

Iteratively coupled mixed finite element solver for thermo-hydro-mechanical modeling of permafrost thaw

Naren Vohra, Malgorzata Peszynska *

Oregon State University, Corvallis, OR, 97331, USA

ARTICLE INFO

Keywords:

Thermo-hydro-mechanical models
Heterogeneous permafrost
Heat conduction with phase change
Poroelectricity
Mixed finite elements
Nonlinear constitutive relationships
Iterative solvers

ABSTRACT

In this paper we consider computational challenges associated with thermo-hydro-mechanical models for simulation of subsidence due to permafrost thaw. The model we outline couples heat conduction with phase change and thermal advection to Biot's poroelasticity equations with attention paid to the dependence of the constitutive parameters on temperature. Our numerical scheme uses the lowest order mixed finite elements for discretization of thermal and hydrological flow, and Galerkin finite elements for mechanics, and uses an implicit-explicit time stepping. We set up an iterative solver that solves the thermal subproblem followed by the hydro-mechanical subproblem, and demonstrate its robustness in practical heterogeneous permafrost scenarios. We also identify the challenges associated with the roughness of the dependence of mechanical parameters on the temperature.

1. Introduction

In this paper we study and address some of the computational mathematics challenges associated with thermo-hydro-mechanical (TpHM) models used to simulate freezing and thawing scenarios in permafrost. In particular, we focus on scenarios involving ground subsidence due to permafrost thaw.

Permafrost is defined as ground that remains frozen for two or more years [1,2]. An increase in global climate temperatures has accelerated thawing of permafrost, and this leads to damage of man-made infrastructure and degradation of natural landscapes. For example, buildings, railway tracks, and pipelines in permafrost regions are affected by the ground subsidence, while the thawing of ground ice and ice-rich terrain features such as pingos leads to the formation of large marshy lakes called thermokarsts [1].

Simulating ground subsidence in thawing permafrost is far from trivial. A holistic approach typically involves the study of multiphysics models that consider the thermal (Tp), hydrological (H), and mechanical (M) aspects of permafrost and their intricate coupling. We use the abbreviation "TpHM" for thermo-hydro-mechanical models, where the "p" serves to emphasize the mathematical complexity due to phase transitions.

We mention two major characteristics for thermo-hydro-mechanical models: (i) the complexity of the governing system of equations and (ii) the complexity of the computational scheme for numerical approximation. For (i), the standard approach is to develop the models for the Tp, H, and M aspects individually using first principles, and then define constitutive relationships to reflect the coupled dynamics of frozen soils. For example, frozen soils have lower hydraulic permeability and increased mechanical strength at subzero temperatures than completely thawed soils [1,3,4]. These relationships are typically obtained empirically after conducting in situ or laboratory experiments, and they are given by nonlinear algebraic expressions. For (ii) an appropriate computational scheme has to be implemented carefully to efficiently and accurately solve the fully coupled TpHM model, and not all schemes available for the individual subproblems work well together.

* Corresponding author.

E-mail addresses: vohran@oregonstate.edu (N. Vohra), mpesz@math.oregonstate.edu (M. Peszynska).

Concerning (i), various TpHM models have been implemented in the applications literature; see for example [5–13]. The models are more or less consistent between each other, insofar the governing systems of equations are concerned. At the Darcy scale, these include heat conduction with phase change (Tp), equations of conservation of mass for the different phases and components (H), and the balance of momentum equations (M). However, the fully coupled TpHM models presented in literature feature different constitutive relationships employed by different authors to represent or emphasize particular attributes of permafrost.

Regarding (ii), a computational scheme has two crucial components: the discretization and the nonlinear solver. For discretization, Galerkin finite elements (P1) are used most commonly for all Tp, H, and M components; see for example [9,11,14]. While P1 based schemes are easy to implement, they are not conservative, an aspect which is especially important in heterogeneous media such as permafrost where the thermal and hydrological fluxes play an important part. Approaches that respect conservation properties include, for example, lowest order mixed finite element methods are used in [15], finite volumes in [16] and mimetic finite differences in [17] for TpH. For thermo-poroelasticity THM models (in the absence of phase change), discontinuous Galerkin method has been used in [18].

Following the numerical discretization, the other important aspect of computational schemes is the nonlinear solver. For example, monolithic solvers in [9,14] use Picard's and Newton's methods, respectively, for fully implicit coupling. However, the computational complexity of monolithic approaches may be inefficient for larger systems, and does not allow to use well tested robust versions of individual model components. An alternative is a sequential (or staggered, or explicit) approach. In this approach, the fully coupled system is broken into subproblems that are solved one after the other (say, the thermal subproblem may be solved first, followed by the hydro-mechanical or vice-versa). The efficiency of such a sequential approach in isothermal fluid flow and geomechanical modeling has been demonstrated in [19]. Sequential approaches may be more efficient and more flexible to implement, but they might also only be conditionally stable. A sequential approach may iterated further at each time step until the solution converges within a specified tolerance. This is commonly known as an iteratively coupled approach, and the solution obtained is, in principle, the same as the that obtained using a monolithic approach.

Although in the applications literature TpHM models have been successfully implemented to simulate practical scenarios, the questions related to the well-posedness or robustness of the computational schemes are not addressed. Quite often, rather few details regarding the scheme are provided, with only a mention of some commercially available software used for implementation. In the mathematical community, such questions have been taken up, but their focus has been on thermo-poroelasticity models [15,20,21]. Moreover, most existing work on thermo-poroelasticity is done in the absence of phase change and by assuming linear constitutive relationships, features which do not completely define TpHM models.

In this paper, we concern ourselves with the challenges raised by point (ii) above and provide a first step towards addressing some of these challenges by building a robust conservative iterative solver for TpHM models. We have introduced and demonstrated the use of the lowest order mixed finite elements for heat conduction models in [22,23], and we now focus on extending their use to the hydro-mechanical aspect of permafrost. More specifically, we approximate the temperature, enthalpy, and pressure in the space of piecewise-constants (P0), the thermal and hydrological flux in the lowest order Raviart–Thomas space $RT_{[0]}$ (RT0), and the displacements in the space of continuous bilinear elements (Q1).

Our iterative solver first applies our existing P0-P0 scheme for heat conduction models, followed by a monolithic P0-RT0-Q1 solver for the hydro-mechanical subproblem. More specifically, our P0-P0 solver for the thermal subproblem is implemented using an implicit–explicit approach via operator splitting, where the thermal advection step is solved explicitly using upwinding, and then the diffusion step is solved implicitly [24]. For our P0-RT0-Q1 hydro-mechanical solver, our implementation is based on a monolithic scheme.

The lowest order mixed finite elements are celebrated for their mass conservation properties and ease of implementation as a cell-centered-finite-difference scheme [25,26]. For isothermal hydro-mechanical systems, P0-RT0-Q1 elements have been widely studied for Biot's system of poroelasticity for single phase systems [27–29] and they are known to lead to spurious pressure oscillations in scenarios where the liquid is nearly incompressible and the hydraulic permeability or time step is small, and may also suffer from Poisson locking when the elastic material reaches the incompressible limit [28,30,31]. However, in mathematical literature such numerical artifacts have only been highlighted through the use of rather special and sometimes artificially constructed non-physical examples with extreme parameter values that are usually not relevant to practical scenarios. In this paper we demonstrate the use of the P0-RT0-Q1 elements in our permafrost modeling applications and show that for practical grid and time step sizes our models do not suffer from the aforementioned artifacts.

Our contributions in this paper are as follows. We first present the TpHM model from first principles and then (a) draw comparisons between the TpHM system and Biot's poroelasticity equations. We highlight the role of the densities of the different permafrost components. (b) We provide a review of the existing numerical methods for TpHM models, and isolate the potential challenges with their extension to TpHM models. (c) We then introduce our iterative solver and prove its convergence. We show that our solver is robust in heterogeneous scenarios and does not suffer from well-known numerical artifacts such as non-physical oscillations or Poisson-locking. (d) Through our numerical experiments, we also identify the importance of Lipschitz continuous regularization of dependence of elasticity parameters on the temperature for convergence of our iterative solver, and the impact of its Lipschitz constant. We further investigate the use of equal and unequal ice and water densities which leads to different solution dynamics in typical permafrost scenarios. To our knowledge, this is a first work undertaking study of computational mathematics aspects of TpHM models.

In this paper we focus on permafrost thaw, and we ignore the effects predominantly associated with freezing such as frost heave and cryosuction. Moreover, we also ignore the effects of mechanics on the heat conduction, since we expect the overburden pressures to be small in permafrost soil scenarios. The study of these aspects is deferred to a future work.

The outline of the paper is as follows. In Section 2 we provide details of the governing equations and constitutive relationships of the TpHM model. In Section 3 we give details of the mixed finite element discretization. Next in Section 4 we review existing numerical methods for TpHM models. In Section 5 we provide the details of our iterative solver and in Section 6 we demonstrate the robustness of our solver in practical permafrost scenarios. We summarize in Section 7, and list acknowledgments of support in Section “Acknowledgements”.

2. Model description

In this section we develop the physical models and make precise various constitutive relationships. We also state the assumptions to be used in our analysis.

Let $\Omega \subset \mathbb{R}^d$ be an open bounded connected set representing a heterogeneous domain that is occupied by permafrost. We denote by ν_n the normal to $\partial\Omega$. We further assume that Ω is occupied by N_r different non-overlapping domains $\Omega^{(j)}$, with each $\Omega^{(j)}$ occupied by a particular soil type or rock with its own physical parameters.

We use the subscripts l, i and r to denote liquid water, ice, and rock grain, respectively. A list of the different parameters used in this paper is tabulated in Table 1.

We denote the function spaces $M = L^2(\Omega)$, $V = (H_0^1(\Omega))^d$, and $X = H_{div}(\Omega)$, where

$$(H_0^1(\Omega))^d = \{f \in (H^1(\Omega))^d \mid f|_{\partial\Omega} = 0\},$$

$$H_{div}(\Omega) = \{f \in (L^2(\Omega))^d \mid \nabla \cdot f \in L^2(\Omega)\}.$$

The L^2 inner product is denoted by $(f_1, f_2) = \int_{\Omega} f_1 f_2$ for scalar-valued $f_1, f_2 \in L^2(\Omega)$ or vector-valued $f_1, f_2 \in (L^2(\Omega))^d$, and we denote the L^2 norm by $\|f\|_2 = (f, f)^{\frac{1}{2}}$, $\forall f \in L^2(\Omega), (L^2(\Omega))^d$. Additionally, we will make use of the following norms

$$\|f\|_{\infty,1} = \sup_{t \in [0,T]} \left(\int_{\Omega} |f(t,x)| dx \right), \quad \forall f \in L^{\infty}(0,T; L^1(\Omega)), \quad (1a)$$

$$\|f\|_{\infty,2} = \sup_{t \in [0,T]} \left(\int_{\Omega} |f(t,x)|^2 dx \right)^{\frac{1}{2}}, \quad \forall f \in L^{\infty}(0,T; L^2(\Omega)), \quad (1b)$$

$$\|f\|_{2,2} = \left(\int_0^T \int_{\Omega} |f(t,x)|^2 dx dt \right)^{\frac{1}{2}}, \quad \forall f \in L^2(0,T; L^2(\Omega)), \quad (1c)$$

for some $T > 0$. More details on the computation of the norms given by (1) in our numerical experiments are provided in Appendix A.2.

We start with an assumption regarding the densities of the permafrost components.

Assumption 2.1. The density of liquid water, ice, and rock grains is constant, i.e., $\rho_{ph} = \text{const}$, $ph \in \{l, i, r\}$, where the constants are not necessarily equal to the same value.

The above assumption implies that liquid water, ice, and rock grains are incompressible, as is frequently assumed for permafrost models; see, e.g., [5–7,13,14,16,32]. However, in [12], ρ_l is assumed to depend linearly on the pressure and quadratically on the temperature; this seems relevant for the hydrate applications studied therein. In [17], the authors consider temperature dependent ρ_l, ρ_i in their TpH models of permafrost, and study the effects of assuming $\rho_l = \rho_i$. We study this aspect in numerical experiments.

2.1. Heat conduction model (Tp)

We now provide details of the thermal subproblem. We start by making the following assumption on the thermal parameters.

Assumption 2.2. The thermal parameters of liquid water, ice, and rock grains of all soil types are constants and uniformly bounded by positive constants, i.e., $\exists c_{\min}, c_{\max} \in \mathbb{R}$ such that

$$0 < c_{\min} \leq c_l, c_i, c_r(x) \leq c_{\max} < \infty, \quad 0 < k_{\min} \leq k_l, k_i, k_r(x) \leq k_{\max} < \infty, \quad \forall x \in \Omega.$$

We also assume that the latent heat satisfies $L \geq 0$.

Heat conduction with phase change and convection in permafrost is modeled as follows [33]

$$\partial_t w - \nabla \cdot (k \nabla \theta) + \nabla \cdot (c_l \theta q_l) = f, \quad w = \alpha(x, \theta, \eta), \quad (2)$$

where θ is the temperature, w is the enthalpy per unit volume, $k = k(x, \theta)$ is the thermal conductivity, c_l is the volumetric heat capacity of liquid water, q_l is the hydrological flux describing the movement of unfrozen liquid water (see Section 2.2), η is the porosity of the soil, f is an external source term, and α is the nonlinear temperature-enthalpy relationship.

The nonlinear relationship $\alpha(\theta)$ is defined as

$$\alpha(x, \theta, \eta) = \int_{\theta_*}^{\theta} c(x, s) ds + L \eta \chi_l(x, \theta), \quad c(x, \theta) = c_f(x) + (c_u(x) - c_f(x)) \chi_l(x, \theta), \quad (3)$$

Table 1

A list of the different variables, parameters, and relationships used in this paper.

Notation	Description/SI Unit
$\{l, i, r\}$	Liquid water l , ice i , and rock grains r
Variable	Description/SI Unit
θ	Temperature [$^{\circ}\text{C}$]
w	Enthalpy per unit volume [J/m^3]
χ_l	Water fraction [—]
q_θ	Thermal flux [$\text{J}/\text{m}^2 \text{ s}$]
p	Pressure [Pa]
q_l	Hydrological flux of liquid water [m/s]
v_{ph}	Velocity of phase/component $ph \in \{l, i, r\}$ [m/s]
u	Displacement [m]
Parameter	Description/SI Unit
θ_*	Freezing point depression [$^{\circ}\text{C}$]
c	Heat capacity per unit volume [$\text{J}/\text{m}^3 \text{ }^{\circ}\text{C}$]
k	Thermal conductivity [$\text{J}/\text{m s }^{\circ}\text{C}$]
η	Porosity [—]
κ	Permeability [m^2]
μ_l	Viscosity [Pa s]
K	Hydraulic conductivity [m/s]
β_l	Fluid compressibility [$1/\text{Pa}$]
ρ	Density [kg/m^3]
γ	Density ratio $\gamma = 1 - \rho_i/\rho_l$
c_0	Specific storage coefficient [—]
E	Young's modulus [Pa]
ν	Poisson's ratio [—]
λ, μ	Lamé parameters [Pa]
G	Acceleration due to gravity [m/s^2]
Function	Description
α	Temperature-enthalpy function; $w = \alpha(\theta)$
ζ	$\zeta(\theta) = 1 - \gamma\eta(1 - \chi_l(\theta))$

where $c_u(x) = \eta(x)c_l + (1 - \eta(x))c_r(x)$ and $c_f(x) = \eta(x)c_i + (1 - \eta(x))c_r(x)$ are the heat capacities of “unfrozen” (completely thawed) and “frozen” soils, $\chi_l = \chi_l(x, \theta)$ is the volumetric water fraction given by the empirically determined soil type specific soil freezing curve (SFC) (see Section 2.4.1). The freezing point depression θ_* is the temperature above which water exists only in the liquid phase. Here we choose $\theta_* = 0$ [$^{\circ}\text{C}$].

In (2), we take the thermal conductivity $k = k(x, \theta)$ to be the harmonic weighted average

$$k = \left[\frac{\eta\chi_l}{k_l} + \frac{\eta(1 - \chi_l)}{k_i} + \frac{1 - \eta}{k_r} \right]^{-1}. \quad (4)$$

but other choices include arithmetic, geometric average weighting or upscaling; see the discussion in [34]. Regardless, from Assumption 2.2, it follows that

$$k_{\min} \leq k(x, \theta) \leq k_{\max}, \quad \forall x \in \Omega, \theta \in \mathbb{R}, \quad (5)$$

We refer the reader to [22,23,34] for complete details on heat conduction models in permafrost, including our results on upscaling the Stefan problem and connecting it to permafrost models.

2.2. Hydro-mechanical model (HM)

We approximate frozen soils as poroelastic materials, and begin by reviewing the linear Biot's poroelasticity equations used for modeling isothermal flow and deformation in saturated porous media.

2.2.1. Linear poroelasticity for single phase system

Suppose Ω is completely saturated with liquid water. In isothermal conditions, Biot's system of poroelasticity is given by [35,36]

$$c_0 \partial_t p + \alpha_B \partial_t (\nabla \cdot u) + \nabla \cdot q_l = g, \quad (6a)$$

$$-\nabla \cdot [\lambda \nabla \cdot u I + 2\mu \epsilon(u)] + \alpha_B \nabla p = l + [\rho_l \eta + \rho_r (1 - \eta)] G, \quad (6b)$$

where u is the displacement, p is the pressure, λ, μ are the Lamé coefficients, $\epsilon(u) = \frac{1}{2} (\nabla u + \nabla u^T)$ is the linearized strain tensor, α_B is the Biot–Willis constant, c_0 is the specific storage coefficient, I is the identity matrix, and g and l are external source and

force terms, and G is the acceleration due to gravity. The hydrological flux $q_l = -\kappa\mu_l^{-1}(\nabla p - \rho_l G \nabla D)$ is defined using Darcy's law, where κ is the permeability of the porous media and μ_l is the viscosity of water, and $D = D(x)$ is the depth at x . For example, if $\Omega = (a, b)$ represents a column of soil with $x = a$ representing its top and $x = b$ its bottom, then $D(x) = x - a$. We also denote by $\tilde{\sigma} = \lambda \nabla \cdot u I + 2\mu \epsilon(u) - p I$ the total stress tensor.

For an elastic material, the Lamé parameters λ and μ in (6b) are determined using the Young's modulus E and Poisson's ratio ν [37](Pg. 146)

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \nu = \frac{E}{2(1+\nu)}. \quad (7)$$

Remark 2.1. Note that in (6a), it is assumed that rock grains are incompressible but liquid water is slightly compressible with a non-constant density $\rho_l = \rho_l(p)$. The specific storage coefficient is defined as $c_0 = \eta\beta_l$ [38](Pg. 122, Eq. 33d) [39]Pg. 170, where β_l is the compressibility of liquid water $\beta_l = \frac{1}{\rho_l} \frac{\partial \rho_l}{\partial p}$. For liquid water $c_0 \approx O(10^{-9})$ [1/Pa] [40]. If the liquid water is assumed to be incompressible, then $c_0 = 0$ in (6a).

Remark 2.2. In this paper we set the Biot-Willis constant $\alpha_B = 1$, consistent with Assumption 2.1 [38](Pg. 122).

2.2.2. Flow and deformation in frozen soils

We now provide the equations for flow and deformation in frozen soils. We closely follow the presentation in [14] and fill in the details. We will assume Assumption 2.1.

