea in the so-called 'L-scheme', which has proven to perform robustly for Richards equation [28, 38], for linear and nonlinear poroelasticity [7, 8], and for nonlinear thermo-poroelasticity [27].

We propose here a variant of the L-scheme, adapted to phase field brittle fracture propagation models. This scheme is based on a partitioning of the model into two subproblems; *phase field* and *mechanics*. Here, the L-scheme acts both as a *stabilization* and as a *linearization* (as a linearization scheme, the stabilization parameters mimics the Jacobian from Newton iteration). Assuming that the mechanical elastic energy remains bounded during the iterations, and that the diffusive zone around crack surfaces is sufficiently thick, we give a proof of monotonic convergence of the proposed scheme by employing a fixed point argument.

The efficiency and robustness of the proposed scheme is demonstrated numerically with several bench-mark problems. Moreover, we compare the number of iterations needed for convergence with 'standard' staggered schemes (i.e., without stabilizing terms), and monolithic schemes in which the fully-coupled system is solved all-at-once. Furthermore, it is well known that when reaching the critical loading steps during the computation of brittle fracture phase field problems (i.e., when the crack is propagating), spikes in iteration numbers appear. For this reason, and thanks to the monotonic convergence property of the proposed scheme, we show that a (low) upper bound on the number of iterations may be enforced, while the computed results are still in very good agreement with the non-truncated solutions. Thus, using this 'truncated L-scheme', we effectively avoid the iteration spikes at the critical loading steps at the cost of negligible loss of accuracy. We mention that this strategy is not available with e.g. Newton iteration, as the iterate solutions may behave erratically for any number of iterations before finally converging. Moreover, the assumption that the mechanical elastic energy remains bounded during the iterations is verified numerically for all tests cases.

The main aims of this work are three-fold: Under a natural assumption, we prove the convergence of a novel iterative staggered scheme, optimized for phase field brittle fracture propagation problems. Based on these theoretical findings, we design a robust solution algorithm with monotonic convergence properties. Finally, several numerical tests are presented in which our variants of the L-scheme are tested in detail.

The outline of this paper is as follows: In Section 2 we present the model equations and coefficients, in Section 3 we introduce the partitioned scheme and derive a convergence proof, in Section 4 we describe in detail our numerical algorithm in pseudo-code, and in Section 5 we provide several numerical experiments, in particular the *single edge notched tension test*, the *single edge notched shear test*, and the *L-shaped panel test*. Finally, in Section 6 we provide some conclusions and summary of the work.

1.1 Preliminaries

In this section we explain the notation used throughout this article, see e.g. [18, 47] for more details. Given an open and bounded set $B \subset \mathbb{R}^d$, $d \in \{2, 3\}$, and $1 \leq p < \infty$, let $L^p(B) = \{f : B \to \mathbb{R} : \int_B |f(x)|^p \mathrm{d}x < \infty\}$. For $p = \infty$, let $L^\infty(B) = \{f : B \to \mathbb{R} : \operatorname{ess\,sup}_{x\in B} |f(x)| < \infty\}$. In particular, $L^2(B)$ is the Hilbert space of square integrable functions with inner product $(\cdot, \cdot)$ and norm $\|f\| := (f, f)^{\frac{1}{2}}$. For $k \in \mathbb{N}$, $k \geq 0$, we denote by $W^{k,p}(B)$ the space of functions in $L^p(B)$ admitting weak derivatives up to k'th order. In particular, $H^1(B) := W^{1,2}(B)$ and we denote by $H^1_0(B)$ its zero trace subspace.

Note that we reserve the use of bold fonts for second order tensors. Hence, if $u, v \in L^2(B)$, their inner product is $(u, v) := \int_B u(x)v(x)\mathrm{d}x$, and similarly, if $u, v \in (L^2(B))^d$ then we take their inner product to be $(u, v) := \int_B u(x) \cdot v(x)\mathrm{d}x$. Finally, if $\mathbf{u}, \mathbf{v} \in (L^2(B))^{d\times d}$ then their inner product is $(\mathbf{u}, \mathbf{v}) := \int_B \mathbf{u}(x) : \mathbf{v}(x)\mathrm{d}x$.

We will also frequently apply several classical inequalities, in particular: *Cauchy-Schwarz, Young, Poincaré*, and *Korn*. See e.g. [15, 24] for a detailed description of these.

2 Governing equations

What follows is a brief description of the phase field approach for quasi-static brittle fracture propagation, see e.g. [20, 30] for more details. Consider a (bounded open) polygonal domain $B \subset \mathbb{R}^d$, wherein $\mathcal{C} \subset \mathbb{R}^{d-1}$ denotes the fracture, and $\Omega \subset \mathbb{R}^d$ is the intact domain, and a time interval $(0, T)$ is given with final time $T > 0$. By introducing the phase field variable $\varphi : B \times (0, T) \to [0, 1]$, which takes the value 0 in the fracture, 1 in the intact domain, and varies smoothly from 0 to 1 in a transition zone of (half-)thickness $\varepsilon > 0$ around $\mathcal{C}$, the evolution of the fracture can be tracked in space and time. Using the phase field approach, the fracture $\mathcal{C}$ is approximated by $\Omega_F \subset \mathbb{R}^d$, where $\Omega_F := \{x \in \mathbb{R}^d : \varphi(x) < 1\}$.

Introducing the displacement vector $u : B\times(0, T) \to \mathbb{R}^d$, the model problem we consider arises as a minimization problem: An energy functional $E(u, \varphi)$ is defined according to Griffith's criterion for brittle fracture [23], which is then sought to be minimized over all admissible $\{u, \varphi\}$. From this minimization problem, the Euler-Lagrange equations are obtained by differentiation with respect to the arguments, yielding a variational equality system. Finally, a crack irreversibility condition must be enforced (the crack is not allowed to heal), which takes the form $\partial_t\varphi \leq 0$. Thus, the variational equality system, which is the previously mentioned Euler-Lagrange equations, is

transformed into a variational inequality system, which reads as follows: Find $(u(t), \varphi(t)) \in V \times W := (H_0^1(B))^d \times W^{1,\infty}(B)$ such that for $t \in (0, T]$ there holds

$$(g(\varphi)\mathbb{C}\mathbf{e}(u), \mathbf{e}(v)) = (b, v), \ \forall v \in V, \tag{2.1a}$$

$$G_c\varepsilon(\nabla\varphi, \nabla\psi) - \frac{G_c}{\varepsilon}(1-\varphi, \psi) + (1-\kappa)(\varphi|\mathbb{C}\mathbf{e}(u)|^2, \psi) \geq 0, \quad \forall \psi \in W, \tag{2.1b}$$

where $G_c > 0$ is the critical elastic energy restitution rate, $0 < \kappa << 1$ is a regularization parameter, the purpose of which is to avoid degeneracy of the elastic energy (equivalent with replacing the fracture with a softer material), and $g(\varphi) := (1-\kappa)\varphi^2 + \kappa$ is a standard choice for the degradation function (see e.g. [39, 45]. Note that $g(\varphi) \to \kappa$ when approaching the fracture zone). The body force acting on the domain B is $b : B \times (0, T) \to \mathbb{R}^d$, and $|\mathbb{C}\mathbf{e}(u)|^2 := \mathbb{C}\mathbf{e}(u) : \mathbf{e}(u)$ is the elastic mechanical energy, where $\mathbf{e}(\cdot) := (\nabla(\cdot) + \nabla(\cdot)^\top)/2$ is the symmetric gradient, and $\mathbb{C} = [C_{ijkl}]_{ijkl}$ is the fourth order tensor containing the elastic material coefficients, where each $C_{ijkl} \in L^\infty(B)$. We assume that $\mathbb{C}$ satisfies the usual *symmetry* and *positive definiteness* properties, i.e., $(\mathbb{C}\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbb{C}\mathbf{v})$, and $(\mathbb{C}\mathbf{u}, \mathbf{u})^{\frac{1}{2}}$ defines an L^2-equivalent norm, i.e., there exists constants $\lambda_m, \lambda_M > 0$ such that

$$\lambda_m\|\mathbf{u}\| \leq (\mathbb{C}\mathbf{u}, \mathbf{u})^{\frac{1}{2}} \leq \lambda_M\|\mathbf{u}\|, \quad \text{for} \quad \mathbf{u}, \mathbf{v} \in (L^2(B))^{d\times d}, \ \mathbf{u}, \mathbf{v} \neq \mathbf{0}. \tag{2.2}$$

In order to facilitate the following developments we assume continuity in time for $\{u, \varphi, b\}$. Let now $0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of the time interval $(0, T)$, with time step $\delta t := t^n - t^{n-1}$, and denote the time discrete solutions by

$$u^n := u(\cdot, t^n), \tag{2.3}$$

$$\varphi^n := \varphi(\cdot, t^n). \tag{2.4}$$

The irreversibility condition now becomes $\varphi^n \leq \varphi^{n-1}$ (using a backward Euler method), and the time-discrete version of the problem (2.1a)–(2.1b) reads as follows: Find $(u^n, \varphi^n) \in V \times W$ such that

$$(g(\varphi^n)\mathbb{C}\mathbf{e}(u^n), \mathbf{e}(v)) = (b^n, v), \qquad \forall v \in V, \tag{2.5a}$$

$$\begin{aligned} &G_c\varepsilon(\nabla\varphi^n, \nabla\psi) - \frac{G_c}{\varepsilon}(1-\varphi^n, \psi) + (1-\kappa)(\varphi^n|\mathbb{C}\mathbf{e}(u^n)|^2, \psi) \\ &\quad + ([\Xi + \gamma(\varphi^n - \varphi^{n-1})]^+, \psi) = 0, \qquad \forall \psi \in W, \end{aligned} \tag{2.5b}$$

where $b^n := b(\cdot, t^n)$. The last term in the phase field equation (2.5b) is a penalization to enforce the irreversibility condition, thus transforming the variational inequality (2.1b) into a variational equality, with penalization parameter $\gamma > 0$, and where $\Xi \in L^2(B)$ is given (in practice Ξ will be obtained by iteration, cf. Section 4). Note that we also used the notation $[x]^+ := \max(x, 0)$. From here on, we shall refer to (2.5a) as the *mechanics subproblem*, and to (2.5b) as the *phase field subproblem*. Regarding the degradation function g, it is easily seen to satisfy the following Lipschitz condition:

$$\|g(\psi) - g(\eta)\| \leq 2(1-\kappa)\|\psi - \eta\|, \quad \forall \psi, \eta \in W. \tag{2.6}$$

The time-discrete system (2.5a)-(2.5b) was analyzed in [36], and there it was shown that at least one global minimizer $(u^n, \varphi^n) \in V \times W$ exists, provided $b^n \in (L^2(B))^d$, for each n. We mention also that the analysis of a pressurized phase field brittle fracture model can be found in [34, 35].

3 Iterative scheme

In this section we introduce the iterative staggered solution procedure for the fully discrete formulation of (2.5a)-(2.5b). To this end, let $\mathcal{T}_h$ be a simplicial mesh of B, such that for any two distinct elements of $\mathcal{T}_h$ their intersection is either an empty set or their common vertex or edge. We denote by h the largest diameter of all the elements in $\mathcal{T}_h$, i.e., $h := \max_{K\in\mathcal{T}_h} \text{diam}(K)$, and let $V_h \times W_h \subset V \times W$ be appropriate (conforming) discrete spaces. We continue now with the same notation for the variables and test functions as before (omitting the usual h-subscript), since we will from here on mostly deal with the discrete solutions.

For each n, the iterative algorithm we propose defines a sequence $\{u^{n,i}, \varphi^{n,i}\}$, for $i \geq 0$, initialized by $\{u^{n-1}, \varphi^{n-1}\}$. The iteration is then done in two steps: First, the mechanics subproblem is solved, with the degradation function held constant. Then, the phase field subproblem is solved, with the elastic energy held constant. Note that there are also artificial stabilizing terms which are held constant during solving of the subproblems. Introducing the stabilization parameters $L_u, L_\varphi > 0$ (to be determined later), the iterative algorithm

reads as follows:

- **Step 1**: Given $(u^{n,i-1}, \varphi^{n,i-1}, b^n)$ find $u^{n,i}$ such that
$$a_u(u^{n,i}, v) := L_u(u^{n,i} - u^{n,i-1}, v) + \big(g(\varphi^{n,i-1})\mathbb{C}\mathbf{e}(u^{n,i}), \mathbf{e}(v)\big) = (b^n, v), \qquad \forall v \in V_h. \tag{3.1a}$$
- **Step 2**: Given $(\varphi^{n,i-1}, u^{n,i}, \varphi^{n-1})$ find $\varphi^{n,i}$ such that
$$\begin{aligned} a_\varphi(\varphi^{n,i}, \psi) &:= L_\varphi(\varphi^{n,i} - \varphi^{n,i-1}, \psi) + G_c\varepsilon(\nabla\varphi^{n,i}, \nabla\psi) - \frac{G_c}{\varepsilon}(1-\varphi^{n,i}, \psi) \\ &\quad + (1-\kappa)(\varphi^{n,i}|\mathbb{C}\mathbf{e}(u^{n,i})|^2, \psi) + (\eta^i(\Xi + \gamma(\varphi^{n,i} - \varphi^{n-1})), \psi) = 0, \qquad \forall \psi \in W_h, \end{aligned} \tag{3.1b}$$

where, in order to avoid the $[\cdot]^+$-bracket, we also introduced the function $\eta^i \in L^\infty(B)$ defined for a.e. $x \in B$ by

$$\eta^i(x) = \begin{cases} 1, & \text{if } \Xi(x) + \gamma(\varphi^{n,i}(x) - \varphi^{n-1}(x)) \geq 0, \\ 0, & \text{if } \Xi(x) + \gamma(\varphi^{n,i}(x) - \varphi^{n-1}(x)) < 0. \end{cases} \tag{3.2}$$

3.1 Convergence analysis

We now proceed to analyze the convergence of the scheme (3.1a)-(3.1b). Our aim is to show a contraction of successive difference functions in energy norms, which implies convergence by the Banach Fixed Point Theorem (see e.g. [14]). To this end we define the following difference functions

$$e_u^i := u^{n,i} - u^n, \tag{3.3}$$
$$e_\varphi^i := \varphi^{n,i} - \varphi^n, \tag{3.4}$$

where $\{u^n, \varphi^n\}$ denotes the (exact) solutions to (2.1a)-(2.1b) at time t^n. Using the symmetry properties of $\mathbb{C}$, the following set of difference equations are then obtained by subtracting (3.1a)-(3.1b) solved by $\{u^n, \varphi^n\}$ from the same equations solved by the iterate solutions:

$$L_u(e_u^i - e_u^{i-1}, v) + (g(\varphi^n)\mathbb{C}\mathbf{e}(e_u^i), \mathbf{e}(v)) + ((g(\varphi^{n,i-1}) - g(\varphi^n))\mathbb{C}\mathbf{e}(u^{n,i}), \mathbf{e}(v)) = 0, \qquad \forall v \in V_h. \tag{3.5a}$$

$$\begin{aligned} &L_\varphi(e_\varphi^i - e_\varphi^{i-1}, \psi) + G_c\varepsilon(\nabla e_\varphi^i, \nabla\psi) + \frac{G_c}{\varepsilon}(e_\varphi^i, \psi) + \gamma(\eta^i e_\varphi^i, \psi) + (1-\kappa)(e_\varphi^i|\mathbb{C}\mathbf{e}(u^{n,i})|^2, \psi) \\ &\qquad + (1-\kappa)\left(\varphi^n\mathbb{C}\mathbf{e}(e_u^i) : \mathbf{e}(u^{n,i} + u^n), \psi\right) = 0, \qquad \forall \psi \in W_h. \end{aligned} \tag{3.5b}$$

Furthermore, we introduce the following assumption related to the elastic mechanical strain.

Assumption 1 (Boundedness of elastic strain)**.** *We assume there exists a constant $M > 0$ such that*

$$\operatorname*{ess\,sup}_{x \in B} |\mathbf{e}(u^n(x))| \leq M, \quad \forall n. \tag{3.6}$$

Moreover, we assume that M is large enough such that the above bound holds also for the iterate elastic strain, i.e.,

$$\operatorname*{ess\,sup}_{x \in B} |\mathbf{e}(u^{n,i}(x))| \leq M, \quad \forall (n,i). \tag{3.7}$$

Note that M is nothing else than an upper bound for the elastic strain in the system for the converged solution, which is arguably finite for any reasonable problem. Note also that with sufficient regularity of the domain, coefficients, source terms, and initial data, the above assumption is satisfied, i.e., the problem (2.5a)-(2.5b) admits a solution $u^n \in (W^{1,\infty}(B))^d$, thus implying the existence of M. Alternatively to introducing the constant M, we could introduce instead a so-called 'cut-off operator' in the iterate equations (3.1a)-(3.1b), as seen in e.g. [40, 41]. Note that in all numerical tests to be done in the next sections, we provide figures validating the second part of this assumption (cf. Section 5.4). With the above definitions, we state our main theoretical result.

Theorem 3.1 (Convergence of the scheme)**.** *The scheme* (3.1b)–(3.1a) *defines a contraction satisfying*

$$\begin{aligned} &\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon} + \frac{G_c\varepsilon}{c_P} - 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^i\|^2 + \left(\frac{L_u}{2} + \frac{\kappa\lambda_{\min}^2}{2c_P c_K}\right)\|e_u^i\|^2 \\ &\qquad \leq \left(\frac{L_\varphi}{2} + 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2, \end{aligned} \tag{3.8}$$

if $L_u, L_\varphi > 0$, and if the model parameter $\varepsilon > 0$ is sufficiently large such that

$$\varepsilon^2 - 16\xi\frac{(1-\kappa)^2}{\kappa}\frac{c_P}{G_c}\varepsilon + c_P > 0, \tag{3.9}$$

where $\xi := (M\lambda_{\max}/\lambda_{\min})^2 > 0$, and where $c_P, c_K > 0$ are the Poincaré and Korn constants, respectively, depending only on the domain B and spatial dimension d.