Let Ω be occupied by permafrost. The equations for conservation of mass of liquid water, ice, and rock grains are given by [14](Eq. (3))

$$\partial_t(\rho_l \eta \chi_l) + \nabla \cdot (\rho_l \eta \chi_l v_l) = \hat{\rho} \quad (8a)$$

$$\partial_t(\rho_i \eta (1 - \chi_l)) + \nabla \cdot (\rho_i \eta (1 - \chi_l) v_i) = -\hat{\rho} \quad (8b)$$

$$\partial_t(\rho_r (1 - \eta)) + \nabla \cdot (\rho_r (1 - \eta) v_r) = 0, \quad (8c)$$

where v_{ph} , $ph \in \{l, i, r\}$ is the velocity of phase/component ph , and $\hat{\rho}$ [kg/m³ s] is the rate of mass exchange between liquid water and ice. We assume $v_i = v_r$, i.e., that the ice moves with the same velocity of rock grains v_r , and that $v_r = \partial_t u$ [14](Pg. 5).

Adding (8a), (8b), and (8c) after dividing them by ρ_l , ρ_i and ρ_r , respectively, we get

$$\partial_t[\eta \chi_l + \eta(1 - \chi_l) + (1 - \eta)] + \nabla \cdot [\eta \chi_l v_l + \eta(1 - \chi_l) v_r + (1 - \eta) v_r] = \frac{\hat{\rho}}{\rho_l} - \frac{\hat{\rho}}{\rho_i}. \quad (9)$$

The first term in (9) is $\partial_t(1) = 0$, and further rearranging the terms we get [14](Eq. (4))

$$\nabla \cdot [\eta \chi_l (v_l - v_r)] + \nabla \cdot v_r = -\gamma \frac{\hat{\rho}}{\rho_i}, \quad (10)$$

where $\gamma = 1 - \rho_i/\rho_l \approx 0.09$ is a constant. Now, we divide (8b) by ρ_i to get

$$-\frac{\hat{\rho}}{\rho_i} = \partial_t(\eta(1 - \chi_l)) + \nabla \cdot (\eta(1 - \chi_l) v_i). \quad (11)$$

Substituting (11) into (10) and using $v_i = v_r$ we get

$$\nabla \cdot (\eta \chi_l (v_l - v_r)) + \nabla \cdot v_r = \gamma [\partial_t(\eta(1 - \chi_l)) + \nabla \cdot (\eta(1 - \chi_l) v_r)]. \quad (12)$$

Rearranging terms in (12) gives us

$$\nabla \cdot [(1 - \gamma \eta(1 - \chi_l)) v_r] + \nabla \cdot [\eta \chi_l (v_l - v_r)] = \gamma \partial_t [\eta(1 - \chi_l)] \quad (13)$$

This can be obtained also using $q_l = \eta \chi_l (v_l - v_r)$ [41](Eq. (7)) and $v_r = \partial_t u$

$$\nabla \cdot [(1 - \gamma \eta(1 - \chi_l)) \partial_t u] + \nabla \cdot q_l = \gamma \partial_t [\eta(1 - \chi_l)] \quad (14)$$

Following notation similar to [14](Pg. 7), and by denoting

$$\zeta = 1 - \gamma \eta(1 - \chi_l), \quad (15)$$

we get from (14)

$$\nabla \cdot (\zeta \partial_t u) + \nabla \cdot q_l = \gamma \partial_t [\eta(1 - \chi_l)] + g, \quad (16)$$

where we have included an additional source term g .

For mechanics in permafrost, we consider linear elasticity [11](Eq. 18)

$$-\nabla \cdot [\lambda \nabla \cdot u I + 2\mu \epsilon(u)] + \nabla p = l + \bar{\rho} G, \quad (17)$$

where $\bar{\rho} = \rho_l \eta \chi_l + \rho_i \eta(1 - \chi_l) + \rho_r(1 - \eta)$, and λ, μ now depend on the temperature.

Table 2

List of primary unknowns as well as the interdependence of Tp, H, and M components through the coupling terms in TpHM model (20).

Tp	
Variable	θ
Coupling terms	q_l from H, η from M, f (source)
H	
Variable	p
Coupling terms	η from M, $\partial_t u$ from M, $\zeta(x, \theta, \eta)$, $\kappa(x, \theta, \eta)$ (data), $\chi_l(x, \theta)$ (data), g (source)
M	
Variable	u
Coupling terms	∇p from H, $\lambda(x, \theta)$ (data), $\mu(x, \theta)$ (data), $\bar{\rho}G$ from Tp, l (source)

The system (16)–(17) is closed with Darcy's law $q_l = -\kappa(\theta, \eta)\mu_l^{-1}(\nabla p - \rho_l G \nabla D)$. Here we assume the permeability to also depend on the temperature and the porosity, and the equation for the porosity which is obtained after dividing (8c) by ρ_r [14](Eq. 18)

$$\partial_t \eta - \nabla \cdot [(1 - \eta)\partial_t u] = 0. \quad (18)$$

Before proceeding further, it is worthwhile to draw comparisons between (16) and (6a). In the region where $\chi_l = 1$ (i.e., completely thawed soils), Eq. (16) reduces to Eq. (6a) in Biot's system with $c_0 = 0$. Similarly, if the density variation between ice and liquid water is ignored, i.e., if we assume $\rho_i = \rho_l$, then $\gamma = 0$, and consequently (16) reduces to (6a) with $c_0 = 0$. This observation makes it clear that any computational challenges associated with isothermal flow and deformation modeling using (6) are expected to arise for frozen soils as well. We discuss this further in Section 4.

Another interesting observation comes from comparing (16) to its counterpart in thermo-hydrological models. For thermo-hydrological model in nondeformable media, the conservation of mass equation is given by [33,42]

$$\nabla \cdot q_l = -\gamma \eta \partial_t \chi_l, \quad q_l = -\frac{\kappa}{\mu_l}(\nabla p - \rho_l G \nabla D). \quad (19)$$

That is the effect of density variation acts only as a source in the conservation of mass equation. Whereas in (16), apart from the source term $\gamma \partial_t [\eta(1 - \chi_l)]$, γ also leads to the coefficient ζ in $\zeta \partial_t u$.

2.3. Fully coupled thermo-hydro-mechanical model (TpHM)

We now summarize our fully coupled TpHM model. We consider

$$\partial_t w - \nabla \cdot (k(x, \theta) \nabla \theta) + \nabla \cdot (c_l \theta q_l) = f, \quad w = \alpha(x, \theta, \eta), \quad (20a)$$

$$\nabla \cdot (\zeta(x, \theta, \eta) \partial_t u) + \nabla \cdot q_l - \gamma \partial_t [\eta(1 - \chi_l)] = g, \quad (20b)$$

$$-\nabla \cdot [\lambda(x, \theta) \nabla \cdot u I + 2\mu(x, \theta) \epsilon(u)] + \nabla p = l + \bar{\rho}G, \quad (20c)$$

$$\partial_t \eta - \nabla \cdot [(1 - \eta)\partial_t u] = 0, \quad (20d)$$

where the hydrological flux is given by $q_l = -\kappa(x, \theta, \eta)\mu_l^{-1}(\nabla p - \rho_l G \nabla D)$. For exposition purposes, we list the variables and interdependent parameters in (20) in the Tp, H, and M regimes in Table 2.

2.4. Constitutive relationships

In this section we provide details of the constitutive relationships used in our TpHM model (20).

2.4.1. Soil freezing curve

In frozen soils the dependence of the water fraction χ_l on temperature θ is expressed by the soil freezing curve (SFC). For our numerical experiments we consider the following expression adapted from [5]

$$\chi_l = \begin{cases} 1; & \theta > \theta_* \\ \chi_{res} + (1 - \chi_{res})e^{b(\theta - \theta_*)}; & \theta \leq \theta_*, \end{cases} \quad (21)$$

where $\chi_{res} \in [-1, 1]$, $b \in 1/^\circ\text{C}$, and $\theta_* \in [^\circ\text{C}]$ are parameters that depend on the soil type. More precisely, θ_* denotes the temperature above which water only exists in the liquid phase (typically $\theta_* \approx 0 \text{ } [^\circ\text{C}]$), χ_{res} denotes the residual water fraction at extremely low temperatures, and b controls the steepness of the curve.

For other SFC expressions, illustrations and comparisons, the reader is referred to [23].

2.4.2. Permeability

The permeability κ of partially frozen soils decreases with water fraction which decreases with θ [1,3,43], and is somewhat analogous to that observed when water fraction decreases in multiphase flow, e.g., in Richards' equation. There is also a dependence of κ on the mechanical deformation frequently modeled with the dependence on the porosity η .

We make the following assumption.

Assumption 2.3. The permeability tensor $\kappa : \Omega \times \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}^{d \times d}$ is symmetric, uniformly bounded, and elliptic, i.e., $\exists \kappa_{\min}, \kappa_{\max} > 0$ such that $\forall x \in \Omega, \theta \in \mathbb{R}, \eta \in [0, 1]$

$$\kappa_{\min} \|\xi\|_2^2 \leq \xi^T \kappa(x, \theta, \eta) \xi \leq \kappa_{\max} \|\xi\|_2^2, \quad \forall \xi \in \mathbb{R}^d, \quad \xi \neq 0, \quad (22)$$

where in (22) $\|\cdot\|_2$ now denotes the standard l^2 norm on $\mathbb{R}^d$ defined as $\|\xi\|_2 = \left(\sum_{j=1}^d |\xi_j|^2 \right)^{\frac{1}{2}}$, $\xi = (\xi_1, \xi_2, \dots, \xi_d)$. Further, we assume κ^{-1} is entry-wise Lipschitz in θ , i.e., $\exists L_{\kappa^{-1}} > 0$ such that $x \in \Omega, \forall \theta_2, \theta_1 \in \mathbb{R}, \eta \in [0, 1]$

$$\left| \kappa_{i,j}^{-1}(x, \theta_2, \eta) - \kappa_{i,j}^{-1}(x, \theta_1, \eta) \right| \leq L_{\kappa^{-1}} |\theta_2 - \theta_1|, \quad \forall 1 \leq i, j \leq d. \quad (23)$$

For our simulations we blend the Carman–Kozeny model [9] with impedance model [44]

$$\kappa(x, \theta, \eta) = \kappa_{0,u} \kappa_{rel} \frac{\eta^3}{(1-\eta)^2} \frac{(1-\eta_{init})^2}{\eta_{init}^3} I, \quad (24a)$$

where $\kappa_{0,u} = \kappa_{0,u}(x) \in \mathbb{R}$ is the intrinsic permeability of thawed soil, $\kappa_{rel} = \kappa_{rel}(x, \theta) \in \mathbb{R}$ is the relative permeability

$$\kappa_{rel}(x, \theta) = \begin{cases} 1; & \theta > \theta_* \\ \max \left\{ (\chi_l(x, \theta))^3, \epsilon_{\kappa_{min}} \right\}; & \theta \leq \theta_*, \end{cases} \quad (24b)$$

and η_{init} is the initial porosity, and $\epsilon_{\kappa_{min}} > 0$ is a constant [44]. For our numerical experiments, we choose $\epsilon_{\kappa_{min}} = 10^{-6}$ in (24b) [42,44]. We further assume μ_l to be a constant. The reader is referred to [45] for a comprehensive list of the different expressions of the permeability.

Note that the porosity is determined using (20d), which does not place any constraints on its maximum or minimum value. For practical purposes, we require $\eta \in (0, 1)$, but as $\eta \rightarrow 0$ or 1, from (24a) we see that $\kappa \rightarrow 0$ or ∞ , respectively. Hence, we make the following a-priori assumption on the porosity.

Assumption 2.4. The porosity $\eta : \Omega \times \mathbb{R} \rightarrow [0, 1]$ is uniformly bounded

$$0 < \eta_{\min} \leq \eta(x, t) \leq \eta_{\max} < 1, \quad \forall x \in \Omega, t > 0. \quad (25)$$

Using Assumption 2.4, we can establish the boundedness of the permeability given by (24a)

$$\kappa_{\min} = \kappa_{0,u} \epsilon_{\kappa_{min}} \frac{\eta_{\min}^3}{(1-\eta_{\min})^2} \frac{(1-\eta_{init})^2}{\eta_{init}^3}, \quad (26a)$$

$$\kappa_{\max} = \kappa_{0,u} \frac{\eta_{\max}^3}{(1-\eta_{\max})^2} \frac{(1-\eta_{init})^2}{\eta_{init}^3}. \quad (26b)$$

Moreover, the Lipschitz continuity assumption (23) is satisfied by (24a) due to the Lipschitz continuity of χ_l in θ [23], and the lower bound $\kappa_{rel} \geq \epsilon_{\kappa_{min}} > 0$.

Remark 2.3. In order to ensure Assumption 2.4 is satisfied, we place additional checks on the porosity in our numerical implementation. In practice, however, we observe that the variations in porosity are very small, and Assumption 2.4 is satisfied without any additional checks.

2.4.3. Elasticity parameters

The mechanical strength of frozen soils increases with decreasing temperature [1,4]. However, in literature, expressions for the dependence of elasticity parameters of frozen soil on temperature or water fraction are scarce. In fact, it is common to consider complex models beyond linear elasticity to simulate the complex mechanics in permafrost. For example, elasto-plastic models are used in [6,7,9,10] and viscoelasticity is considered in [8].

In this paper we consider only elastic properties, and need to identify an appropriate relationship for the Young's modulus E in non-isothermal soils. In literature, piecewise constant $E(\theta) = E_u, \forall \theta > \theta_*$, $E(\theta) = E_f, \forall \theta < \theta_*$ are frequently assumed, where E_u and E_f denote the Young's modulus of unfrozen and frozen soil, respectively. For example, in [7], $E_u = 1000$ [kPa] and $E_f = 5000$ [kPa] are considered for a sandy-silt soil. In [14], the authors consider $E_u = E_f = E \in [2, 20]$ [MPa], but with different values of E for soils under compression and tension. In turn, in [1](Pg. 129), affine expressions $E(\theta) = E_f(1 + a|\theta|)$, $\theta < 0$ are provided for temperatures $\theta \in [0, -10]$ [°C], and under pressures of 100 [kPa]. Similar expressions are provided in [4], with an emphasis on the effect of the pressures under which the data is obtained.

The Poisson's ratio is also affected by the temperature in frozen soils [1](Pg. 130) [4]. For example, it is reported that for frozen sands, ν decreases from 0.2 to 0.1 but no parametric model is given [1](Pg. 130). In experimental results reported in [4] we see $\nu \rightarrow 0.5$ as $\theta \rightarrow 0^-$, however no model is given.

In summary, following [1,4], we consider the following expression

$$E_0(x, \theta) = \begin{cases} E_u(x); & \theta > \theta_* \\ E_f(x) (1 + a(x)|\theta - \theta_*|); & \theta \leq \theta_*, \end{cases} \quad (27)$$

where $\{E_u, E_f, a\}$ [Pa] are soil specific constants.

However, the discontinuity of E_0 in (27) causes difficulties for our iterative solver, and its unboundedness causes difficulties for some theoretical estimates to be established. To address this, we consider a bounded regularization of $E_0(\theta) \approx E_\delta(\theta)$ defined as

$$E_\delta(\theta) = \begin{cases} E_u; & \theta > \theta_* \\ E_u + (E_f + E_f a \delta - E_u) \frac{(\theta_* - \theta)}{\delta}; & \theta \in [\theta_* - \delta, \theta_*] \\ E_f (1 + a|\theta - \theta_*|); & \theta \in [\bar{\theta}, \theta_* - \delta) \\ E_f (1 + a|\bar{\theta} - \theta_*|); & \theta < \bar{\theta}, \end{cases} \quad (28)$$

where δ [°C] > 0 is the regularization parameter and $\bar{\theta} < \theta_* - \delta$ is chosen to have a large absolute value. In practise $\bar{\theta}$ can be chosen outside the range of temperature values considered in our examples, thus we ignore it in practice.

We are now ready to make the following assumptions regarding the elastic soil parameters.

Assumption 2.5. The Young's modulus $E : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ and the Poisson's ratio $\nu : \Omega \rightarrow \mathbb{R}$ are uniformly bounded

$$0 < E_{\min} \leq E(x, \theta) \leq E_{\max} < \infty, \quad 0 < \nu_{\min} \leq \nu(x) \leq \nu_{\max} < 0.5, \quad \forall x \in \Omega, \theta \in \mathbb{R}. \quad (29)$$

Further, E is Lipschitz in θ , i.e., $\exists L_E > 0$ such that $\forall x \in \Omega, \theta_1, \theta_2 \in \mathbb{R}$

$$|E(x, \theta_2) - E(x, \theta_1)| \leq L_E |\theta_2 - \theta_1|. \quad (30)$$

Now from Assumption 2.5 and (7) it follows that $\lambda = \lambda(x, \theta)$ and $\mu = \mu(x, \theta)$ are bounded in $\Omega \times \mathbb{R}$ and Lipschitz in θ

$$0 < \lambda_{\min} \leq \lambda(x, \theta) \leq \lambda_{\max} < \infty, \quad 0 < \mu_{\min} \leq \mu(x, \theta) \leq \mu_{\max} < \infty, \quad \forall x \in \Omega, \theta \in \mathbb{R}. \quad (31)$$

We further denote the Lipschitz constants of λ and μ by L_λ and L_μ , respectively.

We see $E_0(\theta)$ is not Lipschitz, but from (28), the regularized expression $E_\delta(\theta)$ is, with

$$L_{E_\delta} = (E_f + E_f a \delta - E_u) \delta^{-1}. \quad (32)$$

3. Numerical scheme

We now provide details of our numerical scheme. We first rewrite the fully coupled TpHM model (20) in a mixed form as

$$\partial_t w + \nabla \cdot q_\theta = f, \quad w = \alpha(x, \theta, \eta), \quad (33a)$$

$$k^{-1} q_\theta + \nabla \theta - k^{-1} c_l \theta q_l = 0, \quad (33b)$$

$$\nabla \cdot (\zeta \partial_t u) + \nabla \cdot q_l - \gamma \partial_t [\eta(1 - \chi_l)] = g, \quad (33c)$$

$$\left[\frac{\kappa}{\mu_l} \right]^{-1} q_l + (\nabla p - \rho_l G \nabla D) = 0, \quad (33d)$$

$$-\nabla \cdot [\lambda \nabla \cdot u I + 2\mu \epsilon(u)] + \nabla p = l + \bar{\rho} G, \quad (33e)$$

$$\partial_t \eta - \nabla \cdot [(1 - \eta) \partial_t u] = 0, \quad (33f)$$

where q_θ in (33a)–(33b) represents the conductive and convective heat flux $q_\theta = -k \nabla \theta + c_l q_l \theta$.

For simplicity of the presentation of the scheme, we assume below homogeneous Dirichlet boundary conditions $\theta|_{\partial\Omega} = 0$ and $u|_{\partial\Omega} = 0, p|_{\partial\Omega} = 0$. We also assume that $w(x, 0) = w_{\text{init}}(x) \in L^2(\Omega)$ [22], $u(x, 0) = u_{\text{init}}(x) \in H^1(\Omega)^d$ [46](Pg. 148), and $\eta(x, 0) = \eta_{\text{init}} \in L^2(\Omega)$. Note that a different initial condition $\nabla \cdot u_{\text{init}} \in L^2(\Omega)$ is used in [36] for the well-posedness of (6).

The continuous in time variational formulation for (33) is as follows: $\forall t > 0$ we seek $\theta(t), w(t) \in M, q_\theta(t) \in X, u(t) \in V, p(t) \in M, q_l(t) \in X$, and $\eta(t) \in M$ such that

$$(\partial_t w, m) + (\nabla \cdot q_\theta, m) = (f, m), \quad \forall m \in M, \quad w = \alpha(x, \theta, \eta), \quad (34a)$$

$$(k^{-1} q_\theta, \psi) - (\theta, \nabla \cdot \psi) - (k(\theta)^{-1} c_l \theta q_l, \psi) = 0, \quad \forall \psi \in X, \quad (34b)$$

$$\begin{aligned} (\nabla \cdot (\zeta(\theta) \partial_t u), m) + (\nabla \cdot q_l, m) &= (\gamma \partial_t [\eta(1 - \chi_l)], m) \\ &+ (g, m), \quad \forall m \in M, \end{aligned} \quad (34c)$$

$$\left(\left[\frac{\kappa}{\mu_l} \right]^{-1} q_l, \psi \right) - (p, \nabla \cdot \psi) - (\rho_l G \nabla D, \psi_h) = 0, \quad \forall \psi \in X, \quad (34d)$$

$$a_u(\theta, u, \phi) - (p, \nabla \cdot \phi) = (l + \bar{\rho} G, \phi), \quad \forall \phi \in V, \quad (34e)$$

$$(\partial_t \eta, m) - (\nabla \cdot [(1 - \eta) \partial_t u], m) = 0, \quad \forall m \in M, \quad (34f)$$

where the bilinear form $a_u : \mathbb{R} \times V \times V \rightarrow \mathbb{R}$ is defined as

$$a_u(\theta, u, \phi) = \int_{\Omega} \lambda(x, \theta) (\nabla \cdot u) (\nabla \cdot \phi) + \int_{\Omega} 2\mu(x, \theta) \epsilon(u) : \epsilon(\phi), \quad (35)$$

with $A : B = \sum_{i,j} A_{i,j} B_{i,j}$ denoting the double dot product of tensors. We recall now Korn's inequality [47] which yields

$$\exists C_{Korn} > 0 : \|\epsilon(u)\|_2 \geq C_{Korn} \|u\|_{H^1}, \quad \forall u \in V. \quad (36)$$

Further, from (31) and (36), for a given $\theta \in \mathbb{R}$, the bilinear form $a_u(\theta, \cdot, \cdot)$ is continuous, symmetric, and coercive on $V \times V$, i.e., $\forall \theta \in \mathbb{R}, u \in V$

$$a_u(\theta, u, u) \geq \int_{\Omega} \mu(x, \theta) \epsilon(u) : \epsilon(u) \geq \mu_{\min} C_{Korn}^2 \|u\|_{H^1}^2. \quad (37)$$

Remark 3.1. Note that we implicitly assume that θ , u , and η are smooth enough such that the product $\zeta \partial_t u \in X$ in (34c) and $(1 - \eta) \partial_t u \in X$ in (34f).

Numerical discretization. We now present the numerical discretization of (34) working in the finite dimensional subspaces $M_h \subset M$, $X_h \subset X$, and $V_h \subset V$. For simplicity of exposition, we consider $d = 2$ spatial dimensions and we closely follow the notation in our work [22,23]. Let $\Omega \subset \mathbb{R}^2$ be covered by a rectangular grid $\mathcal{T}^h$ with N_{ω} cells $\omega_{i,j}$ such that $\bar{\Omega} = \cup_{i,j} \omega_{i,j}$. Each cell $\omega_{i,j}$ has width $h_{x,i}$ and $h_{y,j}$ in the x and y direction, respectively. We denote the midpoint of the cell $\omega_{i,j}$ by (x_i, y_j) and its four nodes by $(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}})$, $(x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}})$, $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$, and $(x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}})$, listed clockwise starting from the bottom left node. We also denote by $\gamma_{i-\frac{1}{2},j}$, $\gamma_{i,j+\frac{1}{2}}$, $\gamma_{i+\frac{1}{2},j}$, and $\gamma_{i,j-\frac{1}{2}}$ the four sides of $\omega_{i,j}$ listed clockwise starting from the left edge. We consider a time grid $t_n = t_{n-1} + \tau_n$, where $\tau_n > 0$ is the time step.

We denote by $M_h \subset M$ as the space of piecewise-constants (P0), by $X_h = RT_{[0]} \subset X$ the lowest-order Raviart–Thomas space (RT0), and by $V_h \subset V$ the space of continuous bilinear elements (Q1). We also denote by $(\cdot, \cdot)_h$ the use of the Trapezoidal-Midpoint quadrature for numerical integration [25,26]. We also denote $\|f\|_2 = (f, f)^{\frac{1}{2}}$ and $\|f\|_h = (f, f)_h^{\frac{1}{2}}$, $\forall f \in M$.

The basis functions of M_h are the indicator functions $\mathbf{1}_{\omega_{i,j}}$. For any $\theta_h, w_h \in M_h$, we denote by $\theta_{i,j} = \theta_h|_{\omega_{i,j}}$ and $w_{i,j} = w_h|_{\omega_{i,j}}$. The elements of X_h are vector-valued functions, and we denote the basis of X_h by $\psi_{i,j \pm \frac{1}{2}}$ for the first component and $\psi_{i,j \pm \frac{1}{2}}$. For any $q_h = (q_{h1}, q_{h2}) \in X_h$, we denote by $Q_{i \pm \frac{1}{2}, j} = q_{h1}|_{\gamma_{i \pm \frac{1}{2}, j}}$ and $Q_{i, j \pm \frac{1}{2}} = q_{h2}|_{\gamma_{i, j \pm \frac{1}{2}}}$. For the space V_h , the basis functions are denoted by $\phi_{i \pm \frac{1}{2}, j \pm \frac{1}{2}}$, and for any $u_h \in V_h$ we denote by $U_{i \pm \frac{1}{2}, j \pm \frac{1}{2}} = u_h|_{(x_{i \pm \frac{1}{2}}, y_{j \pm \frac{1}{2}})}$. Finally, we let the vectors Θ, W, P, Q_l and U collect the degrees of freedom of θ_h, w_h, p_h, q_l and u_h in their respective basis. We will also use an appropriate superscript to denote the time step and iterate, and we suppress the notation of spatial heterogeneity dependence of the parameters and relationships. For example, instead of $k = k(x_i, y_j, \theta_{i,j})$ we simply write $k = k(\theta_h)$.

We now state the fully discretized scheme for (34) based on operator splitting [24,48] for the thermal equation. Given $\theta_h^{n-1}, w_h^{n-1} \in M_h, p_h^{n-1} \in M_h, u_h^{n-1} \in V_h$, we seek $\theta_h^n, w_h^n \in M_h, q_{\theta_h}^n \in X_h, p_h^n \in M_h, q_l^n \in X_h, u_h^n \in V_h$ such that

$$(w_h^*, m_h) - (w_h^{n-1}, m_h) + \tau_n (\nabla \cdot F_h(\theta_h^{n-1}, q_l^n), m_h) = 0, \quad \forall m_h \in M_h, \quad (38a)$$

$$(w_h^n, m_h) + \tau_n (\nabla \cdot q_{\theta_h}^n, m_h) - (w_h^*, m_h) = \tau_n (f^n, m_h), \quad \forall m_h \in M_h, \quad (38b)$$

$$w_h^n = \alpha(\theta_h^n, \eta_h^n),$$

$$\left((k_h^n)^{-1} q_{\theta_h}^n, \psi_h \right)_h - (\theta_h^n, \nabla \cdot \psi_h) = 0, \quad \forall \psi_h \in X_h, \quad (38c)$$

$$(\zeta_h^n \nabla \cdot u_h^n, m_h) + \tau_n (\nabla \cdot q_l^n, m_h) - (\zeta_h^n \nabla \cdot u_h^{n-1}, m_h) = \gamma(\eta_h^n (\chi_l^{n-1} - \chi_l^n), m_h) + \tau_n (g^n, m_h), \quad \forall m_h \in M_h, \quad (38d)$$

$$\left(\left[\frac{\kappa_h^n}{\mu_l} \right]^{-1} q_l^n, \psi_h \right)_h - (p_h^n, \nabla \cdot \psi_h) - (\rho_l G \nabla D, \psi_h) = 0, \quad \forall \psi_h \in X_h, \quad (38e)$$

$$a_u(\theta_h^n, u_h^n, \phi_h) - (p_h^n, \nabla \cdot \phi_h) = (l^n + \bar{\rho} G, \phi_h), \quad \forall \phi_h \in V_h, \quad (38f)$$

$$(\eta_h^n, m_h) - (\eta_h^{n-1}, m_h) - \tau_n \left((1 - \eta_h^n) \nabla \cdot \left(\frac{u_h^n - u_h^{n-1}}{\tau_n} \right), m_h \right) = 0, \quad \forall m_h \in M_h, \quad (38g)$$

where $k_h^n = k(\theta_h^n)$, $\chi_{lh}^n = \chi_l(\theta_h^n)$, $\zeta_h^n = \zeta(\theta_h^n, \eta_h^n) = 1 - \gamma \eta_h^n (1 - \chi_{lh}^n)$, $\kappa_h^n = \kappa(\theta_h^n, \eta_h^n)$, $f^n = f(\cdot, t_n)$, $g^n = g(\cdot, t_n)$, $l^n = l(\cdot, t_n)$, $\bar{\rho}^n = \rho_l \eta_h^n \chi_{lh}^n + \rho_i \eta_h^n (1 - \chi_{lh}^n) + \rho_r (1 - \eta_h^n)$, and $F_h(\theta_h^{n-1}, q_{lh}^n) \in X_h$ is the upwind flux defined on each edge $\gamma_{i,j+\frac{1}{2}}$ as

$$F_h(\theta_h^{n-1}, q_{lh}^n)_{i,j+\frac{1}{2}} = \begin{cases} c_l \theta_{i,j}^{n-1} Q_{l,i,j+\frac{1}{2}}^n; & Q_{l,i,j+\frac{1}{2}}^n \geq 0 \\ c_l \theta_{i+1,j}^{n-1} Q_{l,i,j+\frac{1}{2}}^n; & Q_{l,i,j+\frac{1}{2}}^n < 0, \end{cases} \quad (39)$$

with the similar extension to edges $\{\gamma_{i,j+\frac{1}{2}}\}_{i,j}$.

Note that in (38a) and (38b) w_h^* denotes the intermediate enthalpy that is calculated explicitly using (38a). In practice (38a) and (38b) may be merged into one equation, but here we list them as separate equations for the sake of readability of the algorithm. Further in (38b)–(38c), the thermal flux $q_{\theta h}^n$ now only represents the conductive flux, as opposed to conductive and advective in (33a)–(33b).

Remark 3.2. Note that in (38d) since $\zeta_h^n \in M_h$, we have $\nabla \cdot (\zeta_h^n u_h^n) = \zeta_h^n \nabla \cdot u_h^n$ in each cell $\omega_{i,j} \in \mathcal{T}^h$. Similarly, with $\eta_h^n \in M_h$, in (38g) we approximate $\nabla \cdot [(1 - \eta_h^n)(u_h^n - u_h^{n-1})] = (1 - \eta_h^n) \nabla \cdot (u_h^n - u_h^{n-1})$. Ideally, an appropriate numerical flux, such as the upwind flux, should be used to handle the terms $\nabla \cdot (\zeta \partial_t u)$ and $\nabla \cdot [(1 - \eta) \partial_t u]$, but we handle these terms approximately.

The system (38) is implicit in time and nonlinear, with the many couplings and interdependencies given in Table 2. A monolithic approach to solve this system is possible but would require a computationally expensive and implementation intensive nonlinear solver. For example, the use of Newton's method would require the explicit calculation of the Jacobian of the system, which is inefficient due to the dependencies of the hydro-mechanical constitutive relationships on temperature. To work around this inefficiency, we consider a sequential approach and an iterative solver in Section 5 which allow to use the individual components of Tp and HM models implemented separately.

4. Literature review of computational schemes for TpHM and THM

In this section we review some relevant literature on numerical methods for TpHM models.

The works on TpHM are primarily from the geotechnical and geophysics literature and focus on the applications to realistic scenarios but do not provide or analyze the details of the computational models. In turn, computational mathematics literature is rather scarce and has focused so far on THM models not involving phase change.

There are many aspects of computational schemes that need to be addressed including well-posedness (solvability) of the discrete system, properties of approximations including conservation, stability, convergence of approximations and of iterative schemes, and efficiency of implementation. In Section 4.1 we overview the literature results on TpHM and THM with regard to these aspects.

The scarcity of literature on the analysis of TpHM models in literature makes it important to study the challenges associated with the thermal Tp and with the hydro-mechanical HM subproblems. This study helps to systematically build a robust solver for the fully coupled system. Since we provide an extensive study of Tp models in permafrost in [22,23], in this paper we focus on the hydro-mechanical models as part of TpHM; we review the relevant literature in Section 4.2. We provide a summary overview of some of the computational schemes used in literature in the Appendix in Table A.10.

4.1. Literature overview

Below we discuss several aspects of computational schemes considered in literature.

Discretization and approximation spaces. In applications literature, most computational schemes involve the use of Galerkin finite elements [8–11,14]. The discrete system is solved monolithically using a nonlinear solver, such as Newton's method [9,10] or Picard's method [14]. Other iterative approaches have also been used; for example in [13] a splitting is followed where the thermal subproblem is solved first followed by the hydro-mechanical subproblem.

In turn, the need for conservative schemes is recognized as crucial. In [15,19,27] conservative approaches for the flow (mixed finite elements) are combined with Galerkin approaches for the mechanics. In particular [27,46] offer the analysis of the use of P0-RT0-Q1 finite elements for HM.

Well-posedness. To our knowledge, no well-posedness results exist yet for the fully coupled system (20), but there is work towards this result, even if formulated under various assumptions. In [49], the semidiscrete Galerkin formulation is considered along with a nonlinear temperature dependent viscosity $\mu_l(\theta)$ and thermal conductivity $k(\theta)$: the authors prove the well-posedness of the semidiscrete formulation under boundedness and growth assumptions on the μ_l and k . For the fully implicit in time discretization of system (34) for thermo-poroelasticity models, the existence of a solution is shown in [15] in the absence of phase change. The authors consider thermal advection given by $c_l q_l \cdot \nabla \theta$, and further use an appropriate bounded Lipschitz “cut-off” operator $\tilde{M}$ to approximate $\tilde{M}(q_l) \cdot \nabla \theta \approx \nabla \theta \cdot q_l$. Moreover, physical parameters are considered to be dimensionless constants.

Stability. An important aspect of iterative schemes is stability. For example, in isothermal linear poroelasticity, such schemes may first solve the mechanics problem followed by the flow problem (undrained and drained split), or the other way around (fixed strain or stress split). The stability of these different schemes has been analyzed in [50] for the two-field formulation of (6) using P0-Q1 elements. The convergence and application of these schemes for (6) using P0-RT0-Q1 elements has also been analyzed in [51,52]. In [53], an iteratively coupled approach for the P0-RT0-Q1 scheme is analyzed for linear poroelasticity, where the

system is decoupled around the mean stress. In [54], the undrained and fixed stress approaches have been extended to include the thermal subproblem, and their stability has been analyzed using the finite volume method for thermo-hydrological flow and P1 elements for mechanics. Specifically, unconditional stability of solving the thermo-hydrological subproblem followed by the mechanical subproblem is proved.

Iterative schemes. Another class of iterative approaches are based on the linearization method called the L-scheme [55]; for example, the application of the L-scheme in proving the convergence of fixed stress splits using P0-RT0-Q1 elements has been analyzed for heterogeneous [56] and nonlinear [57] poroelasticity systems. In [15], the authors present and analyze multiple iteratively coupled schemes for (38) in the absence of phase transition. These iterative schemes are based on the L-scheme, and their convergence is proved.

Limitations of known schemes for TpHM models. The schemes mentioned above are reported efficient, but they do not directly apply for the TpHM model (33). For example, the scheme presented in [53] for hydro-mechanical systems is only convergent if the liquid is assumed to be compressible and its compressibility is large enough. Convergence issues regarding the incompressible assumption of liquid have also been noted in [54] for thermo-poroelasticity models. Moreover, solvers based on the L-schemes introduce a consistency error and are also only linearly convergent. The existing L-scheme approaches for thermo-poroelasticity have also been largely presented and analyzed for models with constant physical parameters, and for linear thermal scenarios, i.e., in the absence of phase transitions [15]. For TpHM models, however, there is an additional dependence of the hydro-mechanical parameters on the temperature and this precludes the extension of existing convergence results of iterative schemes when applied to TpHM models.

4.2. Challenges for hydro-mechanical problems

We discuss now some known difficulties of computational schemes for HM models. These may arise also in TpHM when the temperature is not fixed. Of particular interest to us are the challenges for the schemes based on P0-RT0-Q1 finite elements.

For isothermal hydro-mechanical systems in the incompressible liquid case, i.e., when $c_0 = 0$ in (6), spurious pressure oscillations are known to exist for low permeability κ or small time step size τ [30,31,58,59]. For the P0-RT0-Q1 discretization, these oscillations are linked to the incompatibility of Q1 and P0 spaces, that is, the violation of the inf-sup stability condition [28,29]. In [59], they have been linked to the deficiency of the monotonicity of the discretization. Another well-known challenge is Poisson locking [28,31] associated with the loss of coercivity of a_u in (35) when $\lambda \rightarrow \infty$. This leads to poor approximation of the displacement when coarse spatial meshes are used.

These challenges are well-studied, and are typically handled by an appropriate numerical discretization. For example, the use of discontinuous Galerkin [30] and Bernardi and Raugel [28] element for displacements is shown to eliminate spurious pressure oscillations as well as Poisson locking. In [31], a non-conforming approach using Mardal–Tai–Winther elements for displacement is considered, and the robustness with respect Poisson locking is demonstrated. In [29], a stabilized P0-RT0-Q1 approach is provided where the displacements are bubble-enriched; this approach is shown to be robust with respect to low permeability. In [60], suitable preconditioners are analyzed for the P0-RT0-Q1 scheme.

Path forward for HM in permafrost TpHM models. In this paper we show robustness of the P0-RT0-Q1 elements for the HM portion of TpHM in handling practical permafrost scenarios without any need for special techniques to avoid the aforementioned numerical artifacts. In particular, we demonstrate that unphysical oscillations may arise in dimensions $d \geq 2$, but that this happens for the time steps not relevant in practical scenarios.

5. Iterative solver

In this section we provide details of our iterative solver to seek a numerical solution to (38). At $t = t_n$, given $w_h^{n-1} \in M_h$, and $u_h^{n-1} \in V_h$, we take the following steps. We denote by superscript $n, (m)$ the iterate (m) at time step n .