Proof. We begin by taking $v = e_u^i$ and $\psi = e_\varphi^i$ in (3.5a) and (3.5b), respectively, add the resulting equations together and obtain

$$\begin{aligned}
&\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon}\right)\|e_\varphi^i\|^2 + \frac{L_\varphi}{2}\|e_\varphi^i - e_\varphi^{i-1}\|^2 + G_c\varepsilon\|\nabla e_\varphi^i\|^2 + \gamma(\eta^i e_\varphi^i, e_\varphi^i)\\
&\quad + (1-\kappa)(e_\varphi^i|\mathbb{C}\mathbf{e}(u^{n,i})|^2, e_\varphi^i) + \frac{L_u}{2}\|e_u^i\|^2 + \frac{L_u}{2}\|e_u^i - e_u^{i-1}\|^2 + (g(\varphi^n)\mathbb{C}\mathbf{e}(e_u^i), \mathbf{e}(e_u^i))\\
&\quad = \frac{L_\varphi}{2}\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2 - (1-\kappa)(\varphi^n\mathbb{C}\mathbf{e}(e_u^i) : \mathbf{e}(u^{n,i} + u^n), e_\varphi^i)\\
&\qquad - ((g(\varphi^{n,i-1}) - g(\varphi^n))\mathbb{C}\mathbf{e}(u^{n,i}), \mathbf{e}(e_u^i)),
\end{aligned} \tag{3.10}$$

where we used the following inner product identity

$$2(x-y,x) = \|x\|^2 + \|x-y\|^2 - \|y\|^2. \tag{3.11}$$

Discarding some non-negative terms from the left hand side of (3.10), using the fact that $\operatorname{ess\,sup}_{x\in B}\varphi^n(x) \le 1$, in addition to the Lipschitz property of the degradation function g (2.6), yields

$$\begin{aligned}
&\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon}\right)\|e_\varphi^i\|^2 + G_c\varepsilon\|\nabla e_\varphi^i\|^2 + \frac{L_u}{2}\|e_u^i\|^2 + \kappa(\mathbb{C}\mathbf{e}(e_u^i), \mathbf{e}(e_u^i))\\
&\quad \le \frac{L_\varphi}{2}\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2 + (1-\kappa)\int_B |\mathbb{C}\mathbf{e}(e_u^i) : \mathbf{e}(u^{n,i} + u^n)e_\varphi^i|\mathrm{d}x\\
&\qquad + \int_B |(g(\varphi^{n,i-1}) - g(\varphi^n))\mathbb{C}\mathbf{e}(u^{n,i}) : \mathbf{e}(e_u^i)|\mathrm{d}x\\
&\qquad\quad \le \frac{L_\varphi}{2}\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2 + 2(1-\kappa)\lambda_{\max}M\Big(\|e_\varphi^i\| + \|e_\varphi^{i-1}\|\Big)\|\mathbf{e}(e_u^i)\|,
\end{aligned} \tag{3.12}$$

where we also invoked the Assumption 1 in the last line, and applied the Cauchy-Schwarz inequality. Using the Young inequality, the properties of elastic tensor (2.2), and rearranging, leads to

$$\begin{aligned}
&\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon} - 2(1-\kappa)\lambda_{\max}M\frac{1}{2\delta_1}\right)\|e_\varphi^i\|^2 + G_c\varepsilon\|\nabla e_\varphi^i\|^2\\
&\quad + \frac{L_u}{2}\|e_u^i\|^2 + \Big(\kappa\lambda_{\min}^2 - 2(1-\kappa)\lambda_{\max}M(\delta_1+\delta_2)\Big)\|\mathbf{e}(e_u^i)\|^2\\
&\qquad \le \left(\frac{L_\varphi}{2} + 2(1-\kappa)\lambda_{\max}M\frac{1}{2\delta_2}\right)\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2,
\end{aligned} \tag{3.13}$$

for some constants $\delta_1, \delta_2 > 0$. Choosing $\delta_1 = \delta_2 = \kappa\lambda_{\min}^2/8(1-\kappa)\lambda_{\max}M$ yields (3.13) as

$$\begin{aligned}
&\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon} - 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^i\|^2 + G_c\varepsilon\|\nabla e_\varphi^i\|^2 + \frac{L_u}{2}\|e_u^i\|^2 + \frac{\kappa\lambda_{\min}^2}{2}\|\mathbf{e}(e_u^i)\|^2\\
&\quad \le \left(\frac{L_\varphi}{2} + 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2.
\end{aligned} \tag{3.14}$$

Next, by applying the Poincaré inequality on $\|e_\varphi^i\|$, and by applying successively the Poincaré and Korn inequalities on $\|e_u^i\|$, we obtain

$$\|e_\varphi^i\|^2 \le c_P\|\nabla e_\varphi^i\|^2 \quad \text{and} \quad \|e_u^i\|^2 \le c_P c_K\|\mathbf{e}(e_u^i)\|^2, \tag{3.15}$$

where c_P, c_K are the (squares of the) Poincaré and Korn constants, respectively (depending only on the domain B and spatial dimension d). Finally, employing these bounds on the left hand side of (3.14) yields

$$\begin{aligned}
&\left(\frac{L_\varphi}{2} + \frac{G_c}{\varepsilon} + \frac{G_c\varepsilon}{c_P} - 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^i\|^2 + \left(\frac{L_u}{2} + \frac{\kappa\lambda_{\min}^2}{2c_P c_K}\right)\|e_u^i\|^2\\
&\quad \le \left(\frac{L_\varphi}{2} + 8\xi\frac{(1-\kappa)^2}{\kappa}\right)\|e_\varphi^{i-1}\|^2 + \frac{L_u}{2}\|e_u^{i-1}\|^2.
\end{aligned} \tag{3.16}$$

Thus, for (3.16) to be a contraction estimate, ε must satisfy the following second order inequality

$$P(\varepsilon) := \varepsilon^2 - 16\xi\frac{c_P}{G_c}\frac{(1-\kappa)^2}{\kappa}\varepsilon + c_P > 0. \tag{3.17}$$

Setting the left hand side of (3.17) equal to zero yields a second order polynomial, the discriminant of which must satisfy one of the following three statements:

1. If
$$64\xi^2\frac{(1-\kappa)^4}{\kappa^2} > \frac{G_c^2}{c_P},$$
then $P(\varepsilon) = 0$ has two distinct positive real roots $\varepsilon_1, \varepsilon_2 > 0$, in which case (3.16) is a contraction for $\varepsilon \in (0, \varepsilon_1) \cup (\varepsilon_2, \infty)$.

2. If
$$64\xi^2\frac{(1-\kappa)^4}{\kappa^2} = \frac{G_c^2}{c_P},$$
then $P(\varepsilon) = 0$ has one positive real root, $\varepsilon_0 > 0$, of multiplicity two, in which case (3.16) is a contraction for all $\varepsilon \neq \varepsilon_0, \varepsilon > 0$.

3. If
$$64\xi^2\frac{(1-\kappa)^4}{\kappa^2} < \frac{G_c^2}{c_P},$$
then $P(\varepsilon) = 0$ has two complex roots, in which case (3.16) is a contraction for all $\varepsilon > 0$.

□

Remark 3.1 (Convergence rate)**.** *According to the above proof, if the scheme is not converging for a given value of ε, then a larger or a smaller value may be chosen to rectify the situation. However, since crack surfaces become singular as $\varepsilon \to 0$ (thus necessitating finer meshing, i.e., $h \to 0$), we choose to state Theorem 3.1 with the condition that ε be large enough. We note also that due to some unknown constants in the convergence rate* (3.8)*, it is not known whether this rate is optimal. Furthermore, working with a large ε is substantiated by the theory of phase field fracture being based on Γ convergence [2, 3]. Applying this to phase field fracture was first done in [10]. Specifically, the setting is suitable when $h = o(\varepsilon)$; namely when ε is sufficiently large.*

4 Algorithm

In practice, we apply the stabilizations and penalizations proposed in the previous sections as outlined below. It is well-known (e.g., [37]) that the choice of γ is critical. If γ is too low, crack irreversibility will not be enforced. On the other hand, if γ is too large, the linear equation system is ill-conditioned and influences the performance of the nonlinear solver. For this reason, γ is updated in at each iteration step. Better, in terms of robustness, is the augmention in such an iteration by an additional L^2 function Ξ, yielding a so-called *augmented Lagrangian iteration* going back to [19, 22]. For phase-field fracture this idea was first applied in [42]. Thus, combining the staggered iteration for the solid and phase-field systems with the update of the penalization parameter Ξ yields the following algorithm:

Algorithm 1. *At the loading step t^n.*
Choose initial Ξ^0. Set $\gamma > 0$.
repeat
Iterate on i (augmented Lagrangian loop)
Solve two-field problem, namely
Solve elasticity in Problem (3.1a)
Solve the nonlinear phase-field in Problem (3.1b)
Update
$$\Xi^{i+1} = [\Xi^i + \gamma(\varphi^{n,i+1} - \varphi^{n-1})]^+$$
until
$$\max(\|a_u(u^{n,i}, v_k) - (b^n, v_k)\|, \|a_\varphi(\varphi^{n,i}, \psi_l)\|) \leq \mathrm{TOL}, \quad \text{for } k = 1, \ldots, dim(V_h), \quad l = 1, \ldots, dim(W_h). \tag{4.1}$$
Set: $(u^n, \varphi^n) := (u^{n,i}, \varphi^{n,i})$.
Increment $t^n \to t^{n+1}$.

For the stabilization parameters L_u, L_φ, we have the following requirements (somewhat similar to γ): If the stabilization is too small, the stabilization effects vanish. If the stabilization is too large, we revert to an unacceptably slow convergence, and potentially, may converge to a solution corresponding to an undesirable local minimum of the original problem. In order to deal with these issues, we employ here a simple, yet effective strategy: We draw $L := L_u = L_\varphi$ from a range of suitable values and compare the results, i.e., $L \in \{1.0e{-}6, 1.0e{-}3, 1.0e{-}2, 1.0e{-}1\}$. Moreover, we include also the configurations $L_u = 0$, $L_\varphi > 0$ and $L_u = L_\varphi = 0$ in all the numerical tests to be done in the following.