Step 1: P0-P0 solver for Tp. First, we solve the thermal subproblem: given $q_{lh}^{n,(m-1)} \in X_h$, we calculate $w_h^{*,(m)} \in M_h$ as

$$(w_h^{*,(m)}, m_h) = (w_h^{n-1}, m_h) - \tau_n \left(\nabla \cdot F_h \left(q_{lh}^{n,(m-1)} \right), m_h \right), \quad \forall m_h \in M_h, \quad (40a)$$

Next we seek $w_h^{n,(m)}, \theta_h^n \in M_h$ and $q_{\theta h}^{n,(m)} \in X_h$ such that

$$(w_h^{n,(m)}, m_h) + \tau_n (\nabla \cdot q_{\theta h}^{n,(m)}, m_h) = (w_h^{*,(m)}, m_h) + \tau_n (f^n, m_h), \quad (40b)$$

$$\forall m_h \in M_h, \quad w_h^{n,(m)} = \alpha(\theta_h^{n,(m)}, \eta_h^{n-1}),$$

$$(\tilde{k}^{-1} q_{\theta h}^{n,(m)}, \psi_h)_h - (\theta_h^{n,(m)}, \nabla \cdot \psi_h) = 0, \quad \forall \psi_h \in X_h, \quad (40c)$$

where we use time-lagging and set $\tilde{k} = k_h^{n-1}$. Note that (40b)–(40c) gives rise to a nonlinear system of equations and here $w_h^{n,(m)}, \theta_h^{n,(m)}$ and $q_{\theta h}^{n,(m)}$ denotes the solution to (40b)–(40c) obtained after using the P0-P0 solver. The P0-P0 solver for the thermal implicit diffusion substep (40b)–(40c) is given in Appendix A.1, and involves a separate inner iteration at each iteration (m) .

Step 2: P0-RT0-Q1 solver for HM. Next we solve for the hydro-mechanical subproblem (38d)–(38f): given $\theta_h^{n,(m)} \in M_h$ after solving Step 1, we seek $p_h^{n,(m)} \in M_h$, $q_{lh}^{n,(m)} \in X_h$, and $u_h^{n,(m)} \in V_h$ such that

$$(\epsilon_h^{n,(m)} \nabla \cdot u_h^{n,(m)}, m_h) + \tau_n (\nabla \cdot q_{lh}^{n,(m)}, m_h) = (\epsilon_h^{n,(m)} \nabla \cdot u_h^{n-1}, m_h) + \quad (41a)$$

$$\gamma(\eta_h^{n-1}(\chi_l^{n-1} - \chi_l^{n(m)}), m_h) + \tau_n(g^n, m_h), \forall m_h \in M_h,$$

$$\left(\left[\frac{\kappa_h^{n(m)}}{\mu_l} \right]^{-1} q_l^{n(m)}, \psi_h \right)_h - (p_h^{n(m)}, \nabla \cdot \psi_h) = (\rho_l G \nabla D, \psi_h), \forall \psi_h \in X_h. \quad (41b)$$

$$a_u(\theta_h^{n(m)}, u_h^{n(m)}, \phi_h) - (p_h^{n(m)}, \nabla \cdot \phi_h) = (l^n + \bar{p}^{n(m)} G, \phi_h), \forall \phi_h \in V_h, \quad (41c)$$

where $\zeta_h^{n(m)} = \zeta(\theta_h^{n(m)}, \eta_h^{n-1})$.

Iteration. We iterate Step 1 and Step 2 till convergence (see Section 5.1 for details). We skip this step if we only aim to have a sequential scheme.

Step 3: Porosity update. We update the porosity as follows: we seek $\eta_h^n \in M_h$ such that

$$(\eta_h^n, m_h) = (\eta_h^{n-1}, m_h) + \tau_n \left((1 - \eta_h^{n-1}) \nabla \cdot \left(\frac{u_h^n - u_h^{n-1}}{\tau_n} \right), m_h \right), \forall m_h \in M_h. \quad (42)$$

The value η_h^n is needed in Step 1 and Step 2 to seek a solution at time step t_{n+1} . Here we also ensure that the bound $\eta_h^n \leq \eta_{max}$ required in Assumption 2.4 and discussed in Remark 2.3 holds. If not, we cut the time step τ_n , and redo Steps 1–2.

5.1. Implementation details for the iterative algorithm

The algorithm in Steps 1–2 presented by (40)–(41) is sequential: given $\{\theta_h^{n-1}, p_h^{n-1}, q_l^{n-1}, u_h^{n-1}\}$, it produces $\{\theta_h^n, p_h^n, q_l^n, u_h^n\}$. We can also iterate Steps 1–2 at each time step n as follows.

Iteration of Steps 1–2: We set the initial guess for $\{\theta_h^{n(0)}, q_l^{n(0)}, u_h^{n(0)}\}$ from $\{\theta_h^{n-1}, q_l^{n-1}, u_h^{n-1}\}$.

In each iteration $m = 1, 2, \dots$ we start with $\{\theta_h^{n(m-1)}, p_h^{n(m-1)}, q_l^{n(m-1)}, u_h^{n(m-1)}\}$ and after Steps 1–2 we obtain $\{\theta_h^{n(m)}, p_h^{n(m)}, q_l^{n(m)}, u_h^{n(m)}\}$.

We iterate until we reach an absolute or relative tolerance, i.e.,

$$\max \left\{ \|e_\theta^{(m)}\|_2, \|e_p^{(m)}\|_2, \|e_{q_l}^{(m)}\|_2, \|e_u^{(m)}\|_2 \right\} \leq \epsilon_{abs}, \text{ or,} \quad (43a)$$

$$\max \left\{ \frac{\|e_\theta^{(m)}\|_2}{\|\theta_h^{n(m)}\|_2}, \frac{\|e_p^{(m)}\|_2}{\|p_h^{n(m)}\|_2}, \frac{\|e_{q_l}^{(m)}\|_2}{\|q_l^{n(m)}\|_2}, \frac{\|e_u^{(m)}\|_2}{\|u_h^{n(m)}\|_2} \right\} \leq \epsilon_{rel}, \quad (43b)$$

where

$$\begin{aligned} e_\theta^{(m)} &= \theta_h^{n(m)} - \theta_h^{n(m-1)}, \quad e_p^{(m)} = p_h^{n(m)} - p_h^{n(m-1)}, \\ e_{q_l}^{(m)} &= q_l^{n(m)} - q_l^{n(m-1)}, \quad e_u^{(m)} = u_h^{n(m)} - u_h^{n(m-1)}. \end{aligned} \quad (44)$$

We also denote by

$$e_{q_\theta}^{(m)} = q_{\theta_h}^{n(m)} - q_{\theta_h}^{n(m-1)}, \quad e_w^{(m)} = w_h^{n(m)} - w_h^{n(m-1)}, \quad e_{w^*}^{(m)} = w_h^{*,(m)} - w_h^{*,(m-1)}. \quad (45)$$

If the number of iterations crosses a threshold of $m_{max} = 30$, we report no convergence. In practice, in (43) we choose $\epsilon_{abs} = 10^{-12}$ and $\epsilon_{rel} = 10^{-6}$.

The simulations in Section 6.3 are done using our implementation based on the library deal.II [61], the C++ software library used for finite element code modeling. In deal.II, we use the sparse direct solver UMFPACK which is part of the SuiteSparse library [62] when solving the linear system generated by (41).

5.2. Existence of solution at each iteration

We now prove that at each iteration (m) the system (40)–(41) is well-posed under a specific assumption. Since (42) and (40a) are explicit, the existence and uniqueness of $w_h^{*,(m)}$ and η_h^{n+1} follows trivially.

For the proof for the existence of a solution to (40b)–(40c) for the Stefan problem, see [22](Lemma 7.1). This proof can be adapted to the permafrost models considered here, since α is strictly monotone in θ [23].

We are left with the existence and uniqueness of a solution to (41). We first rewrite (41) in the absence of external sources and gravity in matrix–vector form

$$\begin{bmatrix} \mathcal{A}_u & -\mathcal{B}_{p,u} & 0 \\ Z\mathcal{B}_{p,u}^T & 0 & -\tau_n \mathcal{B}_{q_l,p} \\ 0 & \mathcal{B}_{q_l,p}^T & \widetilde{\mathcal{K}}_{q_l} \end{bmatrix} \begin{bmatrix} U^{n(m)} \\ P^{n(m)} \\ Q_l^{n(m)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{F} \\ 0 \end{bmatrix}, \quad (46)$$

where the matrix blocks in (46) are obtained as follows from the terms in

$$\begin{aligned} a_u(\theta_h^{n(m)}, u_h^{n(m)}, \phi_h) &\rightarrow \mathcal{A}_u U^{n(m)}, \\ (p_h^{n(m)}, \nabla \cdot \phi_h) &\rightarrow \mathcal{B}_{p,u} P^{n(m)} \end{aligned}$$

$$\begin{aligned} \left(\nabla \cdot q_{lh}^{n,(m)}, m_h \right) &\rightarrow -\mathcal{B}_{q_l,p} Q_l^{n,(m)}, \\ \left(\left[\frac{\kappa_h^{n,(m)}}{\mu_l} \right]^{-1} q_{lh}^{n,(m)}, \psi_h \right)_h &\rightarrow \widetilde{\mathcal{K}}_{q_l} Q_l^{n,(m)}, \end{aligned}$$

where a_u is the bilinear form (35), Z is the diagonal matrix with entries corresponding to $\zeta_h^{n,(m)}$, and the vector $\mathcal{F} = Z \mathcal{B}_{p,u}^T U^{n-1} + \gamma H^n (\chi_l(\Theta^{n-1}) - \chi_l(\Theta^{n,(m)}))$, with H^n being the diagonal matrix with entries corresponding to η_h^n . The matrix $\widetilde{\mathcal{K}}_{q_l}$ is diagonal with entries consisting of the transmissibilities due to the use of the trapezoidal-midpoint rule [26,63].

Lemma 5.1. Assume that $\gamma = 0$. Then, for any given $\theta_h^{n,(m)} \in M_h$, the system (41) has a unique solution $p_h^{n,(m)} \in M_h$, $q_{lh}^{n,(m)} \in X_h$, and $u_h^{n,(m)} \in V_h$.

Proof. Since a_u is symmetric and coercive, the matrix $\mathcal{A}_u$ is symmetric and invertible. Similarly, from Assumption 2.3, the diagonal matrix $\widetilde{\mathcal{K}}_{q_l}$ has positive entries and is invertible. Hence, we can rewrite (46) after eliminating $Q_l^{n,(m)}$ as

$$\begin{bmatrix} \mathcal{A}_u & -\mathcal{B}_{p,u} \\ Z \mathcal{B}_{p,u}^T & \tau_n \mathcal{B}_{q_l,p} \widetilde{\mathcal{K}}_{q_l}^{-1} \mathcal{B}_{q_l,p}^T \end{bmatrix} \begin{bmatrix} U^{n,(m)} \\ P^{n,(m)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{F} \end{bmatrix}. \quad (47)$$

We can further rewrite (47) as

$$\begin{bmatrix} \mathcal{A}_u & -\mathcal{B}_{p,u} \\ -\mathcal{B}_{p,u}^T & -Z^{-1} \left(\tau_n \mathcal{B}_{q_l,p} \widetilde{\mathcal{K}}_{q_l}^{-1} \mathcal{B}_{q_l,p}^T \right) \end{bmatrix} \begin{bmatrix} U^{n,(m)} \\ P^{n,(m)} \end{bmatrix} = \begin{bmatrix} 0 \\ -Z^{-1} \mathcal{F} \end{bmatrix}. \quad (48)$$

Note that the matrix in (48) is of a generalized nonsymmetric saddle point form. Under the assumption $\gamma = 0$, we have $Z = I$ (the identity matrix), and we can further eliminate $U^{n,(m)}$ from (48) to get

$$\left(\mathcal{B}_{p,u}^T \mathcal{A}_u^{-1} \mathcal{B}_{p,u} + \tau_n \mathcal{B}_{q_l,p} \widetilde{\mathcal{K}}_{q_l}^{-1} \mathcal{B}_{q_l,p}^T \right) P^{n,(m)} = \mathcal{F}. \quad (49)$$

Now in (49), the matrices $\left(\mathcal{B}_{p,u}^T \mathcal{A}_u^{-1} \mathcal{B}_{p,u} \right)$ and $\left(\mathcal{B}_{q_l,p} \widetilde{\mathcal{K}}_{q_l}^{-1} \mathcal{B}_{q_l,p}^T \right)$ are both square, and symmetric positive semidefinite and symmetric positive definite, respectively. Hence the matrix in (49) is symmetric positive definite and thus invertible, which establishes the existence and uniqueness of $P^{n,(m)}$, and consequently of $U^{n,(m)}$ and $Q_l^{n,(m)}$. $\square$

Remark 5.1. We acknowledge that the assumption of $\gamma = 0$ in Lemma 5.1 is rather limiting, and we plan to include the proof of the general case in a future work. For the case of $\gamma > 0$, Lemma 5.1 still holds if $I - Z$ is small enough compared to the smallest eigenvalue of $Y = \tau_n \mathcal{B}_{q_l,p} \widetilde{\mathcal{K}}_{q_l}^{-1} \mathcal{B}_{q_l,p}^T$. To see this, let $X = \mathcal{B}_{p,u}^T \mathcal{A}_u^{-1} \mathcal{B}_{p,u}$ and $C = ZX + Y$. Then, C is linear and thus Lipschitz. Further, rewriting $C = -(I - Z)X + X + Y$, and recalling that the entries of $I - Z$ are bounded pointwise below by 0 and above by $\gamma\eta$, C is strongly monotone if $\gamma\eta\|X\|_2$ is sufficiently small compared with the smallest eigenvalues of Y . Following this, its nonsingularity follows from [64](Thm. 5.1.4), and thus Lemma 5.1 holds in this case.

The preceding discussion shows the theoretical sensitivity of the nonsingularity of C to the value of γ . However, in practice, we have not observed C to be nonsingular for the physical value of $\gamma \approx 0.09$.

Corollary 5.1. Under the assumption $\gamma = 0$, the system (40)–(42) has a unique solution at each time step n and iteration (m) .

Remark 5.2. As in the derivation of (34), the results here are proven for homogeneous Dirichlet boundary conditions, but can be extended to the mixed boundary conditions, with Dirichlet conditions imposed on at least on some portion of $\partial\Omega$, for each of H and M problems.

5.3. Stability of time stepping scheme and convergence of the iterative solver.

We first address stability of the explicit steps and then prove convergence of our iterative solver from Section 5.1.

Stability of explicit steps. The implicit–explicit time stepping scheme (38a)–(38b) is stable as long as the following CFL condition is respected [48]

$$\tau_n \leq \frac{\min_{i,j} \{h_{x,i}, h_{y,j}\}}{\max \left\{ \sup |\partial_w F_1|, \sup |\partial_w F_2| \right\}}, \quad (50)$$

where $F = (F_1, F_2)$ is the convective flux $F(\theta) = c_l q_{lh}^n \theta$. Using $\theta = \alpha^{-1}(w)$, we get

$$\partial_w F_1(w) = \frac{1}{\alpha'(\alpha^{-1}(w))} c_l q_{lh1}^n, \quad \partial_w F_2(w) = \frac{1}{\alpha'(\alpha^{-1}(w))} c_l q_{lh2}^n. \quad (51)$$

Now we use the bounds $c_{\min} \leq |\alpha'| \leq c_{\max} + L\eta L_{\chi_l}$ which follow from the property proven in [23] (Lemma 2.1) that α is monotone with bounded derivatives. Here L_{χ_l} is the Lipschitz constant of χ_l , and $c_l \geq c_{\min}$ from 2.2. We get

$$\sup |\partial_w F_1| \geq \frac{c_{\min} q_{h1}^n}{c_{\max} + L\eta L_{\chi_l}}, \quad \sup |\partial_w F_2| \geq \frac{c_{\min} q_{h2}^n}{c_{\max} + L\eta L_{\chi_l}}. \quad (52)$$

Substituting (52) into (50) and using $\eta < 1$ from Assumption 2.4, we get

$$\tau_n \leq \left(\frac{c_{\max} + L\eta L_{\chi_l}}{c_{\min}} \right) \frac{\min_{i,j} \{h_{x,i}, h_{y,j}\}}{\max \left\{ \sup |q_{h1}^n|, \sup |q_{h2}^n| \right\}} \quad (53)$$

In practice the CFL condition (53) is not a significant restriction due to the low order of magnitude of the hydrological flux in permafrost scenarios.

For the porosity update (42) we first rewrite (20d) as

$$\partial_t(1 - \eta) + \nabla \cdot ((1 - \eta)\partial_t u) = 0. \quad (54)$$

and now check that the CFL-like condition involving $\partial_t u_h^n = (u_h^n - u_h^{n-1})/\tau_n$ holds

$$\bar{\tau}_n \leq \frac{\min_{i,j} \{h_{x,i}, h_{y,j}\}}{\max \left\{ \sup |\partial_t u_{h1}^n|, \sup |\partial_t u_{h2}^n| \right\}}, \quad (55)$$

If either of the conditions (53) or (55) is not satisfied, we reduce the time step and repeat Steps 1–2 until these hold. In practice, however, we observe neither of these conditions presents a significant restriction.

Convergence of iteration of Steps 1–2 from Section 5.1.

Our proof has similar setup as that in [15]. However, a direct application of the approach from [15] does not apply due to $E = E(\theta)$ in our model. We provide a proof which is similar but not identical to that in [15]. For the purpose of the proof, we consider the Eqs. (40)–(41) in Steps 1–2 to be in a dimensionless form.

We start by proving a simple estimate for the upwind flux (39). In what follows, for simplicity we assume a uniform square grid $\mathcal{T}^h$, i.e., $h_{x,i} = h_{y,j} = h$, $\forall i, j$. We also assume a uniform time step $\tau_n = \tau \forall n$.

Lemma 5.2. *Let $\theta_h \in M_h$. Then $\forall q_h, r_h \in X_h$ we have on each edge $\gamma_{i+\frac{1}{2},j}$*

$$\left| F_h(\theta_h, q_h)_{\gamma_{i+\frac{1}{2},j}} - F_h(\theta_h, r_h)_{\gamma_{i+\frac{1}{2},j}} \right| \leq c_l \|\theta_h\|_\infty \left| Q_{i+\frac{1}{2},j} - R_{i+\frac{1}{2},j} \right|, \quad (56)$$

where $Q_{i+\frac{1}{2},j} = q_h|_{\gamma_{i+\frac{1}{2},j}}$ and $R_{i+\frac{1}{2},j} = r_h|_{\gamma_{i+\frac{1}{2},j}}$. The estimate (56) holds for all edges $\gamma_{i-\frac{1}{2},j}, \gamma_{i,j\pm\frac{1}{2}}$ as well.

Proof. For simplicity of exposition, we present the proof for an interior edge $\gamma_{i+\frac{1}{2},j}$. Suppose $Q_{i+\frac{1}{2},j}, R_{i+\frac{1}{2},j} \geq 0$. Then it follows trivially that

$$\left| F_h(\theta_h, q_h)_{\gamma_{i+\frac{1}{2},j}} - F_h(\theta_h, r_h)_{\gamma_{i+\frac{1}{2},j}} \right| = c_l \left| Q_{i+\frac{1}{2},j} \Theta_{i,j} - R_{i+\frac{1}{2},j} \Theta_{i,j} \right| \leq c_l \max\{|\Theta_{i,j}|\} \left| Q_{i+\frac{1}{2},j} - R_{i+\frac{1}{2},j} \right|. \quad (57)$$

Now suppose $Q_{i+\frac{1}{2},j} > 0$ and $R_{i+\frac{1}{2},j} < 0$. Then

$$\begin{aligned} \left| F_h(\theta_h, q_h)_{\gamma_{i+\frac{1}{2},j}} - F_h(\theta_h, r_h)_{\gamma_{i+\frac{1}{2},j}} \right| &= c_l \left| Q_{i+\frac{1}{2},j} \Theta_{i,j} - R_{i+\frac{1}{2},j} \Theta_{i+1,j} \right| \\ &\leq c_l |\Theta_{i,j}| \left| Q_{i+\frac{1}{2},j} \right| + c_l |\Theta_{i+1,j}| \left| R_{i+\frac{1}{2},j} \right| \leq c_l \max\{|\Theta_{i,j}|, |\Theta_{i+1,j}|\} \left(\left| Q_{i+\frac{1}{2},j} \right| + \left| R_{i+\frac{1}{2},j} \right| \right) \\ &\leq c_l \max\{|\Theta_{i,j}|\} \left(Q_{i+\frac{1}{2},j} + (-R_{i+\frac{1}{2},j}) \right) = c_l \max\{|\Theta_{i,j}|\} \left| Q_{i+\frac{1}{2},j} - R_{i+\frac{1}{2},j} \right| \end{aligned} \quad (58)$$

Similar steps can be followed for the remaining cases $\gamma_{i-\frac{1}{2},j}$ and $\gamma_{i,j\pm\frac{1}{2}}$. This proves the result. $\square$

We next prove an auxiliary intermediate result.

Lemma 5.3. *At time step n and for each iteration (m) , the following estimate holds*

$$\left| \left(e_{w^*}^{(m)}, e_{\theta}^{(m)} \right) \right| \leq \frac{\tau \epsilon_1}{2} \left\| e_{\theta}^{(m)} \right\|_2^2 + \frac{\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2}{h^2 \epsilon_1} \left(e_{q_l}^{(m-1)}, e_{q_l}^{(m-1)} \right)_h, \quad \forall \epsilon_1 > 0. \quad (59)$$

Proof. Taking the difference of the consecutive iterates in (40a) we get

$$\left(e_{w^*}^{(m)}, m_h \right) = -\tau \left(\nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{l_h}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{l_h}^{n,(m-2)} \right) \right], m_h \right). \quad (60)$$

Using $m_h = e_{\theta}^{(m)}$ in (60) we get after taking using the Cauchy–Schwarz and Young inequality

$$\left| \left(e_{w^*}^{(m)}, e_{\theta}^{(m)} \right) \right| = \left| \tau \left(\nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{l_h}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{l_h}^{n,(m-2)} \right) \right], e_{\theta}^{(m)} \right) \right| \quad (61)$$

$$\leq \frac{\tau \epsilon_1}{2} \|e_\theta^{(m)}\|_2^2 + \frac{\tau}{2\epsilon_1} \left\| \nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-2)} \right) \right] \right\|_2^2,$$

for any $\epsilon_1 > 0$. Now on each cell $\omega_{i,j}$ we have

$$\begin{aligned} & \left| \nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-2)} \right) \right] \right| \\ & \leq h^{-1} \sum_{f \in \partial \omega_{i,j}} \left| F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-1)} \right) |_f - F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-2)} \right) |_f \right| \\ & \leq c_l \|\theta_h^{n-1}\|_\infty h^{-1} \sum_{f \in \partial \omega_{i,j}} \left| q_{lh}^{n,(m-1)} |_f - q_{lh}^{n,(m-2)} |_f \right| \\ & \leq 2c_l \|\theta_h^{n-1}\|_\infty h^{-1} \left(\sum_{f \in \partial \omega_{i,j}} \left| q_{lh}^{n,(m-1)} |_f - q_{lh}^{n,(m-2)} |_f \right|^2 \right)^{\frac{1}{2}}, \end{aligned} \quad (62)$$

where we have used Lemma 5.2 in the second last inequality, and in the last inequality we have used the arithmetic-quadratic mean inequality. Thus, we get from (62)

$$\begin{aligned} & \left\| \nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-2)} \right) \right] \right\|_2^2 \\ & = \sum_{\omega_{i,j}} \left\| \nabla \cdot \left[F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-1)} \right) - F_h \left(\theta_h^{n-1}, q_{lh}^{n,(m-2)} \right) \right] \right\|_2^2 h^2 \\ & \leq 4c_l^2 \|\theta_h^{n-1}\|_\infty^2 h^{-2} \sum_{\omega_{i,j}} \sum_{f \in \partial \omega_{i,j}} \left| q_{lh}^{n,(m-1)} |_f - q_{lh}^{n,(m-2)} |_f \right|^2 h^2 \\ & = 2c_l^2 \|\theta_h^{n-1}\|_\infty^2 h^{-2} \left(e_{q_l}^{(m-1)}, e_{q_l}^{(m-1)} \right)_h. \end{aligned} \quad (63)$$

Substituting (63) into (61) we get

$$\left| \left(e_{w^*}^{(m)}, e_\theta^{(m)} \right) \right| \leq \frac{\tau \epsilon_1}{2} \|e_\theta^{(m)}\|_2^2 + \frac{\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2}{h^2 \epsilon_1} \|e_{q_l}^{(m-1)}\|_h^2, \quad (64)$$

which proves the result. $\square$

We now proceed to the convergence proof of the iterative solver.

Theorem 5.1. Assume the following.

- (1) The parameter $\gamma = 0$ and $G = 0$.
- (2) The volumetric strain and strain tensors are uniformly bounded at each time step, i.e.,

$$\left\| \nabla \cdot u_h^{n,(m)} \right\|_\infty, \left\| e \left(u_h^{n,(m)} \right) \right\|_\infty \leq M_u, \quad \forall m, \quad (65)$$

for some $M_u > 0$.

- (3) The hydrological fluxes are uniformly bounded at each time step, i.e.,

$$\|q_{lh}^{n,(m)}\|_\infty \leq M_{q_l}, \quad \forall m, \quad (66)$$

for some $M_{q_l} > 0$.

Then, the iterative solver from Section 5.1 iterating Steps 1–2 (40)–(41) converges provided

$$c_{\min} - \frac{2\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2 \kappa_{\max}}{h^2 \mu_l} - \frac{\tau \mu_l M_{q_l}^2 L_{\kappa^{-1}}^2 \kappa_{\max}}{2} - \frac{M_u^2 L_\lambda^2}{4\lambda_{\min}} - \frac{M_u^2 L_\mu^2}{2\mu_{\min}} > 0. \quad (67)$$

Proof. We take difference of the consecutive iterates in (40b)–(40c) to get

$$(e_w^{(m)}, m_h) + \tau (\nabla \cdot e_{q_\theta}^{(m)}, m_h) = (e_{w^*}^{(m)}, m_h), \quad (68a)$$

$$(\tilde{k}^{-1} e_{q_\theta}^{(m)}, \psi_h)_h = (e_\theta^{(m)}, \nabla \cdot \psi_h). \quad (68b)$$

Taking $m_h = e_\theta^{(m)}$ in (68a) and $\psi_h = e_{q_\theta}^{(m)}$ in (68b), and adding the resulting equations we get

$$(e_w^{(m)}, e_\theta^{(m)}) + \tau (\tilde{k}^{-1} e_{q_\theta}^{(m)}, e_{q_\theta}^{(m)})_h = (e_{w^*}^{(m)}, e_\theta^{(m)}). \quad (69)$$

We now use the monotonicity of α [23](Lemma 2.1), [Assumption 2.2](#), and the estimate given by [Lemma 5.3](#) to get

$$c_{\min} \|e_{\theta}^{(m)}\|_2^2 + \tau k_{\max}^{-1} (e_{q_{\theta}}^{(m)}, e_{q_{\theta}}^{(m)})_h \leq \frac{\tau \epsilon_1}{2} \|e_{\theta}^{(m)}\|_2^2 + \frac{\tau c_l^2 \|\theta_h^{n-1}\|_{\infty}^2}{h^2 \epsilon_1} \|e_{q_l}^{(m-1)}\|_h^2, \quad (70)$$

where we choose $\epsilon_1 > 0$ later in the proof.

The difference of the consecutive iterates of (41) gives

$$(\nabla \cdot e_u^{(m)}, m_h) + \tau (\nabla \cdot e_{q_l}^{(m)}, m_h) = 0, \quad (71a)$$

$$\left(\left[\frac{\kappa^{n,(m)}}{\mu_l} \right]^{-1} q_l^{n,(m)} - \left[\frac{\kappa^{n,(m-1)}}{\mu_l} \right]^{-1} q_l^{n,(m-1)}, \psi_h \right)_h = (e_p^{(m)}, \nabla \cdot \psi_h), \quad (71b)$$

$$\begin{aligned} & \int_{\Omega} (\lambda^{n,(m)} \nabla \cdot u^{n,(m)} - \lambda^{n,(m-1)} \nabla \cdot u^{n,(m-1)}) \nabla \cdot \phi_h \\ & + \int_{\Omega} (\mu^{n,(i)} \epsilon(u_h^{n,(m)}) - \mu^{n,(m-1)} \epsilon(u_h^{n,(m-1)})) : \epsilon(\phi_h) = (e_p^{(i)}, \nabla \cdot \phi_h), \end{aligned} \quad (71c)$$

where $\lambda^{n,(m)} = \lambda(\theta_h^{n,(m)})$ and $\mu^{n,(m)} = \mu(\theta_h^{n,(m)})$. Choosing $m_h = e_p^{(m)}$ in (71a), $\psi_h = e_{q_l}^{(m)}$ in (71b), and $\phi_h = e_u^{(m)}$ in (71c), and adding the equations we get

$$\begin{aligned} & \tau \left(\left[\frac{\kappa^{n,(m)}}{\mu_l} \right]^{-1} q_l^{n,(m)} - \left[\frac{\kappa^{n,(m-1)}}{\mu_l} \right]^{-1} q_l^{n,(m-1)}, e_{q_l}^{(m)} \right)_h \\ & + \int_{\Omega} (\lambda^{n,(m)} \nabla \cdot u^{n,(m)} - \lambda^{n,(m-1)} \nabla \cdot u^{n,(m-1)}) \nabla \cdot e_u^{(m)} \\ & + 2 \int_{\Omega} (\mu^{n,(m)} \epsilon(u_h^{n,(m)}) - \mu^{n,(m-1)} \epsilon(u_h^{n,(m-1)})) : \epsilon(e_u^{(m)}) = 0. \end{aligned} \quad (72)$$

After some algebraic manipulations (72) can be rewritten as

$$\begin{aligned} & \tau \left(\left[\frac{\kappa^{n,(m)}}{\mu_l} \right]^{-1} e_{q_l}^{(m)}, e_{q_l}^{(m)} \right)_h + (\lambda^{n,(m)} \nabla \cdot e_u^{(m)}, \nabla \cdot e_u^{(m)}) + 2 (\mu^{n,(m)} \epsilon(e_u^{(m)}), \epsilon(e_u^{(m)})) = \\ & - \tau \left(\left(\left[\frac{\kappa^{n,(m)}}{\mu_l} \right]^{-1} - \left[\frac{\kappa^{n,(m-1)}}{\mu_l} \right]^{-1} \right) q_l^{n,(m-1)}, e_{q_l}^{(m)} \right)_h \\ & - (\lambda^{n,(m)} - \lambda^{n,(m-1)}) \nabla \cdot u_h^{n,(m-1)}, \nabla \cdot e_u^{(m)} \\ & - 2 (\mu^{n,(m)} - \mu^{n,(m-1)}) \epsilon(u_h^{n,(m-1)}), \epsilon(e_u^{n,(m)}) \). \end{aligned} \quad (73)$$

Now, using [Assumptions 2.3](#) and [2.5](#), the LHS of (73) is non-negative and can be bounded from below by

$$\tau \left[\frac{\kappa_{\max}}{\mu_l} \right]^{-1} \|e_{q_l}^{(m)}\|_h^2 + \lambda_{\min} \|\nabla \cdot e_u^{(m)}\|_2^2 + 2\mu_{\min} \|\epsilon(e_u^{(m)})\|_2^2. \quad (74)$$

The absolute value of the RHS of (73) can be bounded from above from [Assumption 2.5](#), [Assumption 2.3](#), and the uniform bounds (65)–(66) by

$$\begin{aligned} & \tau \mu_l M_{q_l} L_{\kappa^{-1}} \left(e_{q_l}^{(m)}, e_{q_l}^{(m)} \right)_h^{\frac{1}{2}} \|e_{\theta}^{(m)}\|_2 \\ & + M_u L_{\lambda} \|\nabla \cdot e_u^{(m)}\|_2 \|e_{\theta}^{(m)}\|_2 + 2M_u L_{\mu} \|\epsilon(e_u^{(m)})\|_2 \|e_{\theta}^{(m)}\|_2. \end{aligned} \quad (75)$$

Using Young's inequality for the first term in (75) we get

$$\tau \mu_l M_{q_l} L_{\kappa^{-1}} \left(e_{q_l}^{(m)}, e_{q_l}^{(m)} \right)_h^{\frac{1}{2}} \|e_{\theta}^{(m)}\|_2 \leq \tau \mu_l M_{q_l} L_{\kappa^{-1}} \left(\frac{\epsilon_2}{2} \|e_{q_l}^{(m)}\|_h^2 + \frac{1}{2\epsilon_2} \|e_{\theta}^{(m)}\|_2^2 \right). \quad (76)$$

Finally, adding (70) to (73) and using the estimates (74), (75) and (76) we get

$$\begin{aligned} & c_{\min} \|e_{\theta}^{(m)}\|_2^2 + \tau k_{\max}^{-1} \|e_{q_{\theta}}^{(m)}\|_h^2 + \tau \left[\frac{\kappa_{\max}}{\mu_l} \right]^{-1} \|e_{q_l}^{(m)}\|_h^2 \\ & + \lambda_{\min} \|\nabla \cdot e_u^{(m)}\|_2^2 + \mu_{\min} \|\epsilon(e_u^{(m)})\|_2^2 \\ & \leq \frac{\tau \epsilon_1}{2} \|e_{\theta}^{(m)}\|_2^2 + \frac{\tau c_l^2 \|\theta_h^{n-1}\|_{\infty}^2}{h^2 \epsilon_1} \|e_{q_l}^{(m-1)}\|_h^2 \\ & + \tau \mu_l M_{q_l} L_{\kappa^{-1}} \left(\frac{\epsilon_2}{2} \|e_{q_l}^{(m)}\|_h^2 + \frac{1}{2\epsilon_2} \|e_{\theta}^{(m)}\|_2^2 \right) \\ & + M_u L_{\lambda} \|\nabla \cdot e_u^{(m)}\|_2 \|e_{\theta}^{(m)}\|_2 + M_u L_{\mu} \|\epsilon(e_u^{(m)})\|_2 \|e_{\theta}^{(m)}\|_2 \end{aligned}$$

Rearranging terms in (77) and completing the squares we get

$$\begin{aligned}
& \left(c_{\min} - \frac{\tau \epsilon_1}{2} - \frac{\tau \mu_l M_{q_l} L_{\kappa^{-1}}}{2 \epsilon_2} - \frac{M_u^2 L_\lambda^2}{4 \lambda_{\min}} - \frac{M_u^2 L_\mu^2}{2 \mu_{\min}} \right) \|e_\theta^{(m)}\|_2^2 + \tau k_{\max}^{-1} \|e_{q_\theta}^{(m)}\|_h^2 \\
& + \tau \left(\left[\frac{\kappa_{\max}}{\mu_l} \right]^{-1} - \mu_l M_{q_l} L_{\kappa^{-1}} \frac{\epsilon_2}{2} \right) \|e_{q_l}^{(m)}\|_h^2 \\
& + \left(\sqrt{\lambda_{\min}} \|\nabla \cdot e_u^{(m)}\|_2 - \frac{M_u L_\lambda}{2 \sqrt{\lambda_{\min}}} \|e_\theta^{(m)}\|_2 \right)^2 + 2 \left(\sqrt{\mu_{\min}} \|\epsilon(e_u^{(m)})\|_2 - \frac{M_u L_\mu}{2 \sqrt{\mu_{\min}}} \|e_\theta^{(m)}\|_2 \right)^2 \\
& \leq \frac{\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2}{h^2 \epsilon_1} \|e_{q_l}^{(m)}\|_h^2.
\end{aligned} \tag{77}$$

We now choose $\epsilon_1 = (4c_l^2 \|\theta_h^{n-1}\|_\infty^2 \kappa_{\max})(h^2 \mu_l)^{-1}$, $\epsilon_2 = (M_{q_l} L_{\kappa^{-1}} \kappa_{\max})^{-1}$ to get

$$\begin{aligned}
& \left(c_{\min} - \frac{2\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2 \kappa_{\max}}{h^2 \mu_l} - \frac{\tau \mu_l M_{q_l}^2 L_{\kappa^{-1}}^2 \kappa_{\max}}{2} - \frac{M_u^2 L_\lambda^2}{4 \lambda_{\min}} - \frac{M_u^2 L_\mu^2}{2 \mu_{\min}} \right) \|e_\theta^{(m)}\|_2^2 + \tau k_{\max}^{-1} \|e_{q_\theta}^{(m)}\|_h^2 \\
& + \frac{\tau}{2} \left[\frac{\kappa_{\max}}{\mu_l} \right]^{-1} \|e_{q_l}^{(m)}\|_h^2 + \left(\sqrt{\lambda_{\min}} \|\nabla \cdot e_u^{(m)}\|_2 - \frac{M_u L_\lambda}{2 \sqrt{\lambda_{\min}}} \|e_\theta^{(m)}\|_2 \right)^2 \\
& + 2 \left(\sqrt{\mu_{\min}} \|\epsilon(e_u^{(m)})\|_2 - \frac{M_u L_\mu}{2 \sqrt{\mu_{\min}}} \|e_\theta^{(m)}\|_2 \right)^2 \leq \frac{\tau}{4} \left[\frac{\kappa_{\max}}{\mu_l} \right]^{-1} \|e_{q_l}^{(m)}\|_h^2.
\end{aligned} \tag{78}$$

Assuming that

$$c_{\min} - \frac{2\tau c_l^2 \|\theta_h^{n-1}\|_\infty^2 \kappa_{\max}}{h^2 \mu_l} - \frac{\tau \mu_l M_{q_l}^2 L_{\kappa^{-1}}^2 \kappa_{\max}}{2} - \frac{M_u^2 L_\lambda^2}{4 \lambda_{\min}} - \frac{M_u^2 L_\mu^2}{2 \mu_{\min}} > 0, \tag{79}$$

we have from (78),

$$\|e_{q_l}^{(m)}\|_h^2 \leq \frac{1}{2} \|e_{q_l}^{(m-1)}\|_h^2, \tag{80}$$

and thus

$$\|e_{q_l}^{(m)}\|_h \rightarrow 0 \text{ as } m \rightarrow \infty, \tag{81}$$

and there is a limit of the sequence $\{q_l^{n,(m)}\}_m$ which we call $q_l^n \in X_h$. Thus the right hand side of (77) vanishes in the limit, and thus we obtain existence of the limits of $\theta_h^{n,(m)}$, $u_h^{n,(m)}$ and $p_h^{n,(m)}$, which we call θ_h^n , u_h^n , and p_h^n , respectively. $\square$

Before we proceed to the numerical experiments, we make a few remarks. The quantity in (67) is considered in its dimensionless form and we show that the estimate (67) is satisfied for practical permafrost scenarios, however, we observe that in our numerical experiments the maximum number of iterations required by the solver increases with a decreasing time step, even if the average number of iterations decreases with smaller time step. The first feature seems contrary to what the estimate (67) dictates, but is perhaps rather due to the enhanced coupling enabled for smaller time steps.

6. Numerical experiments

In this section, we verify the basic properties of our algorithm; this is done in Section 6.1.