Remark 4.1. *In this paper we use* $\mathrm{TOL} = 10^{-6}$.

4.1 Nonlinear solution, linear subsolvers and programming code

Both subproblems (phase field and mechanics) may be nonlinear. In our theory presented above, we assumed a standard elasticity tensor. However, the model (3.1a)–(3.1b) is too simple for most mechanical applications. More realistic phase-field fracture applications require a splitting of the stress tensor (based on an energy split) in order to account for fracture development only under tension, but not under compressive forces. Consequently, we follow here [31] and split $\boldsymbol{\sigma}$ into tensile $\boldsymbol{\sigma}^+$ and compressive parts $\boldsymbol{\sigma}^-$:

$$\begin{aligned}\boldsymbol{\sigma}^+ &:= 2\mu_s \mathbf{e}^+ + \lambda_s < \mathrm{tr}(\mathbf{e}) > \mathbf{I},\\ \boldsymbol{\sigma}^- &:= 2\mu_s(\mathbf{e} - \mathbf{e}^+) + \lambda_s(\mathrm{tr}(\mathbf{e}) - < \mathrm{tr}(\mathbf{e}) >)\mathbf{I},\end{aligned}$$

and

$$\mathbf{e}^+ = \mathbf{P}\boldsymbol{\Lambda}^+\mathbf{P}^T,$$

where the elasticity tensor $\mathbb{C}$ has been replaced by the Lamé parameters, μ_s and λ_s. Moreover, $\mathbf{I}$ is the $d \times d$ identity matrix, and $< \cdot >$ is the positive part of a function. In particular, for $d = 2$, we have

$$\boldsymbol{\Lambda}^+ := \boldsymbol{\Lambda}^+(u) := \begin{pmatrix} < \lambda_1(u) > & 0 \\ 0 & < \lambda_2(u) > \end{pmatrix},$$

where $\lambda_1(u)$ and $\lambda_2(u)$ are the eigenvalues of the strain tensor $\mathbf{e} := \mathbf{e}(u)$, and $v_1(u)$ and $v_2(u)$ the corresponding (normalized) eigenvectors. Finally, the matrix $\mathbf{P}$ is defined as $\mathbf{P} := \mathbf{P}(u) := [v_1|v_2]$; namely, it consists of the column vectors $v_i,\ i = 1, 2$. We notice that another frequently employed stress-splitting law was proposed in [4].

The modified scheme reads:

- **Step 1**: given $(u^{n,i-1}, \varphi^{n,i-1}, b^n)$ find $u^{n,i}$ such that
$$L_u(u^{n,i} - u^{n,i-1}, v) + \big(g(\varphi^{n,i-1})\boldsymbol{\sigma}^+(u^{n,i}), \mathbf{e}(v)\big) + \big(\boldsymbol{\sigma}^-(u^{n,i}), \mathbf{e}(v)\big) = (b^n, v), \qquad \forall v \in V_h, \tag{4.2a}$$
- **Step 2**: given $(\varphi^{n,i-1}, u^{n,i}, \varphi^{n-1})$ find $\varphi^{n,i}$ such that
$$\begin{aligned}L_\varphi(\varphi^{n,i} - \varphi^{n,i-1}, \psi) + G_c\varepsilon(\nabla\varphi^{n,i}, \nabla\psi) - \frac{G_c}{\varepsilon}(1 - \varphi^{n,i}, \psi)\\ + (1 - \kappa)(\varphi^{n,i}\boldsymbol{\sigma}^+(u^{n,i}) : \mathbf{e}(u^{n,i}), \psi) + (\eta^i(\Xi + \gamma(\varphi^{n,i} - \varphi^{n-1})), \psi) = 0, \qquad \forall \psi \in W_h.\end{aligned} \tag{4.2b}$$

These modifications render the displacement system (4.2a) nonlinear, for which we use a Newton-type solver. The phase field equation is also nonlinear due to the penalization term and the stress splitting. Our version of Newton's method is based on a residual-based monotonicity criterion (e.g., [17]) outlined in [45][Section 3.2]. Inside Newton's method, the linear subsystems are solved with a direct solver; namely UMFPACK [16]. All numerical tests presented in Section 5 are implemented in the open-source finite element library deal.II [5, 6]. Specifically, the code is based on a simple adaptation of the multiphysics template [43] in which specifically the previously mentioned Newton solver is implemented.

5 Numerical experiments

In this section, we present several numerical tests to substantiate our algorithmic developments. The goals of all three numerical examples are comparisons between an unlimited number of staggered iterations (although bounded by 500) denoted by 'L', and a low, fixed number, denoted by 'LFI', where we use 30 (Ex. 1 and Ex. 2), and 20 (Ex. 3) staggered iterations, respectively. These comparisons are performed in terms of the number of iterations and the correctness of the solutions in terms of the so-called *load-displacement curve*, measuring the stresses of the top boundary versus the number of loading steps.

5.1 Single edge notched tension test

This test was applied for instance in [31]. The configuration is displayed in Figure 1. We use the system (3.1a)-(3.1b). Specifically, we study our proposed iterative schemes on different mesh levels, denoted as refinement (Ref.) levels 4, 5, 6 (uniformly refined), with 1024 elements (2210 Dofs for the displacements, 1 105 Dofs for the phase-field, $h = 0.044$), 4 096 elements (8 514 Dofs for the displacements, 4257 Dofs for the phase-field, $h = 0.022$), and 16 384 elements (33 410 Dofs for the displacements, 16 705 Dofs for the phase-field, $h = 0.011$).

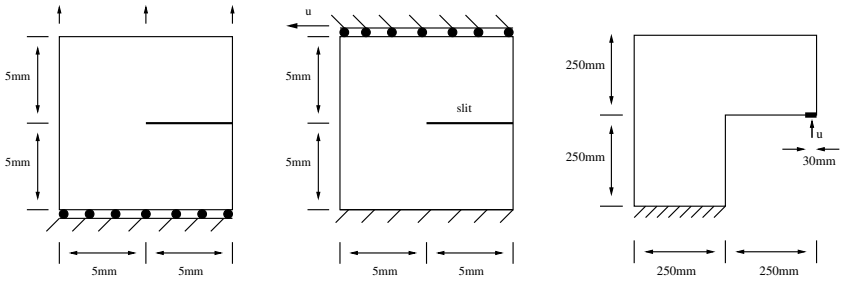


Figure 1: Examples 1,2,3: Configurations. Left: single edge notched tension test. In detail, the boundary conditions are: $u_y = 0\,\text{mm}$ (homogeneous Dirichlet) and traction free (homogeneous Neumann conditions) in x-direction on the bottom. On the top boundary Γ_{top}, we prescribe $u_x = 0\,\text{mm}$ and u_y as provided in (5.3). All other boundaries including the slit are traction free (homogeneous Neumann conditions). Single edge notched shear test (middle) and L-shaped panel test (right). We prescribe the following conditions: On the left and right boundaries, $u_y = 0\,\text{mm}$ and traction-free in x-direction. On the bottom part, we use $u_x = u_y = 0\,\text{mm}$ and on Γ_{top}, we prescribe $u_y = 0\,\text{mm}$ and u_x as stated in (5.3). Finally, the lower part of the slit is fixed in y-direction, i.e., $u_y = 0\,\text{mm}$. For the L-shaped panel test (at right), the lower left boundary is fixed: $u_x = u_y = 0\,\text{mm}$. A displacement condition for u_y is prescribed by (5.4) in the right corner on a section Γ_u that has $30\,\text{mm}$ length.

Specifically, we use $\mu_s = 80.77\,\text{kN/mm}^2$, $\lambda_s = 121.15\,\text{kN/mm}^2$, and $G_c = 2.7\,\text{N/mm}$. The crack growth is driven by a non-homogeneous Dirichlet condition for the displacement field on Γ_{top}, the top boundary of B . We increase the displacement on Γ_{top} over time, namely we apply non-homogeneous Dirichlet conditions:

$$u_y = t\bar{u}, \quad \bar{u} = 1\,\text{mm/s}, \tag{5.1}$$

where t denotes the current loading time. Furthermore, we set $\kappa = 10^{-10}$ [mm] and $\varepsilon = 2h$ [mm]. We evaluate the surface load vector on the Γ_{top} as

$$\tau = (F_x, F_y) := \int_{\Gamma_{\text{top}}} \boldsymbol{\sigma}(u)\nu \, \mathrm{d}s, \tag{5.2}$$

with normal vector ν, and we are particularly interested in F_y for Example 1 and F_x for Example 2 (Section 5.2). Graphical solutions are displayed in the Figures 2 and 3 showing the phase-field variable and the discontinuous displacement field. Our findings of using different stabilization parameters L are compared in the Figures 4, 5, 6, 7, and 8. Different mesh refinement studies are shown in the Figures 7 and 8. Here, the number of staggered iterations does not increase with finer mesh levels, which shows the robustness of our proposed methodology.