Next we apply our solver to practical permafrost scenarios. We consider practical one-dimensional soil consolidation scenarios using homogeneous and heterogeneous domains; this is done in Section 6.2. Through these examples we demonstrate the robustness of the solver. We also study the role of regularization of Young's modulus $E = E(\theta)$. We also study the importance of $\rho_l \neq \rho_i$.

Finally, in Section 6.3 we consider a two-dimensional isothermal soil consolidation example, where we demonstrate the robustness of our solver with respect to numerical artifacts such as nonphysical pressure oscillations or Poisson locking.

6.1. Order of convergence

We start with a convergence study to obtain the order of convergence of our fully discrete scheme by adapting an example from [15]. Our example features a known analytic solution, is not connected to any physical scenario, and we do not consider any particular physical units.

Example 6.1 (Order of convergence in the absence of phase transition with manufactured solutions). Let $x \in \Omega = (0, 1)$, $t > 0$, and assume that the temperature, pressure, and displacement profiles are given by

$$\theta(x, t) = tx(1 - x) + 1, \tag{82a}$$

Table 3
Physical parameters in Example 6.1.

Parameter	Value
$c_{ph}, ph \in \{l, i, r\}$	1
L	0
$k_{ph}, ph \in \{l, i, r\}$	1
η_{init}	0.5
κ_0	1
$\rho_{ph}, ph \in \{l, i, r\}$	1
γ	0
G	0

Table 4

Results for Example 6.1 showing errors and convergence rates obtained using our scheme and iterative solver.

h	τ	$\ \theta_{err}\ _{\infty,2}$	$\ \theta_{err}\ _{\infty,1}$	$\ \theta_{err}\ _{2,2}$
2×10^{-2}	1×10^{-1}	1.3047×10^{-3}	9.9090×10^{-4}	8.6609×10^{-4}
1×10^{-2}	5×10^{-2}	6.1370×10^{-4}	4.7019×10^{-4}	3.9582×10^{-4}
5×10^{-3}	2.5×10^{-2}	2.9751×10^{-4}	2.2947×10^{-4}	1.8906×10^{-4}
Rate		1.06	1.05	1.09
h	τ	$\ p_{err}\ _{\infty,2}$	$\ p_{err}\ _{\infty,1}$	$\ p_{err}\ _{2,2}$
2×10^{-2}	1×10^{-1}	6.8927×10^{-5}	5.3755×10^{-5}	4.1168×10^{-5}
1×10^{-2}	5×10^{-2}	2.5580×10^{-5}	2.1143×10^{-5}	1.6728×10^{-5}
5×10^{-3}	2.5×10^{-2}	1.2053×10^{-5}	1.0319×10^{-5}	7.8455×10^{-6}
Rate		1.25	1.19	1.19
h	τ	$\ q_{l, err}\ _{\infty,2}$	$\ q_{l, err}\ _{\infty,1}$	$\ q_{l, err}\ _{2,2}$
2×10^{-3}	1×10^{-1}	2.5407×10^{-4}	2.2274×10^{-4}	1.3625×10^{-4}
1×10^{-2}	5×10^{-2}	7.6339×10^{-5}	5.8784×10^{-5}	4.2855×10^{-5}
5×10^{-3}	2.5×10^{-2}	2.8702×10^{-5}	2.1753×10^{-5}	1.7291×10^{-5}
Rate		1.57	1.67	1.48
h	τ	$\ u_{err}\ _{\infty,2}$	$\ u_{err}\ _{\infty,1}$	$\ u_{err}\ _{2,2}$
2×10^{-2}	1×10^{-1}	1.2029×10^{-4}	1.0883×10^{-4}	8.4838×10^{-5}
1×10^{-2}	5×10^{-2}	3.7250×10^{-5}	3.0097×10^{-5}	2.6039×10^{-5}
5×10^{-3}	2.5×10^{-2}	1.3951×10^{-5}	1.0389×10^{-5}	9.6545×10^{-6}
Rate		1.55	1.69	1.56

$$u(x, t) = tx(1 - x) + 1, \quad (82b)$$

$$p(x, t) = tx(1 - x) + 1, \quad (82c)$$

We consider the Young's modulus and the permeability as

$$E(x, \theta) = \frac{5\theta^2}{6}, \quad v(x, \theta) = 0.25, \quad \kappa(x, \theta) = \theta^2, \quad \forall x \in \Omega, \quad \theta \in \mathbb{R}. \quad (83)$$

Consider other data listed in Table 3; in addition, assume $\eta = \eta_{init} = 0.5$ in the thermal model. Further, take $\chi_l = 1$ and assume no phase transition. We consider homogeneous Dirichlet boundary conditions for all $t > 0$

$$\theta(0, t) = \theta(1, t) = 1, \quad (84a)$$

$$p(0, t) = p(1, t) = 1, \quad (84b)$$

$$u(0, t) = u(1, t) = 1, \quad (84c)$$

and initial conditions $\theta_{init} \equiv 1$, and $u_{init} \equiv 1$. The sources are calculated accordingly

$$f(x, t) = x(1 - x) + 2t - 3[tx(1 - x) + 1]^2[t(1 - 2x)]^2 + (2t)[tx(1 - x) + 1]^3, \quad (85a)$$

$$g(x, t) = (1 - 2x) - 2tx(1 - x)[t(1 - 2x)]^2 + (2t)[tx(1 - x) + 1]^2, \quad (85b)$$

$$l(x, t) = -2tx(1 - x)[t(1 - 2x)]^2 + (2t)[tx(1 - x) + 1]^2 + t(1 - 2x). \quad (85c)$$

We let $t \in (0, 1)$ and compute the order of convergence of θ, p, u, q_l in the $\|\cdot\|_{\infty,2}$, $\|\cdot\|_{\infty,1}$, and $\|\cdot\|_{2,2}$, with results tabulated in Table 4.

We see that we obtain order at least 1 for θ, p, q_l , and u .

6.2. One-dimensional soil consolidation

We now test robustness of our solver in scenarios involving subsidence due to the permafrost thaw. Such scenarios have applications for structures constructed on ground underlain with permafrost [1,4,9]. In our examples, we do not include gravity

Table 5
Physical parameters used in Example 6.2. Units are as in Table 1.

Parameter	Value	Reference
c_l	4.19×10^6	[65](Section 4)
c_i	1.90×10^6	[65](Section 4)
c_r	2.385×10^6	[43](Pg. 90)
L	306×10^6	[65](Section 4)
k_l	0.58	[65](Section 4)
k_i	2.30	[65](Section 4)
k_r	2.92	[43](Pg. 90)
SFC parameters in (21): $\{b, \theta_s, \chi_{res}\} = \{0.2, 0, 0.2\}$		
E	$E_u = 20 \times 10^6$ $E_f = 400 \times 10^6, a = 3.5$ $\delta = 0.1$ [°C]	[66](Pg. 407, Table 14.2) [1](Pg. 129)
ν	0.30	[66](Pg. 407, Table 14.2)
η_{init}	0.45	[39](Pg. 74)
$\kappa_{0,u}$	10^{-13}	[39](Pg. 119)
μ_l	1.0005×10^{-3}	[40] at 20 [°C]
β_l	0 (incompressible assumption)	
ρ_l	1000	[43](Pg. 90)
ρ_i	917	[43](Pg. 90)
ρ_r	2650	[43](Pg. 90)

terms to avoid having to adjust boundary and initial conditions: we focus on consolidation due to the external traction alone. We start with a homogeneous soil example and next move to a heterogeneous case.

6.2.1. Homogeneous domain

Here the goal is to test the robustness of our solver with respect to different spatial and temporal grid sizes, and depending on the degree of the regularization of Young's modulus $E(\theta)$ measured by its Lipschitz constant. We also compare the iterative vs the sequential approach for (40)–(41).

Example 6.2 (One-dimensional homogeneous soil consolidation). Let $\Omega = (0, 1)$ [m] be occupied by a soil with parameters as in Table 5. For the thermal component Tp, we consider the boundary and initial conditions

$$\theta(0, t) = 10, \quad \theta(1, t) = -5, \quad (86a)$$

$$\theta_{init}(x) = -5 \text{ [°C]} \quad (86b)$$

For the flow and deformation model HM we consider mixed boundary conditions

$$\tilde{\sigma}(0, t)v_n = 10^5, \quad u(1, t) = 0, \quad (87a)$$

$$p(0, t) = 0, \quad q_l(1, t) \cdot v_n = 0, \quad (87b)$$

$$u_{init}(x) = 0. \quad (87c)$$

The simulation is run over $t \in (0, 30)$ [day], and we show some simulation results in Fig. 1. We report on the number of iterations taken by the iterative solver with respect to the regularization parameter δ in Table 6 and different discretization parameters.

We see that the solver performs robustly and converges within an average of 3–4 iterations for the grid sizes $h \in \{1, 2, 4\} \times 10^{-2}$ [m] and the time steps $\tau \in \{1, 24, 120\}$ [hr]. The smooth evolution of the solution can be observed, without any spurious nonphysical oscillations.

However, the solver does not converge for small time step sizes when $E(\theta)$ is discontinuous (27), but it does converge for sufficiently smooth regularization when $E = E_\delta(\theta)$ with $\delta \geq 0.01$.

With this evidence, we now examine the convergence from Theorem 5.1. We recall now that the only potentially troublesome terms in the assumption (79) are the last two terms which depend on the data (elasticity parameters), and the others can be made arbitrarily small with an appropriately small τ . Hence, we need in practice

$$c_{min} - \frac{M_u^2 L_\lambda^2}{4\lambda_{min}} - \frac{M_u^2 L_\mu^2}{2\mu_{min}} \geq c_0 > 0. \quad (88)$$

We check now if these are realistic. From Example 6.2, we observe that $M_u = O(10^{-3})$. Further, from Table 5, we have $c_{min} = O(10^6)$, $\lambda_{min} = \mu_{min} = O(10^7)$, and $L_\lambda = L_\mu = O(10^8)$. With this data, we see that (88) is satisfied, and the result of Theorem 5.1 holds.

We next study the role of the parameter γ in (20).

Example 6.3. We consider the same scenario as in Example 6.2, and we re-run the simulation by setting $\gamma = 0$. We plot the hydrological fluxes to compare the results with those obtained from Example 6.2.

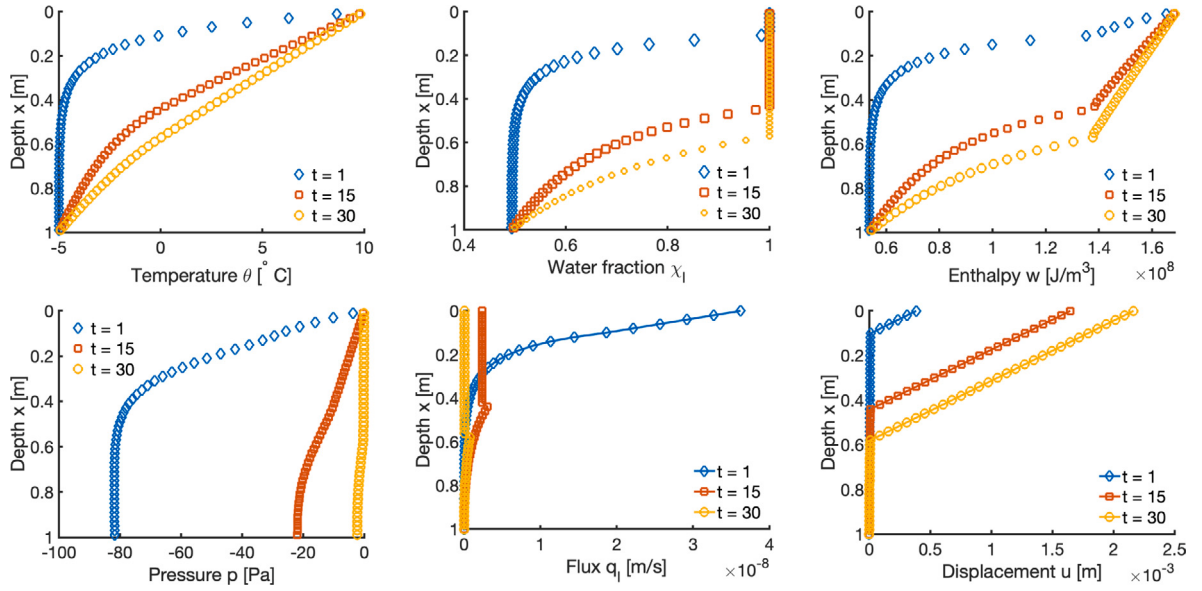


Fig. 1. Results for Example 6.2 showing the solution profiles at different times (units of t in the legend are [day]). Top row: temperature (left), water fraction (middle) and enthalpy (right). Bottom row: pressure (left) and hydrological flux (middle), and displacement (right). Here we have used $h = 0.02$ [m] and $\tau = 24$ [hr].

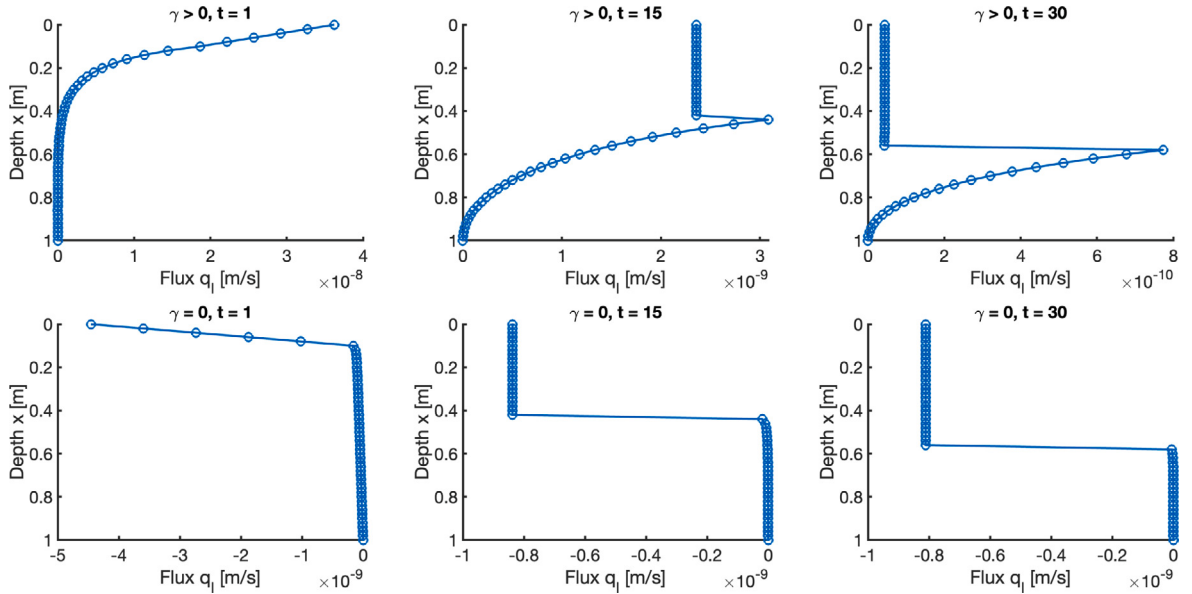


Fig. 2. Results for Example 6.2 showing the effect of γ on the flux q_l . Top row: flux profiles at $t = 1$ (left), $t = 15$ (middle), and $t = 30$ (right) when $\gamma > 0$. Bottom row: flux profiles at $t = 1$ (left), $t = 15$ (middle), and $t = 30$ (right) when $\gamma = 0$. Here we have used $h = 0.02$ [m] and $\tau = 24$ [hr].

The results are shown in Fig. 2. When $\gamma = 0$, we see that the water moves out of Ω through the boundary $x = 0$. However, when $\gamma \neq 0$, the opposite happens. This is due to the fact as the frozen soil thaws, $\partial_t \chi_l > 0$, and thus the change in the volume of ice and liquid water acts as a negative source term in (20b) nonzero only when $\gamma \neq 0$.

Next we study effectiveness of iterative coupling over sequential coupling (i.e., iterative coupling with only one iteration) as well as the effect of regularization $E_\delta(\theta)$.

Example 6.4. Consider the same scenario as in Example 6.2. We re-run our simulation allowing only one iteration in (40a)–(41) at each time step, i.e., we consider sequential coupling only. We also use $E_0(\theta)$ given by (27). We denote this solution $\{\tilde{\theta}, \tilde{p}, \tilde{u}, \tilde{q}_l\}$,

Table 6

Results for [Example 6.2](#) and [Example 6.5](#) showing the number of iterations taken by the iterative solver for regularized E_δ (28) or when using a discontinuous E_0 (27). Note that the solver performs robustly for the homogeneous and heterogeneous cases, but does not converge for δ small, or when using E_0 .

		Example 6.2 (homogeneous case)		Example 6.5 (heterogeneous case)	
h [m]	τ [hr]	Convergence?	Max/min/mean iter.	Convergence	Max/min/mean iter.
E_0					
4×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.6	Yes	5/3/3.6
	1	No	-/-/-	No	-/-/-
2×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.5	Yes	5/3/3.7
	1	No	-/-/-	No	-/-/-
1×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.3	Yes	5/3/3.6
	1	No	-/-/-	No	-/-/-
E_δ with $\delta = 1$					
4×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.6	Yes	5/3/3.7
	1	Yes	7/3/3.0	Yes	5/2/3.0
2×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.6
	24	Yes	5/3/3.4	Yes	5/3/3.6
	1	Yes	6/3/3.1	Yes	5/2/3.0
1×10^{-2}	120	Yes	4/3/3.8	Yes	5/3/3.6
	24	Yes	4/3/3.2	Yes	5/3/3.6
	1	Yes	6/2/3.1	Yes	5/2/3.1
E_δ with $\delta = 0.1$					
4×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.6	Yes	6/3/3.7
	1	Yes	9/3/3.0	Yes	9/2/3.0
2×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	5/3/3.5	Yes	7/3/3.7
	1	Yes	6/3/3.1	Yes	8/2/3.0
1×10^{-2}	120	Yes	4/3/3.8	Yes	5/3/3.7
	24	Yes	4/3/3.2	Yes	7/3/3.6
	1	Yes	7/3/3.1	No	-/-/-
E_δ with $\delta = 0.01$					
4×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.8
	24	Yes	4/3/3.6	Yes	5/3/3.6
	1	No	-/-/-	No	-/-/-
2×10^{-2}	120	Yes	4/4/4	Yes	5/3/3.7
	24	Yes	4/3/3.5	Yes	6/3/3.7
	1	Yes	13/3/3.0	No	-/-/-
1×10^{-2}	120	Yes	4/4/4	Yes	6/3/3.7
	24	Yes	4/3/3.3	Yes	9/3/3.6
	1	Yes	12/3/3.1	No	15/2/3.0

and next we compare with the solution $\{\theta, p, u, q_l\}$ of [Example 6.2](#). We study the difference

$$\Delta f = \|f - \tilde{f}\|_{\infty, \infty}, \Delta f_{rel}(\%) = \frac{\Delta f}{\|f\|_{\infty, \infty}} \times 100, f = \theta, p, q_l, u. \quad (89)$$

The results tabulated in [Table 7](#) show that the difference between the solutions is only significant for small time step $\tau = 1$ [hr]. Further, maximum difference is observed in p and q_l .

6.2.2. Heterogeneous domain

We now consider scenarios to demonstrate the robustness of the solver in heterogeneous domains.