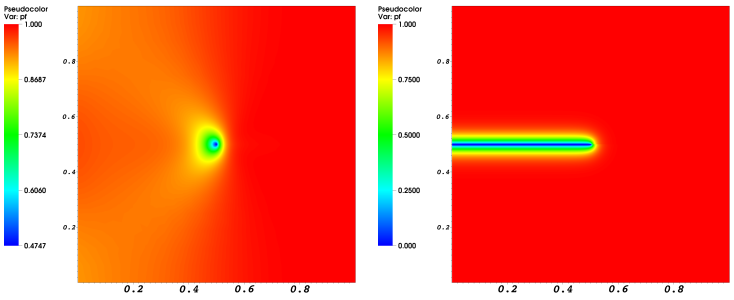


Figure 2: Example 1: Single edge notched tension test: crack path at loading step 59 (left) and 60 (right). We see brutal crack growth in which the domain is cracked within one loading step.

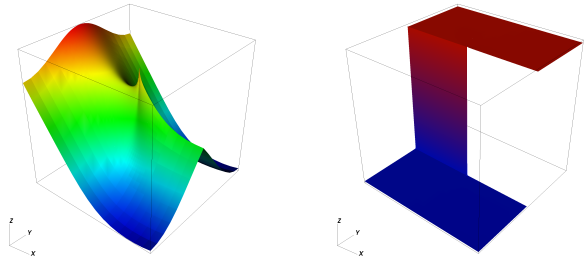


Figure 3: Example 1: Single edge notched tension test: 3D plot of the displacement variable u_x at the loading steps 59 and 60. At right, the domain is totally fractured. In particular, we see the initial crack build in the geometry in the right part where the domain has a true discontinuity. In the left part, the domain is cracked using the phase-field variable. Here, the displacement variable is still continuous since we are using C^0 finite elements for the spatial discretization.

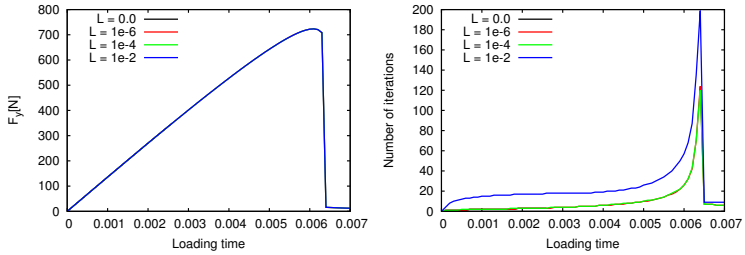


Figure 4: Example 1: Comparison of different L. At left, the stresses are shown. At right, the number of staggered iterations is displayed.

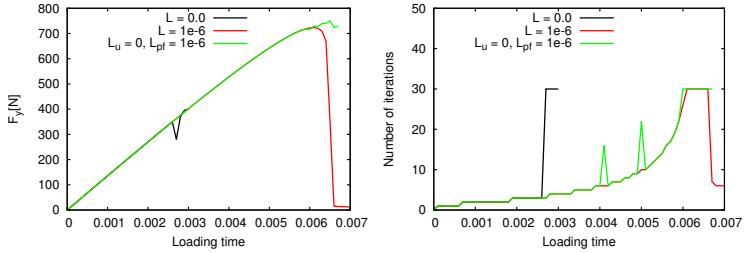


Figure 5: Example 1: Comparison of different L. In this example, possibly due to brutal crack growth, stabilizing only phase field subproblem does not work.

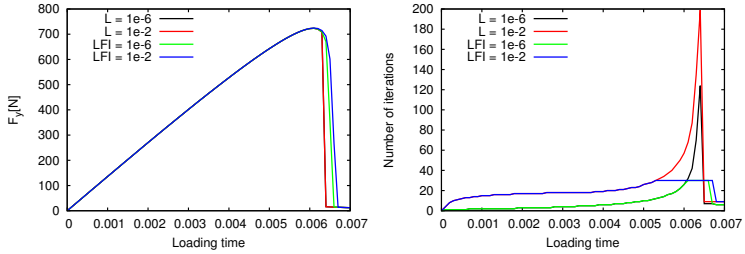


Figure 6: Example 1: Comparison of different L for an open number of iterations and a fixed number of iterations (LFI) with a maximum of 30 iterations. At left, the stresses are shown. At right, the number of staggered iterations is displayed.

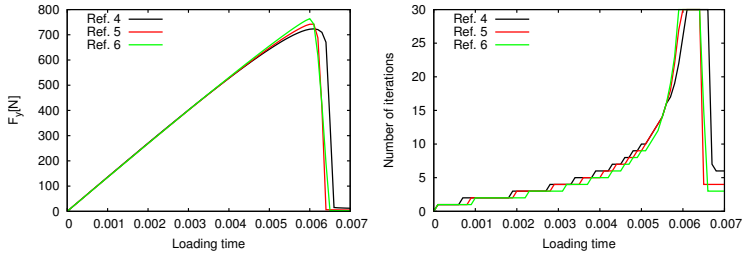


Figure 7: Example 1: Using $L = 1e-6$ comparing different mesh refinement levels $4, 5, 6$. At left, the stresses are shown. At right, the number of staggered iterations is displayed.

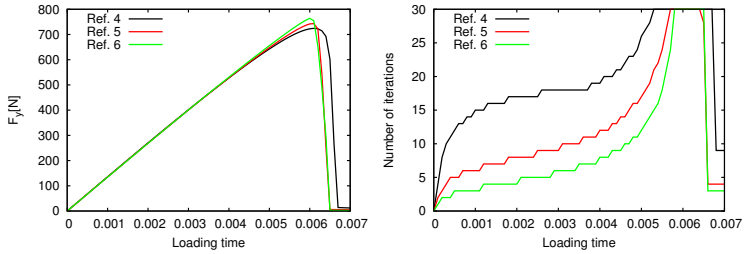


Figure 8: Example 1: Using $L = 1e-2$ comparing different mesh refinement levels $4, 5, 6$. At left, the stresses are shown. At right, the number of staggered iterations is displayed.

5.2 Single edge notched shear test

The configuration of this second setting is very similar to Example 1 and was first proposed in a phase-field context in [31]. We now use the model with strain-energy split (4.2a)–(4.2b). The parameters and the geometry (see Figure 1) are the same as in the previous test case. The boundary condition is changed from tensile forces to a shear condition (see also again Figure 1):

$$u_x = t\bar{u}, \quad \bar{u} = 1\,\mathrm{mm/s}, \tag{5.3}$$

As quanitity of interest we evaluate F_x in (5.2). Our findings are shown in the Figures 9, 10, 11, 12, 13, 14, and 15. The major difference to Example 1 is that the scheme is converging even with $L_u = 0$, as computationally justified in Figure 11. As in Example 1, the load-displacement curves are close to the published literature and, again, the proposed L scheme is robust under mesh refinement (see Figures 12 - 15).

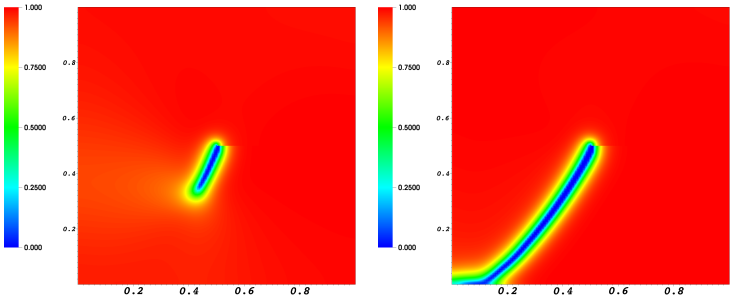


Figure 9: Example 2: Single edge notched shear test: Crack path at loading step 110 (left) and 135 (right).

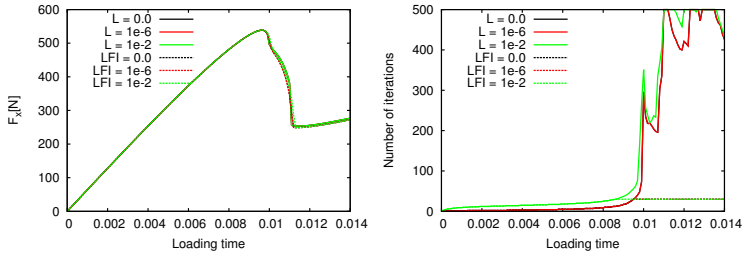


Figure 10: Example 2: Comparison of different L with an open number of staggered iterations (fixed by 500) and a fixed number (LFI) with 30 iterations per loading step. At left, the load-displacement curves displaying the evolution of F_x versus the loading time. At right, the number of iterations is displayed.

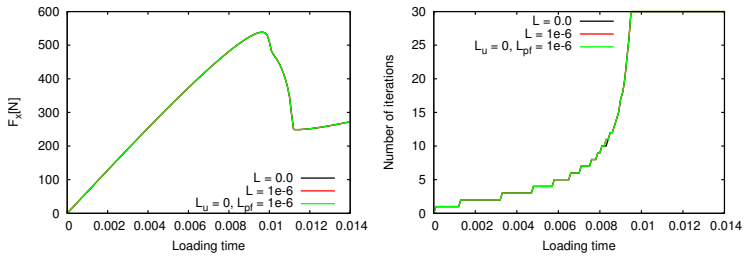


Figure 11: Example 2: Comparison of different L. Observe that stabilizing the mechanics subproblem in this example has no or little effect. At left, the load-displacement curves displaying the evolution of F_x versus u_y are shown. At right, the number of staggered iterations is displayed.