Example 6.5 (One-dimensional heterogeneous soil consolidation). Let $\Omega = (0, 2)$ [m] be occupied by sand in $\Omega^{(1)} = (0, 0.4)$ [m] and clay in $\Omega^{(2)} = (0.4, 2)$ [m] with physical parameters as in [Table 8](#). For the thermal component,

$$\theta(0, t) = -2.5 + 17.5 \sin\left(2\pi\left(t + \frac{7}{12}\right)\right) + \mathcal{F} \text{ on } , \theta(2, t) = -8, \quad (90a)$$

$$\theta_{init}(x) = -8. \quad (90b)$$

Table 7

Results for [Example 6.4](#) comparing the solution obtained using iterative (with E_δ) and sequential (with E_0) coupling. The quantities presented are defined in (89). The difference is most prominent in the pressure and only when τ is small.

h [m]	τ [hr]	Δp [Pa], $\Delta p_{rel}(\%)$	Δq_t [m/s], $\Delta q_{t,rel}(\%)$	Δu [m], $\Delta u_{rel}(\%)$	$\Delta \theta$ [°C], $\Delta \theta_{rel}(\%)$
4×10^{-2}	120	1.33×10^{-1} , 0.21	2.29×10^{-11} , 0.14	7.51×10^{-8} , 0.003	2.57×10^{-2} , 0.26
	24	1.60, 1.95	3.12×10^{-10} , 0.88	2.62×10^{-5} , 1.25	2.22×10^{-2} , 0.23
	1	2.13×10^2 , 194.75	3.92×10^{-8} , 22.23	1.41×10^{-4} , 6.73	6.69×10^{-2} , 0.69
2×10^{-2}	120	1.46×10^{-1} , 0.23	2.74×10^{-11} , 0.17	1.68×10^{-7} , 0.008	2.51×10^{-2} , 0.25
	24	1.90, 2.32	3.37×10^{-10} , 0.93	2.83×10^{-5} , 1.30	2.50×10^{-2} , 0.25
	1	1.13×10^2 , 141.43	1.98×10^{-8} , 11.79	7.05×10^{-5} , 3.26	3.86×10^{-2} , 0.39
1×10^{-2}	120	1.65×10^{-1} , 0.26	4.28×10^{-11} , 0.27	1.89×10^{-5} , 0.90	2.54×10^{-2} , 0.25
	24	6.39×10^{-1} , 0.78	1.35×10^{-10} , 0.37	1.12×10^{-5} , 0.52	2.39×10^{-2} , 0.24
	1	5.44×10^1 , 71.87	9.97×10^{-9} , 5.73	3.52×10^{-5} , 1.62	2.52×10^{-2} , 0.25

Table 8

Physical parameters used in [Example 6.5](#). Units are as in [Table 1](#).

Parameter	Value	Reference
Sand		
c_r	2.128×10^6	[43](Pg. 90) (Quartz)
k_r	8.80	[43](Pg. 90)
ρ_r	2660	[43](Pg. 90)
SFC parameters in (21): $\{b, \theta_s, \chi_{res}\} = \{0.5, 0, 0\}$		
E	$E_u = 20 \times 10^6$	[66](Pg. 407)
	$E_f = 500 \times 10^6$, $a = 4.2$	[1](Pg. 129)
	$\delta = 1$	
ν	0.25	[66](Pg. 407)
$\kappa_{0,u}$	10^{-10}	[39](Pg. 119)
η_{init}	0.4	[39](Pg. 74)
Clay		
c_r	2.385×10^6	[43](Pg. 90)
k_r	2.92	[43](Pg. 90)
ρ_r	2650	[43](Pg. 90)
SFC parameters in (21): $\{b, \theta_s, \chi_{res}\} = \{0.15, 0, 0.2\}$		
E	$E_u = 15 \times 10^6$	[66](Pg. 406)
	$E_f = 500 \times 10^6$, $a = 0.46$	[1](Pg. 129)
	$\delta = 1$	
ν	0.30	[66](Pg. 406)
$\kappa_{0,u}$	10^{-14}	[39](Pg. 119)
η_{init}	0.50	[39](Pg. 74)

Here $\mathcal{F} \sim N(0, 1)$ in (90a) is the Gaussian noise added to replicate the oscillatory nature of the in situ measured surface temperature data; see [Fig. 3](#) for a plot of the surface temperature. For flow and deformation, we consider the boundary and initial conditions

$$\tilde{\sigma}(0, t)v_n = 10^5, \quad u(2, t) = 0 \quad (91a)$$

$$p(0, t) = 100, \quad q_t(2, t) \cdot v_n = 0, \quad (91b)$$

$$u_{init}(x) = 0. \quad (91c)$$

where the Dirichlet pressure condition on $x = 0$ imitates rainfall of approximately 10 [mm].

We run the simulation over $t \in (0, 12)$ [month], where we take 1 [month] = 30 [day], and some plots are included in [Fig. 4](#). We observe no nonphysical oscillations in any of the solution profiles.

We also tabulate the number of iterations taken for different h and τ in [Table 6](#). It appears that the solver struggles to converge when E_0 , or E_δ with small δ , are used. Convergence is achieved within an average of 3–4 iterations for $\delta \geq 0.1$, similarly to the homogeneous case in [Example 6.2](#).

6.3. Two-dimensional isothermal soil consolidation

At this time we are not able to report on simulations of TpHM in $d = 2$, since our current work on TpH is ongoing [33]. In this paper we report only our preliminary results on robustness of HM for heterogeneous soils. We expect this aspect to be most significant when considering partially frozen and thawed soil layers when their material properties dramatically change.

In particular, we investigate the robustness of the P0-RT0-Q1 formulation for an example of HM featuring heterogeneity, with focus on the appropriate time step choice. The goal is to study the possible appearance of Poisson locking or spurious pressure oscillations. We consider a two-dimensional isothermal soil consolidation problem.

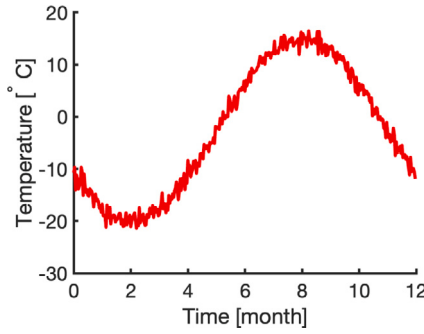


Fig. 3. Illustration for Example 6.5 showing the surface temperature Dirichlet boundary condition.

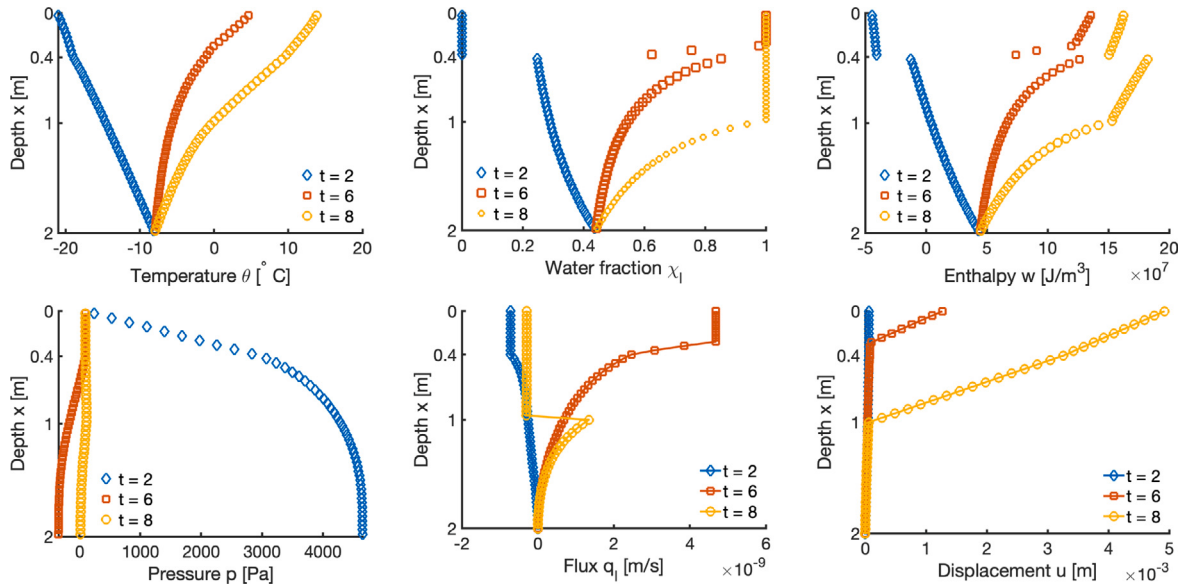


Fig. 4. Results for Example 6.5. Top row: the temperature (left), water fraction (middle), and enthalpy (right) profiles at different times. Bottom row: the pressure (left), flux (middle), and displacement (right) profiles at different times. The units of time in the legend are in [month] = 30[day]. Here we use $h = 4 \times 10^{-2}$ [m] and $\tau = 1$ [day].

Example 6.6 (*Two-dimensional heterogeneous isothermal soil consolidation*). Let $\Omega = (0, 1) \times (0, 1)$ [m]² be occupied by thawed sand in $\Omega^{(1)} = (0, 1) \times (0.5, 1)$ [m]² and frozen clay in $\Omega^{(2)} = (0, 1) \times (0, 0.5)$ [m]²; see Fig. 5. We consider physical parameters as in Table 8, and consider frozen clay in $\Omega^{(2)}$ to be at $\theta = -10$ [°C], and thawed sand to be at $\theta = 5$ [°C].

We consider mixed boundary conditions as follows

$$\tilde{\sigma} v_n = \begin{cases} -10^5; & (x, y) \in (0, 0.5) \times \{1\} \\ 0; & (x, y) \in (0.5, 1) \times \{1\}, (x, y) \in \{0\} \times (0, 1), \end{cases} \quad (92a)$$

$$u_1(1, y) = 0, \quad y \in (0, 1), \quad u(x, 0) = 0, \quad x \in (0, 1), \quad (92b)$$

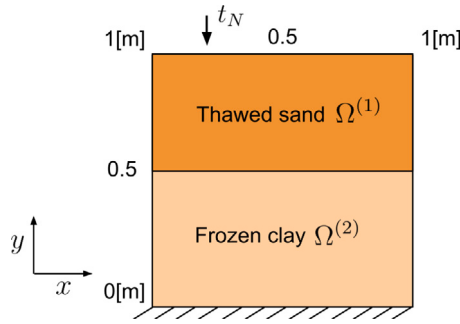
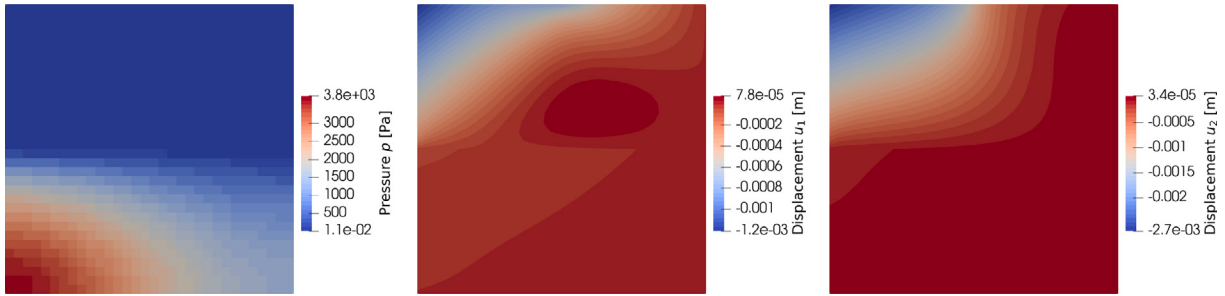
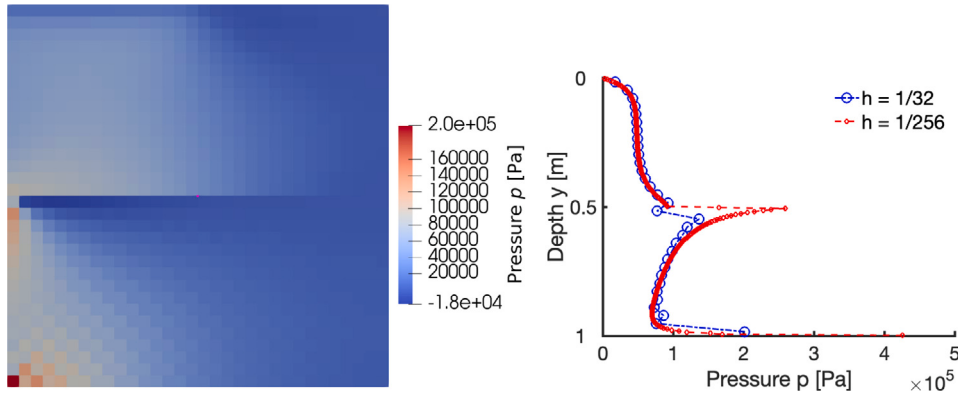
$$p(x, 1) = 0, \quad x \in (0, 1), \quad (92c)$$

$$q_l(0, y) \cdot v_n = 0, \quad q_l(1, y) \cdot v_n = 0, \quad y \in (0, 1), \quad q_l(x, 0) \cdot v_n = 0, \quad x \in (0, 1), \quad (92d)$$

$$u_{init}(x, y) = 0. \quad (92e)$$

We run the simulation using different grid sizes h_x, h_y and time steps τ . Since locking effects are most prominent near $t = 0$, we run the simulation over one time step as done in [28,30].

In Fig. 6 we show solution plots corresponding to $h_x = h_y = 3.125 \times 10^{-2}$ [m] and $\tau = 1$ [hr]; the solution profiles do not feature oscillations. Similarly, no oscillations were observed for $\tau \in [1, 120]$ [hr]. However, when $\tau \lesssim 10^{-2}$ [s], pressure oscillations do appear; this is shown in Fig. 7; this is a well known effect discussed, e.g., in [28,30].

Fig. 5. Illustration for Example 6.6 showing the domain Ω .Fig. 6. Illustration for Example 6.6 showing the plots of the pressure p (left) and displacement components u_1 (middle) and u_2 (right) at the $t = 1$ [hr]. Here $h_x = h_y = 3.125 \times 10^{-2}$ [m] and $\tau = 1$ [hr].Fig. 7. Results for Example 6.6. Here we compare coarse grid ($h_x = h_y = 3.125 \times 10^{-2}$) and fine grid ($h_x = h_y = 3.90625 \times 10^{-3}$) solutions, with $\tau = 10^{-3}$. Left: the pressure profile is shown at the first time step when using coarse grid. Note the oscillations in the profile near the lower left corner and the interface $y = 0.5$. Right: pressure profile plotted along the line $x = 10^{-3}$ at the first time step. It can be seen that applying spatial mesh refinement eliminates the pressure oscillations.

To investigate further the occurrence of Poisson locking, we consider the displacement $u_h^1(0.5, 0.5)$ values for different grid sizes; see Table 9. We see that no abrupt change in the displacement values occurs.

Although the pressure oscillations do arise, as anticipated, at small time steps $O(10^{-2})$ [s], such small time steps are infeasible for practical permafrost simulations; usually, only the time steps of $O(1)$ [hr]– $O(1)$ [day] are practical, with the dynamics of thermal and hydrological processes being fairly slow in usual scenarios [17]. We recall also that as is well known [28,30], for the cases when small time step is required, spatial mesh refinement can be used as a remedy. We illustrate this feature in Fig. 7. For this particular example, when the grid size is refined down to $h_x = h_y = 3.90625 \times 10^{-3}$ [m], the pressure oscillations disappear.

6.4. Summary notes regarding performance of the iterative solver

We now summarize our observations following from the tests of performance of our iterative solver (40)–(41). We see that we can expect our TpHM solver to perform robustly in homogeneous and heterogeneous permafrost scenarios, without any non-physical oscillations and locking.

Table 9

Results for [Example 6.6](#) showing the effect of grid sizes on the displacement $u_h^1(0.5, 0.5)$. It can be observed that the scenarios does not suffer from any significant Poisson locking. In the case of locking, an abrupt change in the values would be observed, as expected from the results reported in [\[28\]](#). Here $\tau = 1$ [hr].

$h_x = h_y$ [m]	$u_h^1 = (u_{h_1}^1, u_{h_2}^1)$ [m]
1/4	$(-1.00337 \times 10^{-5}, -2.31849 \times 10^{-5})$
1/8	$(-1.10757 \times 10^{-5}, -2.39167 \times 10^{-5})$
1/16	$(-1.13748 \times 10^{-5}, -2.41439 \times 10^{-5})$
1/32	$(-1.14727 \times 10^{-5}, -2.42196 \times 10^{-5})$
1/64	$(-1.15060 \times 10^{-5}, -2.42457 \times 10^{-5})$
1/128	$(-1.15178 \times 10^{-5}, -2.42549 \times 10^{-5})$
1/256	$(-1.15221 \times 10^{-5}, -2.42581 \times 10^{-5})$

In particular, we observe that we can successfully use reasonable grid sizes of $h \in [1 \times 10^{-2}, 4 \times 10^{-2}]$ [m] and time steps $\tau \in [1, 120]$ [hr] to simulate realistic scenarios involving TpHM in permafrost. With these, the solver converges within 3–4 iterations on average. However, there are difficulties with convergence when using a discontinuous Young's modulus $E_0(\theta)$, or a regularization E_δ of E_0 with a small δ resulting in a large Lipschitz constant L_{E_δ} . This practical experience agrees with the theory in [Theorem 5.1](#). On the plus side, using a discontinuous Young's modulus E_0 and a sequential algorithm may be effective when a coarse time step and grid size is used. This is exemplified by [Example 6.4](#) where we compare the solution obtained using our iterative solver and a sequential approach. In this particular example, we observe that the difference between the approaches is prominent only for a fine grid size $h = 1 \times 10^{-2}$ [m] and a small time step $\tau = 1$ [hr]. This also informs us that for finer grids using a sequential approach with a discontinuous Young's modulus may not be accurate.

More work remains to study the delicate aspects of the coupled dynamics and the iterative solver. In particular, we see that occasionally the maximum number of iterations at some time steps increases with a decreasing time step, even if the average iterations count decreases. These effects are exacerbated when E_0 is used.

7. Summary and conclusion

In this paper we propose a discretization scheme and a solver in two variants: sequential and iterative for thermo-hydro-mechanical models to simulate ground subsidence due to permafrost thaw. We also recommend regularization for the elastic parameters which seems to aid the solver.

Although there are no a-priori results regarding the order of convergence of the numerical scheme, we provide a convergence study where we show orders of convergence typical for the approximating polynomial orders when applied to the individual model components.

We prove convergence of the iterative solver and demonstrate its robustness in practical heterogeneous soil consolidation scenarios. We show that the use of discontinuous Young's modulus causes difficulties for the iterative solver, but for a regularized model with a sufficiently small Lipschitz constant, convergence is achieved.

We also compare the solutions of the iterative (with E_δ) and sequential (with E_0) variants of the algorithm. We observe that the difference between the two approaches is significant only for small time steps, which is surprising. Furthermore, we investigate the effects of assuming equal liquid water and ice densities in soil consolidation scenarios, and show that the difference is seen predominantly in the hydrological fluxes alone.

Finally, we show that the solver does not suffer from numerical artifacts such as nonphysical pressure oscillations and locking, which are anticipated when using mixed finite elements for hydro-mechanical systems.

More work is needed to investigate the convergence of the solver. In this paper we have provided a convergence proof under certain assumptions, which could be perhaps relaxed or lifted. In particular, we considered equal liquid water and ice densities, and boundedness of the displacement and flux. We also do not see significant effects of the decrease in permeability in [\(67\)](#) directly on the solver either theoretically or in practice, but this aspect deserves further investigation.