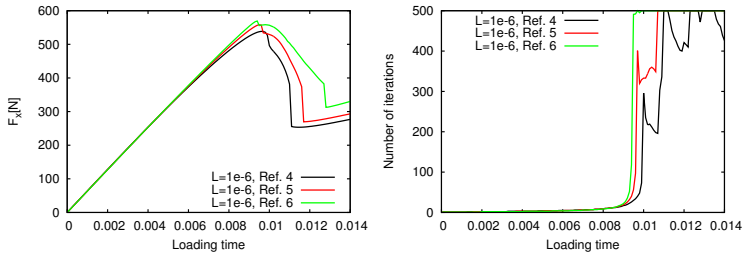


Figure 12: Example 2: Using $L = 1e-6$, comparing different mesh refinement levels $4, 5, 6$. At left, the load-displacement curves displaying the evolution of F_x versus the loading time. At right, the number of iterations is displayed.

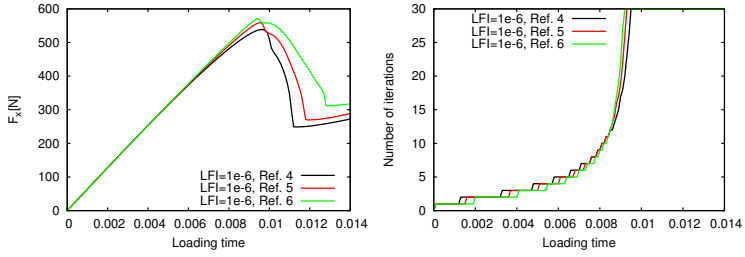


Figure 13: Example 2: Using $L = 1e-6$ and fixing the number of iterations by 30, we compare different mesh refinement levels $4, 5, 6$. At left, the load-displacement curves displaying the evolution of F_x versus the loading time. At right, the number of iterations is displayed.

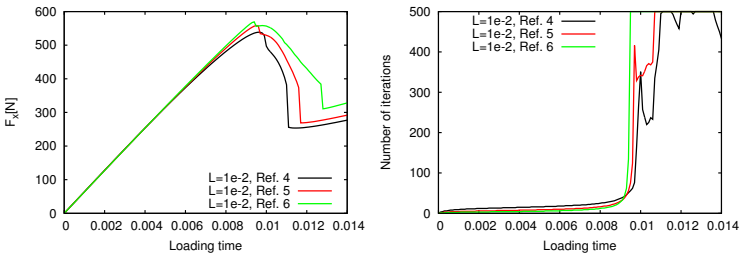


Figure 14: Example 2: Using $L = 1e-2$, we compare different mesh refinement levels $4, 5, 6$. At left, the load-displacement curves displaying the evolution of F_x versus the loading time. At right, the number of iterations is displayed.

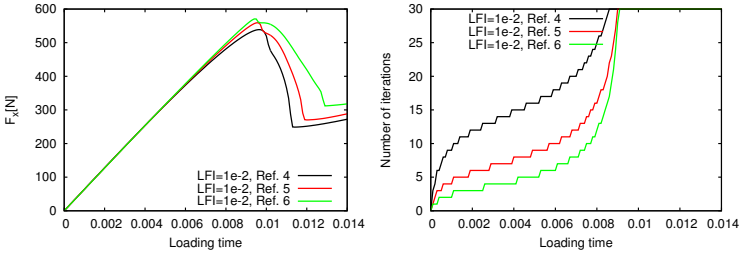


Figure 15: Example 2: Using $L = 1e-2$ and fixing the number of iterations by 30, we compare different mesh refinement levels $4, 5, 6$. At left, the load-displacement curves displaying the evolution of F_x versus the loading time. At right, the number of iterations is displayed.

The results are very comparable to the published literature. In particular, it is nowadays known that the proposed Miehe et al. stress splitting does not release all stresses once the specimen is broken (see [1]) and it is also known that we do not see convergence of the curves when both h and ε are refined (see [25]).

5.3 L-shaped panel

For the configuration of this third example we refer to [1, 29, 44], which are based on an experimental setup [46]. We use again the model with strain-energy split; namely (4.2a)-(4.2b). Moreover, in this test a carefully imposed irreversibility constraint is important since the specimen is pushed, pulled, and again pushed (see Figure 16 for the loading history on the small boundary part Γ_u). In the pulling phase the fracture vanishes if the penalization is not strong enough.

The geometry and boundary conditions are displayed in Figure 1. In contrast to the previous examples, no initial crack prescribed. The initial mesh is $1, 2$ and 3 times uniformly refined, leading to $300, 1200, 4800$ mesh elements, with $h = 29.1548\,\text{mm}$, $14.577\,\text{mm}$, $7.289\,\text{mm}$, respectively.

We increase the displacement $u_D := u_y = u_y(t)$ on $\Gamma_u := \{(x, y) \in B | \, 470\,\text{mm} \leq x \leq 500\,\text{mm},\ y = 250\,\text{mm}\}$ over time, where Γ_u is a section of $30\,\text{mm}$ length on the right corner of the specimen. We apply a loading-dependent, non-homogeneous Dirichlet condition (see also Figure 16):

$$\begin{aligned} u_y &= t \cdot \bar{u}, & \bar{u} &= 1\,\text{mm/s}, \quad 0.0\,\text{s} \leq t < 0.3\,\text{s}, \\ u_y &= (0.6 - t) \cdot \bar{u}, & \bar{u} &= 1\,\text{mm/s}, \quad 0.3\,\text{s} \leq t < 0.8\,\text{s}, \\ u_y &= (-1 + t) \cdot \bar{u}, & \bar{u} &= 1\,\text{mm/s}, \quad 0.8\,\text{s} \leq t \leq 2.0\,\text{s}, \end{aligned} \tag{5.4}$$

where t denotes the total loading time. Due to this cyclic loading the total displacement at the end time $T = 2\,\text{s}$ is $1\,\text{mm}$.

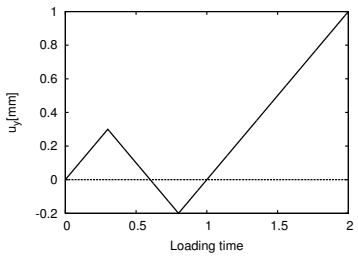


Figure 16: Example 3: Loading history on Γ_u for the L-shaped panel test.

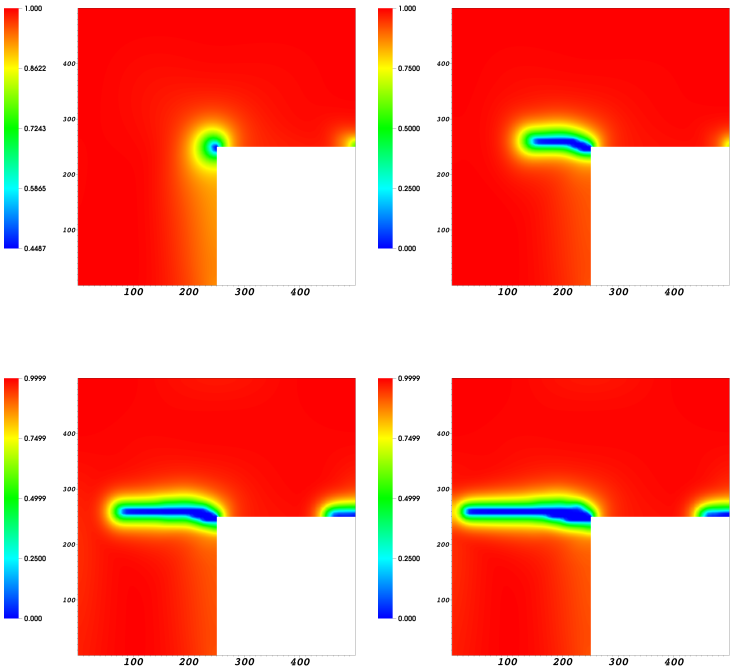


Figure 17: Example 3: crack path of the L-shaped panel test at the loading steps $220, 300, 1450, 2000$.

We use $\mu_s = 10.95\,\mathrm{kN/mm^2}$, $\lambda_s = 6.16\,\mathrm{kN/mm^2}$, and $G_c = 8.9 \times 10^{-5}\,\mathrm{kN/mm}$. The time (loading) step size is $\delta t = 10^{-3}\,\mathrm{s}$. Furthermore, we set $k = 10^{-10}h[\mathrm{mm}]$ and $\varepsilon = 2h$. As before, we observe the number of Newton iterations and we evaluate the surface load vector

$$\tau = (F_x, F_y) := \int_{\Gamma_u} \boldsymbol{\sigma}(u)\nu \, \mathrm{d}s,$$

with normal vector ν, and now we are particularly interested in F_y. The crack path at the chosen time step snapshots in Figure 17 corresponds to the published literature [44, 29, 1]. The load-displacement curves and the number of iterations for different L and corresponding mesh refinement studies are displayed in the Figures 18, 19, 20, 21 and 22.

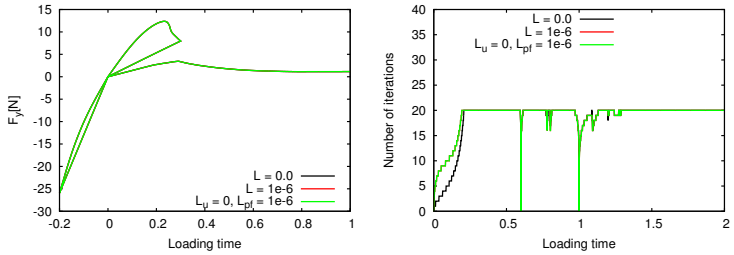


Figure 18: Example 3: Comparison of different L. Observe that stabilizing the mechanics subproblem has no effect in this example. At left, the load-displacement curves displaying the evolution of F_y versus u_y are shown. At right, the number of staggered iterations is displayed.

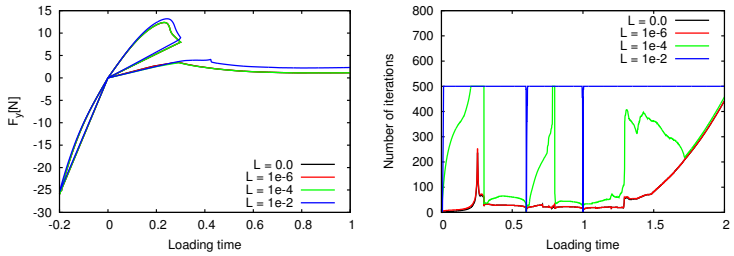


Figure 19: Example 3: Comparison of different L with an open number of staggered iterations (fixed by 500 though). At left, the load-displacement curves displaying the evolution of F_y versus u_y are shown. At right, the number of staggered iterations is displayed.

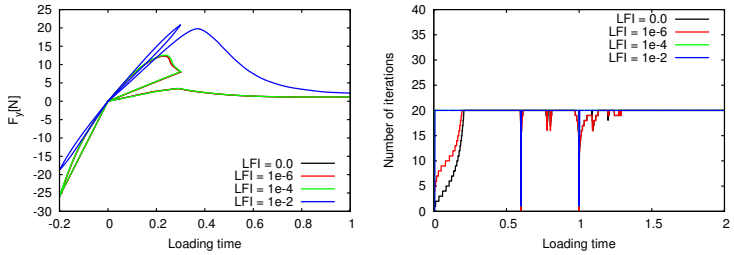


Figure 20: Example 3: Comparison of different L with a fixed number of 20 staggered iterations. At left, the load-displacement curves displaying the evolution of F_y versus u_y are shown. At right, the number of staggered iterations is displayed.

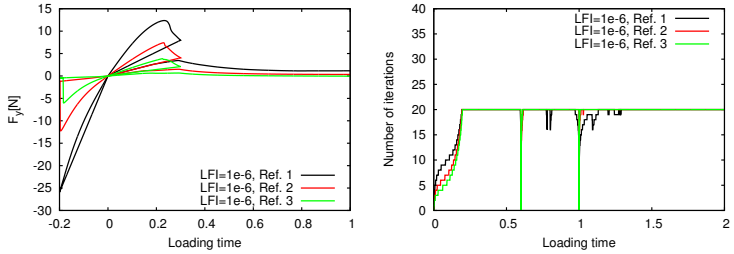


Figure 21: Example 3: Using $L = 1e-6$ and a fixed number of 20 staggered iterations, we compare the results on different refinement levels $1, 2, 3$. At left, the load-displacement curves displaying the evolution of F_y versus u_y are shown. At right, the number of staggered iterations is displayed.

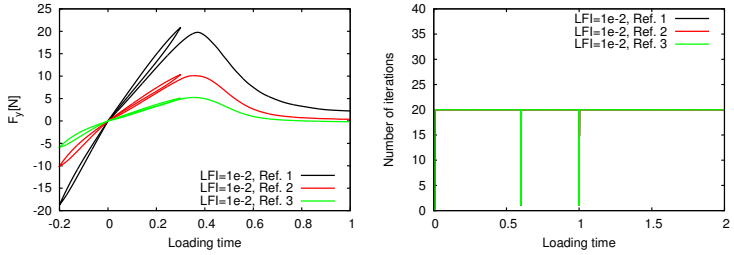


Figure 22: Example 3: Using $L = 1e-2$ and a fixed number of 20 staggered iterations, we compare the results on different refinement levels $1, 2, 3$. At left, the load-displacement curves displaying the evolution of F_y versus u_y are shown. At right, the number of staggered iterations is displayed.

5.4 Verification of Assumption 1

In this last set of computations, we verify whether Assumption 1 holds true in our computations. We choose some prototype settings, namely on the coarest mesh level Ref. 4 and $L_u = L_\varphi = 1e-6$. In Figure 23, we observe that $\operatorname{ess\,sup}_{x\in B} |\mathbf{e}(u^n(x))|$ varies, but always can be bounded from above with $M > 0$. The value of $\operatorname{ess\,sup}_{x\in B} |\mathbf{e}(u^n(x))|$ is the final strain when the L-scheme terminates. The minimum and maximum values shows that there are no significant variations in $\operatorname{ess\,sup}_{x\in B} |\mathbf{e}(u^n(x))|$ during the L-scheme iterations with respect to the finally obtained value.

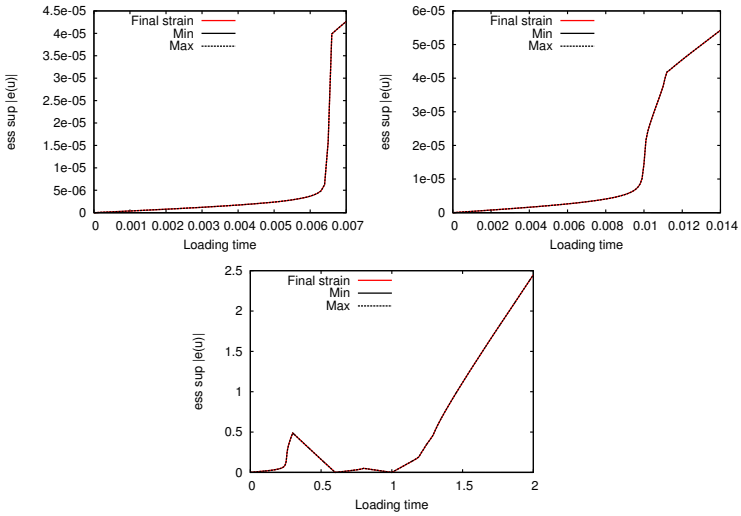


Figure 23: Comparison of $\operatorname{ess\,sup}_{x\in B} |\mathbf{e}(u^n(x))|$ and the minimal/maximal $\operatorname{ess\,sup}_{x\in B} |\mathbf{e}(u^n(x))|$ per loading time step.

6 Conclusions

We have proposed a novel staggered iterative algorithm for brittle fracture phase field models. This algorithm is employing stabilization and linearization techniques known in the literature as the 'L-scheme', which is a generalization of the Fixed Stress Splitting algorithm coming from poroelasticity. Through theory and numerical examples we have investigated the performance of our proposed variants of the L-scheme for brittle fracture phase field problems.

Under natural constraints that the elastic mechanical energy remains bounded, and that the model parameter ε is sufficiently large (i.e., that the diffusive zone around crack surfaces must be sufficiently thick), we have shown that a contraction of successive difference functions in energy norms can be obtained from the proposed scheme. This result implies the algorithm is converging monotonically with a linear convergence rate. However, in the convergence analysis there appears some unknown constants which makes the precise convergence rate, as well as the precise lower bound on ε unknown.

We provide detailed numerical tests where our proposed scheme is employed on several phase field brittle fracture bench-mark problems. For each numerical example we provide findings for different values of stabilization parameters. For most cases we let $L_u = L_\varphi > 0$, but for comparison we include also for the stabilization configurations $L_u = 0$ with $L_\varphi > 0$, and $L_u = L_\varphi = 0$. For the test cases presented here, there is only Example 1 where $L_u = 0$ does not work. This might be due to the very rapid crack growth, which sets Example 1 apart from Examples 2 and 3. In this regard, we conclude that further work is needed to find an optimal configuration of L_u and L_φ. For all numerical test we also provide computational justification for the assumption of bounded elastic mechanical energy. Furthermore, a slight dependency on h in the iteration counts is observed in the numerical

tests, but this is expected since we use $\varepsilon = 2h$, and as our analysis demonstrates, the convergence rate is dependent on ε. The variation in iteration numbers with mesh refinement is in any case sufficiently small enough that we conclude our algorithm is robust with respect to mesh refinement.

Moreover, due to the iteration spikes at the critical loading steps, we have included, for comparison, several results in which the iteration has been truncated (labeled LFI in Examples 1-3). Due to the monotonic convergence of the scheme, this strategy still produces acceptable results, while effectively avoiding the iteration spikes. We therefore conclude, at least for the particular examples presented here, that a truncation of the L-scheme can be employed for greatly improved efficiency with only negligible (depending on the situation at hand, of course) loss of accuracy.

Acknowledgements

This work forms part of Research Council of Norway project 250223. The authors also acknowledges the support from the University of Bergen. The first author, MKB, thanks the group 'Wissenschaftliches Rechnen' of the Institute of Applied Mathematics of the Leibniz University Hannover for the hospitality during his research stay from Oct - Dec 2018. The second author, TW, has been supported by the German Research Foundation, Priority Program 1748 (DFG SPP 1748) named *Reliable Simulation Techniques in Solid Mechanics. Development of Non-standard Discretization Methods, Mechanical and Mathematical Analysis.* The subproject within the SPP1748 reads *Structure Preserving Adaptive Enriched Galerkin Methods for Pressure-Driven 3D Fracture Phase-Field Models* (WI 4367/2-1).

References

[1] M. Ambati, T. Gerasimov, and L. De Lorenzis, *A review on phase-field models of brittle fracture and a new fast hybrid formulation*, Computational Mechanics, 55 (2015), pp. 383–405.