For the governing equations of the thermo-hydro-mechanical model, our future work involves including the effects of cryosuction and frost heave. We also plan to consider the variations of the density of liquid water with pressure and temperature.

Lastly, in this paper we only consider rectangular meshes, partly due to the easiness of connecting this work at Darcy work to our pore-scale work and the use of voxel geometries from xray-CT tomography. However, we plan to consider unstructured meshes in the future work to accommodate more complicated geometries if they arise in practical scenarios.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was partially supported by NSF DMS-1912938 “Modeling with Constraints and Phase Transitions in Porous Media”, NSF DMS-1522734 “Phase transitions in porous media across multiple scales”, and NSF DMS-2309682 “Computational mathematics of Arctic processes”, PI: Malgorzata Peszynska. We would like to thank the anonymous reviewers whose comments helped to improve the paper.

Appendix

A.1. Details of the P0-P0 solver for T_p

Here we provide details of our P0-P0 solver for solving the system of equations

$$(w_h^n, m_h) + \tau_n(\nabla \cdot q_{\theta h}^n, m_h) = \tau_n(f, m_h) + (w_h^{n-1}, m_h), \quad (93a)$$

$$(\tilde{k}^{-1} q_{\theta h}^n, \psi_h)_h - (\theta_h^n, \nabla \cdot \psi_h) = 0, \quad (93b)$$

where $\tilde{k} = k(\theta_h^{n-1})$. We first rewrite (93) in matrix-vector form. Let the matrices $\mathcal{M}$, $\mathcal{B}_{q_{\theta}, \theta}$, and $\widetilde{\mathcal{K}}_{q_{\theta}}$ be obtained as follows

$$\begin{aligned} (w_h^n, m_h) &\rightarrow \mathcal{M} W^n \\ (\nabla \cdot q_{\theta h}^n, m_h) &\rightarrow -\mathcal{B}_{q_{\theta}, \theta} Q_{\theta}^n \\ (\tilde{k}^{-1} q_{\theta h}^n, \psi_h)_h &\rightarrow \widetilde{\mathcal{K}}_{q_{\theta}} Q_{\theta}^n, \end{aligned}$$

where Q_{θ}^n collects the degrees of freedom of $q_{\theta h}^n$ in its basis. Then, we can rewrite (93) as

$$\mathcal{M} W^n - \tau_n \mathcal{B}_{q_{\theta}, \theta} Q_{\theta}^n = \tau_n F^n + \mathcal{M} W^{n-1}, \quad W^n = \alpha(\Theta^n), \quad (94a)$$

$$\mathcal{B}_{q_{\theta}, \theta}^T \Theta^n + \widetilde{\mathcal{K}}_{q_{\theta}} Q_{\theta}^n = 0, \quad (94b)$$

or after eliminating Q_{θ}^n as

$$\mathcal{M} W^n + \tau_n \mathcal{A} \Theta^n = \mathcal{M} W^{n-1} + \tau_n F^n, \quad (95)$$

where $\mathcal{A} = \mathcal{B}_{q_{\theta}, \theta} \widetilde{\mathcal{K}}_{q_{\theta}}^{-1} \mathcal{B}_{q_{\theta}, \theta}^T$, and $F^n \in \mathbb{R}^{N_{\omega}}$ collects the entries $(f^n, 1_{\omega_{i,j}})$. To seek a solution W^n, Θ^n to (95), we consider the following algorithm [22,23]. We obtain a sequence $\{\Theta^{n,(i)}\}_i$ as follows

$$\begin{cases} R(\Theta^{n,(i-1)}) = \mathcal{M} \alpha(\Theta^{n,(i-1)}) + \tau_n \mathcal{A} \Theta^{n,(i-1)} - \mathcal{M} W^{n-1} - \tau_n F^n, & (a) \\ (\mathcal{M} \mathcal{J}_{\alpha}^{n,(i-1)} + \tau_n \mathcal{A}) \delta \Theta^{n,(i)} = -R(\Theta^{n,(i-1)}), & (b) \\ \Theta^{n,(i)} = \Theta^{n,(i-1)} + \delta \Theta^{n,(i)}, & (c) \end{cases} \quad (96)$$

where $\mathcal{J}_{\alpha}^{n,(i-1)} = \partial \alpha(\Theta^{n,(i-1)})$ is a diagonal matrix. We denote by $\partial \alpha$ the Clarke’s generalized Jacobian, which is defined as the convex hull $\partial \alpha = \text{co}(\partial_B \alpha)$, with the B-subdifferential

$$\partial_B \alpha(\theta) = \{J_{\alpha} \in \mathbb{R} \mid \exists \{\theta_k\}_k \in D_{\alpha}, \theta_k \rightarrow \theta, (\alpha)'(\theta_k) \rightarrow J_{\alpha}\}, \quad (97)$$

where $D_{\alpha} \subset \mathbb{R}$ is the set where α admits a Fréchet derivative. In our implementation, we use $\partial \alpha(\theta) = \alpha'(\theta), \forall \theta \neq \theta_*$, and we set $\partial \alpha(\theta_*) = c_u$.

In (96), we set $\Theta^{n,(0)} = \Theta^{n-1}$, and we iterate till the residuals R achieve an absolute tolerance of 10^{-12} or a relative tolerance of 10^{-6} (with respect to the first iterate).

A.2. Error norms

In Section 6.1, we compute the norms $\|\cdot\|_{\infty, q}$, $q \in \{1, 2\}$ and $\|\cdot\|_{2,2}$ as follows

$$(f, g) = \sum_{j=1}^{N_{\omega}} f(x_j) g(x_j) h_j, \quad (98a)$$

$$\|f\|_{2,2} = \left(\sum_{n=1}^N \tau_n \|f(\cdot, t_n)\|_2^2 \right)^{\frac{1}{2}}, \quad \|f(\cdot, t_n)\|_2 = \left(\sum_{j=1}^{N_{\omega}} h_j |f(x_j, t_n)|^2 \right)^{\frac{1}{2}}, \quad (98b)$$

Table A.10

A brief review of the various computational schemes for thermo-hydro-mechanical problems.

Reference	Governing equations	Discretization	Nonlinear solver
Nicolsky et al.' 08 [14]	TpHM	Galerkin finite elements	Monolithic, Picard's method
Nishimura et al.' 09 [10]	TpHM	Galerkin finite elements	Monolithic, Newton's method, CODE_BRIGHT framework
Thomas et al.' 09 [11]	TpHM	Galerkin finite elements	Monolithic; COMPASS framework
Zhang, Michalowski' 15 [6]	TpHM	–	ABAQUS 6.12
Zhang et al.'16 [8]	TpHM	–	COMSOL Multiphysics
Liu et al.'19 [7]	TpHM	Quadratic Lagrange for Tp, H; quadratic serendipity for M	COMSOL Multiphysics
Yu et al.'20 [13]	TpHM	Galerkin finite elements	Sequential; MATLAB code
Shastri et al.' 21 [9]	TpHM	Galerkin finite elements	Monolithic; CODE_BRIGHT framework
Liu et al.' 09 [18]	THM	Discontinuous Galerkin	Direct iteration
Brun et al.' 20 [15]	THM	Mixed finite elements	Iterative L-scheme; Python code
Beddrich et al.' 22 [16]	TpH	Finite volumes	Monolithic, semismooth Newton's method; DUNE-PDELab framework

$$\|f\|_{\infty,q} = \max_{1 \leq n \leq N} \left[\left(\sum_{j=1}^{N_\omega} h_j |f(x_j, t_n)|^q \right)^{\frac{1}{q}} \right], \quad q \in \{1, 2\}, \quad (98c)$$

where x_j is the center of the cell ω_j and $h_j = |\omega_j|$ is the size of the cell.

A.3. Literature review table

Here we provide a brief overview of the different numerical discretizations and nonlinear solvers used for TpHM models. See Table A.10.

References

- [1] Andersl OB, Ladanyi B. Frozen ground engineering. 2nd ed.. ASCE, Hoboken, NJ [Reston, Va.]: Wiley; 2004.
- [2] Sandells M, Flocco D. Introduction to the Physics of the Cryosphere. Morgan and Claypool; 2014.
- [3] Burt TP, Williams PJ. Hydraulic conductivity in frozen soils. Earth Surf. Process. 1976;1(4):349–60.
- [4] Tsytovich NA. In: Swinzow George K, Tschebotarioff Gregory P, editors. The mechanics of frozen ground. McGraw-Hill series in modern structures, Scripta Book Co Washington; 1975.
- [5] Michalowski RL. A constitutive model of saturated soils for frost heave simulations. Cold Reg Sci & Technol 1993;22(1):47–63.
- [6] Zhang Y, Michalowski R. Thermal-hydro-mechanical analysis of frost heave and thaw settlement. J Geotech Geoenviron Eng 2015.
- [7] Liu H, Maghoul P, Shalaby A, Bahari A. Thermo-hydro-mechanical modeling of frost heave using the theory of poroelasticity for frost-susceptible soils in double-barrel culvert sites. Transp. Geotech. 2019;20:100251.
- [8] Zhang H, Zhang J, Zhang Z, Chen J, You Y. A consolidation model for estimating the settlement of warm permafrost. Comput Geotech 2016;76:43–50.
- [9] Shastri A, Sánchez M, Gai X, Lee MY, Dewers T. Mechanical behavior of frozen soils: Experimental investigation and numerical modeling. Comput Geotech 2021;138:104361.
- [10] Nishimura S, Gens A, Olivella S, Jardine RJ. THM-coupled finite element analysis of frozen soil: formulation and application. Géotechnique 2009;59(3):159–71.
- [11] Thomas HR, Cleall P, Li Y-C, Harris C, Kern-Luetschg M. Modelling of cryogenic processes in permafrost and seasonally frozen soils. Géotechnique 2009;59(3):173–84.
- [12] Sánchez M, Santamarina C, Teymouri M, Gai X. Coupled numerical modeling of gas hydrate-bearing sediments: From laboratory to field-scale analyses. J Geophys Res Solid Earth 2018;123(12).
- [13] Yu F, Guo P, Lai Y, Stolle D. Frost heave and thaw consolidation modelling. part 2: One-dimensional thermohydrummechanical (THM) framework. Can Geotech J 2020;57(10):1595–610.
- [14] Nicolsky DJ, Romanovsky VE, Tipenko GS, Walker DA. Modeling biogeophysical interactions in nonsorted circles in the Low Arctic. J. Geophys. Res. Biogeosciences 2008;113(G3).
- [15] Brun MK, Ahmed E, Berre I, Nordbotten JM, Radu FA. Monolithic and splitting solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. Comput Math Appl 2020;80(8):1964–84.
- [16] Beddrich J, Gupta S, Wohlmuth B, Chiogna G. The importance of topographic gradients in alpine permafrost modeling. Adv Water Resour 2022;170:104321.
- [17] Gao B, Coon ET. Evaluating simplifications of subsurface process representations for field-scale permafrost hydrology models. Cryosphere 2022;16(10):4141–62.
- [18] Liu R, Wheeler M, Dawson C, Dean R. Modeling of convection-dominated thermoporomechanics problems using incomplete interior penalty Galerkin method. Comput Methods Appl Mech Engrg 2009;198(9):912–9.
- [19] Minkoff S, Stone C, Bryant S, Peszynska M, Wheeler M. Coupled fluid flow and geomechanical deformation modeling. J Pet Sci Eng 2003;38.
- [20] Brun MK, Ahmed E, Nordbotten JM, Radu FA. Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport. J Math Anal Appl 2019;471(1):239–66.
- [21] van Duijn C, Mikelić A, Wheeler MF, Wick T. Thermoporoeasticity via homogenization: Modeling and formal two-scale expansions. Internat J Engrg Sci 2019;138:1–25.

- [22] Bigler L, Peszynska M, Vohra N. Heterogeneous Stefan problem and permafrost models with P0-P0 finite elements and fully implicit monolithic solver. *Electron. Res. Arch.* 2022;30(4):1477–531.
- [23] Vohra N, Peszynska M. Robust conservative scheme and nonlinear solver for phase transitions in heterogeneous permafrost. *J Comput Appl Math* 2024;442:115719.
- [24] Dawson C. Analysis of an upwind-mixed finite element method for nonlinear contaminant transport equations. *SIAM J Numer Anal* 1998;35(5):1709–24.
- [25] Russell TF, Wheeler MF. 2. Finite element and finite difference methods for continuous flows in porous media. 1983, p. 35–106.
- [26] Peszynska M, Jenkins E, Wheeler MF. Boundary conditions for fully implicit two-phase flow model. In: Feng X, Schulze TP, editors. *Recent advances in numerical methods for partial differential equations and applications*. Contemporary mathematics series, vol. 306, American Mathematical Society; 2002, p. 85–106.
- [27] Phillips P, Wheeler M. A coupling of mixed and continuous Galerkin finite element methods for poroelasticity I: The continuous-in-time case. *Comput Geosci* 2007;11:145–58.
- [28] Yi S-Y. Convergence analysis of a new mixed finite element method for Biot's consolidation model. *Numer Methods Partial Differential Equations* 2014;30.
- [29] Rodrigo C, Hu X, Ohm P, Adler J, Gaspar F, Zikatanov L. New stabilized discretizations for poroelasticity and the Stokes' equations. *Comput Methods Appl Mech Engrg* 2018;341:467–84.
- [30] Phillips P, Wheeler M. Overcoming the problem of locking in linear elasticity and poroelasticity: An heuristic approach. *Comput Geosci* 2009;13:5–12.
- [31] Lee J. Robust three-field finite element methods for Biot's consolidation model in poroelasticity. *BIT Numer Math* 2017;58.
- [32] Hansson K, Simunek J, Jiri, Mizoguchi M, Lundin L-C, van Genuchten M. Water flow and heat transport in frozen soil. *Vadose Zone J* 2004;3:527–33.
- [33] Peszynska M, Hilliard Z, Vohra N. Coupled energy and flow models in permafrost from pore- to Darcy scale: model variants and approximations. 2023 [in preparation].
- [34] Peszynska M, Vohra N, Bigler L. Upscaling an extended heterogeneous stefan problem from the pore-scale to the Darcy scale in permafrost. *SIAM Multiscale Model Simul* 2023. accepted.
- [35] Biot MA. General theory of three-dimensional consolidation. *J Appl Phys* 1941;12(2):155–64.
- [36] Showalter R. Diffusion in poro-elastic media. *J Math Anal Appl* 2000;251(1):310–40.
- [37] Shames IH, Cozzarelli FA. Elastic and inelastic stress analysis. CRC Press; 1997.
- [38] Detournay E, Cheng Alexander H-D. Fundamentals of poroelasticity. In: Fairhurst C, editor. *Analysis and design methods*, vol. 2. Pergamon, Oxford: Pergamon Press; 1993, p. 113–71.
- [39] Bear J, Cheng A. Modeling groundwater flow and contaminant transport, vol. 23, 2010.
- [40] Engineering Toolbox. <https://www.engineeringtoolbox.com>. Retrieved in 2023.
- [41] Verruijt A. Elastic storage of aquifers. In: Wiest RJM De, editor. *Flow through porous media*. New York: Academic Press; 1969, p. 331–76.
- [42] McKenzie JM, Voss CI, Siegel DI. Groundwater flow with energy transport and water–ice phase change: Numerical simulations, benchmarks, and application to freezing in peat bogs. *Adv Water Resour* 2007;30(4):966–83.
- [43] Williams PJ, Smith MW. The frozen earth: Fundamentals of geocryology. In: *Studies in polar research*, Cambridge University Press; 1989.
- [44] Grenier C, Anbergen H, Bense V, Chanzy Q, Coon E, Collier N, et al. Groundwater flow and heat transport for systems undergoing freeze-thaw: Intercomparison of numerical simulators for 2d test cases. *Adv Water Resour* 2018;114:196–218.
- [45] Kurylyk BL, Watanabe K. The mathematical representation of freezing and thawing processes in variably-saturated, non-deformable soils. *Adv Water Resour* 2013;60:160–77.
- [46] Phillips P, Wheeler M. A coupling of mixed and continuous Galerkin finite element methods for poroelasticity II: The discrete-in-time case. *Comput Geosci* 2007;11:145–58.
- [47] Brenner SC, Scott LR. The mathematical theory of finite element methods. *Texts in applied mathematics*, vol. 15, Springer; 2008.
- [48] Dawson C. Godunov-mixed methods for advection-diffusion equations in multidimensions. *SIAM J Numer Anal* 1993;30(5):1315–32.
- [49] van Duijn CJ, Mikelić A, Wick T. Mathematical theory and simulations of thermoporoelasticity. *Comput Methods Appl Mech Engrg* 2020;366:113048.
- [50] Kim J, Tchelepi H, Juanes R. Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics. *SPE J* 2009;16.
- [51] Mikelić A, Wang B, Wheeler MF. Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Comput Geosci* 2014;18(3):325–41.
- [52] Jha B, Juanes R. A locally conservative finite element framework for the simulation of coupled flow and reservoir geomechanics. *Acta Geotech* 2007;2:139–53.
- [53] Wheeler MF, Gai X. Iteratively coupled mixed and Galerkin finite element methods for poro-elasticity. *Numer Methods Partial Differential Equations* 2007;23(4):785–97.
- [54] Kim J. Unconditionally stable sequential schemes for all-way coupled thermoporoelasticity: Undrained-adiabatic and extended fixed-stress splits. *Comput Methods Appl Mech Engrg* 2018;341:93–112.
- [55] Pop I, Radu F, Knabner P. Mixed finite elements for the richards' equation : linearization procedure. *J Comput Appl Math* 2004;168(1–2):365–73.
- [56] Both JW, Borregales M, Nordbotten JM, Kumar K, Radu FA. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Appl Math Lett* 2017;68:101–8.
- [57] Borregales Reverón M, Radu F, Kumar K, Nordbotten J. Robust iterative schemes for non-linear poromechanics. *Comput Geosci* 2018;22.
- [58] Murad M, Loula A. On stability and convergence of finite element approximations of Biot's consolidation theory. *Internat J Numer Methods Engrg* 1994;37:645–67.
- [59] Rodrigo C, Gaspar F, Hu X, Zikatanov L. Stability and monotonicity for some discretizations of the Biot's consolidation model. *Comput Methods Appl Mech Engrg* 2016;298:183–204.
- [60] Adler JH, He Y, Hu X, MacLachlan S, Ohm P. Monolithic multigrid for a reduced-quadrature discretization of poroelasticity. *SIAM J Sci Comput* 2023;45(3):S54–81.
- [61] Arndt D, Bangerth W, Bergbauer M, Feder M, Fehling M, Heinz J, et al. The deal.II library, version 9.5.. *J Numer Math* 2023 [in press].
- [62] Davis TA. Algorithm 832: Umfpack v4.3—an unsymmetric-pattern multifrontal method. *ACM Trans Math Software* 2004;30(2):196–9.
- [63] Weiser A, Wheeler MF. On convergence of block-centered finite differences for elliptic problems. *SIAM J Numer Anal* 1988;25(2):351–75.
- [64] Han W, Atkinson KE. Theoretical numerical analysis. NY: Springer New York; 2009.
- [65] Rogers JC, Berger AE, Ciment M. The alternating phase truncation method for numerical solution of a Stefan problem. *SIAM J Numer Anal* 1979;16(4):563–87.
- [66] Briaud J-L. Geotechnical engineering: unsaturated and saturated soils. Wiley; 2013.