[2] L. Ambrosio and V. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via γ-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

[3] L. Ambrosio and V. Tortorelli, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B, 6 (1992), pp. 105–123.

[4] H. Amor, J.-J. Marigo, and C. Maurini, *Regularized formulation of the variational brittle fracture with unilateral contact: Numerical experiments*, J. Mech. Phys. Solids, 57 (2009), pp. 1209–1229.

[5] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells, *The `deal.II` library, version 8.5*, Journal of Numerical Mathematics, (2017), doi:10.1515/jnma-2016-1045.

[6] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II – a general purpose object oriented finite element library*, ACM Trans. Math. Softw., 33 (2007), pp. 24/1–24/27.

[7] M. Borregales, F. A. Radu, K. Kumar, and J. M. Nordbotten, *Robust iterative schemes for non-linear poromechanics*, Comput. Geosci., 22 (2018), pp. 1021–1038, doi:10.1007/s10596-018-9736-6, https://doi.org/10.1007/s10596-018-9736-6.

[8] J. W. Both, M. Borregales, J. M. Nordbotten, K. Kumar, and F. A. Radu, *Robust fixed stress splitting for Biot's equations in heterogeneous media*, Appl. Math. Lett., 68 (2017), pp. 101–108, doi:10.1016/j.aml.2016.12.019, https://doi.org/10.1016/j.aml.2016.12.019.

[9] B. Bourdin, *Numerical implementation of the variational formulation for quasi-static brittle fracture*, Interfaces and free boundaries, 9 (2007), pp. 411–430.

[10] B. Bourdin, G. Francfort, and J.-J. Marigo, *Numerical experiments in revisited brittle fracture*, J. Mech. Phys. Solids, 48 (2000), pp. 797–826.

[11] B. Bourdin, G. Francfort, and J.-J. Marigo, *The variational approach to fracture*, J. Elasticity, 91 (2008), pp. 1–148.

[12] S. Burke, C. Ortner, and E. Süli, *An adaptive finite element approximation of a variational model of brittle fracture*, SIAM J. Numer. Anal., 48 (2010), pp. 980–1012.

[13] N. Castelletto, J. White, and H. Tchelepi, *Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics*, International Journal for Numerical and Analytical Methods in Geomechanics, 39 (2015), pp. 1593–1618.

[14] W. Cheney, *Analysis for applied mathematics*, vol. 208 of Graduate Texts in Mathematics, Springer-Verlag, New York, 2001, doi:10.1007/978-1-4757-3559-8, https://doi.org/10.1007/978-1-4757-3559-8.

[15] D. Cioranescu and P. Donato, *An introduction to homogenization*, vol. 17 of Oxford Lecture Series in Mathematics and its Applications, The Clarendon Press, Oxford University Press, New York, 1999.

[16] T. A. Davis and I. S. Duff, *An unsymmetric-pattern multifrontal method for sparse LU factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 140–158.

[17] P. Deuflhard, *Newton Methods for Nonlinear Problems*, vol. 35 of Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2011.

[18] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[19] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems*, Stud. Math. Appl. 15, North Holland, Amsterdam, 1983.

[20] G. A. Francfort and J.-J. Marigo, *Revisiting brittle fracture as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342, doi:10.1016/S0022-5096(98)00034-9, https://doi.org/10.1016/S0022-5096(98)00034-9.

[21] T. Gerasimov and L. D. Lorenzis, *A line search assisted monolithic approach for phase-field computing of brittle fracture*, Computer Methods in Applied Mechanics and Engineering, 312 (2016), pp. 276 – 303.

[22] R. Glowinski and P. L. Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM Stud. Appl. Math. 9, SIAM, Philadelphia, 1989.

[23] A. A. Griffith and M. Eng, *Vi. the phenomena of rupture and flow in solids*, Phil. Trans. R. Soc. Lond. A, 221 (1921), pp. 163–198.

[24] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities. cambridge mathematical library series*, 1967.

[25] T. Heister, M. F. Wheeler, and T. Wick, *A primal-dual active set method and predictor-corrector mesh adaptivity for computing fracture propagation using a phase-field approach*, Comp. Meth. Appl. Mech. Engrg., 290 (2015), pp. 466 – 495.

[26] J. Kim, H. A. Tchelepi, R. Juanes, et al., *Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics*, in SPE reservoir simulation symposium, Society of Petroleum Engineers, 2009.

[27] M. Kirkesæther Brun, E. Ahmed, I. Berre, J. M. Nordbotten, and F. A. Radu, *Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport*, arXiv e-prints, (2019), arXiv:1902.05783, p. arXiv:1902.05783, arXiv:1902.05783.

[28] F. List and F. A. Radu, *A study on iterative methods for solving Richards' equation*, Comput. Geosci., 20 (2016), pp. 341–353, doi:10.1007/s10596-016-9566-3, https://doi.org/10.1007/s10596-016-9566-3.

[29] A. Mesgarnejad, B. Bourdin, and M. Khonsari, *Validation simulations for the variational approach to fracture*, Computer Methods in Applied Mechanics and Engineering, 290 (2015), pp. 420 – 437.

[30] C. Miehe, M. Hofacker, and F. Welschinger, *A phase field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits*, Comput. Methods Appl. Mech. Engrg., 199 (2010), pp. 2765–2778, doi:10.1016/j.cma.2010.04.011, https://doi.org/10.1016/j.cma.2010.04.011.

[31] C. Miehe, F. Welschinger, and M. Hofacker, *Thermodynamically consistent phase-field models of fracture: variational principles and multi-field fe implementations*, Int. J. Numer. Methods Engrg., 83 (2010), pp. 1273–1311.

[32] A. Mikelić, B. Wang, and M. F. Wheeler, *Numerical convergence study of iterative coupling for coupled flow and geomechanics*, Comput. Geosci., 18 (2014), pp. 325–341, doi:10.1007/s10596-013-9393-8, https://doi.org/10.1007/s10596-013-9393-8.

[33] A. Mikelić and M. F. Wheeler, *Convergence of iterative coupling for coupled flow and geomechanics*, Comput. Geosci., 17 (2013), pp. 455–461, doi:10.1007/s10596-012-9318-y, https://doi.org/10.1007/s10596-012-9318-y.

[34] A. Mikelić, M. F. Wheeler, and T. Wick, *A quasi-static phase-field approach to pressurized fractures*, Nonlinearity, 28 (2015), pp. 1371–1399.

[35] A. Mikelić, M. F. Wheeler, and T. Wick, *Phase-field modeling through iterative splitting of hydraulic fractures in a poroelastic medium*, GEM - International Journal on Geomathematics, 10 (2019).

[36] I. Neitzel, T. Wick, and W. Wollner, *An optimal control problem governed by a regularized phase-field fracture propagation model*, SIAM J. Control Optim., 55 (2017), pp. 2271–2288, doi:10.1137/16M1062375, https://doi.org/10.1137/16M1062375.

[37] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Ser. Oper. Res. Financial Engrg., 2006.

[38] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373, doi:10.1016/j.cam.2003.04.008, https://doi.org/10.1016/j.cam.2003.04.008.

[39] J. M. Sargado, E. Keilegavlen, I. Berre, and J. M. Nordbotten, *High-accuracy phase-field models for brittle fracture based on a new family of degradation functions*, J. Mech. Phys. Solids, 111 (2018), pp. 458–489, doi:10.1016/j.jmps.2017.10.015, https://doi.org/10.1016/j.jmps.2017.10.015.

[40] S. Sun, B. Rivière, and M. F. Wheeler, *A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media*, in Recent progress in computational and applied PDEs (Zhangjiajie, 2001), Kluwer/Plenum, New York, 2002, pp. 323–351.

[41] S. Sun and M. F. Wheeler, *Discontinuous Galerkin methods for coupled flow and reactive transport problems*, Appl. Numer. Math., 52 (2005), pp. 273–298, doi:10.1016/j.apnum.2004.08.035, https://doi.org/10.1016/j.apnum.2004.08.035.

[42] M. Wheeler, T. Wick, and W. Wollner, *An augmented-Lagangrian method for the phase-field approach for pressurized fractures*, Comp. Meth. Appl. Mech. Engrg., 271 (2014), pp. 69–85.

[43] T. Wick, *Solving monolithic fluid-structure interaction problems in arbitrary Lagrangian Eulerian coordinates with the deal.II library*, Archive of Numerical Software, 1 (2013), pp. 1–19, http://www.archnumsoft.org.

[44] T. Wick, *An error-oriented Newton/inexact augmented Lagrangian approach for fully monolithic phase-field fracture propagation*, SIAM Journal on Scientific Computing, 39 (2017), pp. B589–B617, doi:10.1137/16M1063873.

[45] T. Wick, *Modified Newton methods for solving fully monolithic phase-field quasi-static brittle fracture propagation*, Comput. Methods Appl. Mech. Engrg., 325 (2017), pp. 577–611, doi:10.1016/j.cma.2017.07.026, https://doi.org/10.1016/j.cma.2017.07.026.

[46] B. Winkler, *Traglastuntersuchungen von unbewehrten und bewehrten Betonstrukturen auf der Grundlage eines objektiven Werkstoffgesetzes fuer Beton*, PhD thesis, University of Innsbruck, 2001.

[47] K. Yosida, *Functional analysis*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, doi:10.1007/978-3-642-61859-8, https://doi.org/10.1007/978-3-642-61859-8. Reprint of the sixth (1980) edition.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230844427 (print)
9788230853443 (PDF)