EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media

Citation for published version (APA):
Radu, F. A., Kumar, K., Nordbotten, J. M., & Pop, I. S. (2015). *A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media*. (CASA-report; Vol. 1539). Technische Universiteit Eindhoven.

Document status and date:
Published: 01/12/2015

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 04. Oct. 2023

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computer Science

CASA-Report 15-39
December 2015

A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media

by

F.A. Radu, K. Kumar, J.M. Nordbotten, I.S. Pop

Centre for Analysis, Scientific computing and Applications
Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven, The Netherlands
ISSN: 0926-4507

A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media

Florin A. Radu[1], Kundan Kumar[1], Jan M. Nordbotten[1], Iuliu S. Pop[1,2]

[1] Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020 Bergen, Norway
[2] Department of Mathematics and Computer Science, Eindhoven University of Technology,
P. O. Box 513, 5600 MB Eindhoven, The Netherlands

e-mails: {*florin.radu, jan.nordbotten, kundan.kumar*}*@math.uib.no, i.pop@tue.nl*

Abstract. In this work we present a mass conservative numerical scheme for two-phase flow in porous media. The model for flow consists on two fully coupled, non-linear equations: a degenerate parabolic equation and an elliptic equation. The proposed numerical scheme is based on backward Euler for the temporal discretization and mixed finite element method (MFEM) for the discretization in space. Continuous, semi-discrete (continuous in space) and fully discrete variational formulations are set up and the existence and uniqueness of solutions is discussed. Error estimates are presented to prove the convergence of the scheme. The non-linear systems within each time step are solved by a robust linearization method. This iterative method does not involve any regularization step. The convergence of the linearization scheme is rigorously proved under the assumption of a Lipschitz continuous saturation. The case of a Hölder continuous saturation is also discussed, a rigorous convergence proof being given for Richards' equation. Numerical results are presented to sustain the theoretical findings.

Keywords: linearization, two-phase flow, mixed finite element method, convergence analysis, a priori error estimates, porous media, Richards' equation, degenerate parabolic problems, coupled problems.

1 Introduction

Two-phase porous media flow models are widely encountered in real-life applications of utmost societal relevance, including water and soil pollution, oil recovery, geological carbon dioxide sequestration, or nuclear waste management [31, 21]. Such complex problems admit only in very simplified situations analytical solutions, therefore numerical methods for solving multiphase flow in porous media are playing a determining role in understanding and solving the problems. Nevertheless, the design and analysis of robust, accurate and efficient numerical schemes is a very challenging task.

Here we discuss a numerical scheme for a two-phase porous media flow model. The fluids are assumed immiscible and incompressible and the solid matrix is non-deformable. The adopted formulation uses the global pressure and a complementary pressure, obtained by using the Kirchhoff transformation, as primary unknowns (see [11, 3, 12]). This leads to a system of two coupled non-linear partial differential equations, a degenerate elliptic - parabolic one and an elliptic one.

Numerical methods for two-phase flow have been the object of intensive research in the last decades. The major challenge in developing efficient schemes is related to the degenerate nature of the problem. Due to this, the solution typically lacks regularity, which makes lower order finite elements or finite volumes a natural choice for the spatial discretization. In this respect, we refer to [20, 32] for Galerkin finite elements, to [19, 34, 30] for finite volumes, to [16, 13, 14] for methods combining Galerkin finite elements combined with the mixed finite element method (MFEM), and to [17, 43] for the discontinuous Galerkin method. In all cases, the convergence of the numerical schemes is proved rigorously either by compactness arguments, or by obtaining *a priori* error estimates. *A posteriori* error estimates are obtained e.g. in [9]. Furthermore, similar issues appear for the Richards equation, which is a simplified model for saturated/unsaturated flow in the case when the pressure of one phase is supposed to be constant. In this context we mention Galerkin finite elements [32, 36], MFEM based works [4, 38, 40, 48, 50], multipointflux approximation (MPFA) [28] or finite volume - MFEM combined methods [18].

In this paper we propose a mass conservative scheme based on MFEM (lowest order Raviart-Thomas elements [7]) and backward Euler for numerical simulation of the two-phase flow in porous media. Continuous, semi-discrete (continuous in space) and fully discrete mixed variational formulations are defined. Existence and uniqueness of solutions is discussed, the equivalence with a conformal formulation being involved in the proof. We show the convergence of the numerical scheme and provide explicit order of convergence estimates. The analysis is inspired by similar results in [4, 14, 38, 40].

Typical problems involving flow in porous media, like e.g. water and soil pollution or nuclear waste management are spread over decades or even centuries, so that the use of relatively large time steps is a necessity. Due to this, implicit methods are a necessity (our choice here being the first-order backward Euler method, due to the low regularity of the considered problem). Since the original model is non-linear, at each time step one needs to solve non-linear algebraic systems. In this work we propose a robust linearization scheme for the systems appearing at each time step, as a valuable alternative to modified Picard method [10] or Newton's method [5, 35, 39, 33, 27] or iterative IMPES [22, 23]. Although the applicability of Newton's method for parabolic equations is well recognized, its convergence is not straightforward for degenerate equations, where the Jacobian might become singular. A possible way to overcome this is to regularize the problem. However, even in this case convergence is guaranteed only under a severe stability condition for the discretization parameters, see [39]. This has motivated the alternative, robust linearization scheme proposed in this work. The new scheme, called $L-$scheme from now on, does not involve the calculations of any derivatives and does not need a regularization step. The L-scheme combines the idea of a classical Picard method and the scheme presented in [37] for MFEM or [49, 45] for Galerkin finite elements. The L-scheme was proposed for two-phase flow in combination with the MPFA method in [42], the proof of convergence there being only sketched and not made completely rigorous. We show here that the L-scheme for MFEM based discretizations converges linearly if the time step satisfies a mild condition. This robustness is the main advantage of the scheme when compared to the quadratic, but locally convergent Newton method.

All the papers quoted above are considering Lipschitz continuous nonlinearities, like the dependency of the saturation on the complementary pressure. This is due to the fact that the L-scheme as proposed there involves the constants that need to be larger than the Lipschitz constants of the nonlinear functions in the models. If the nonlinearities are only Hölder continuous but not Lipschitz, the derivatives become unbounded. Then the convergence proof for the L-scheme, as presented in [37] for the MFEM discretization of the Richards equation, or in [49, 45] for the Galerkin finite elements also for the Richards equation, or in [42] for MPFA/two-phase flow, is not valid anymore. Commonly, one is regularizing first the problem by approximating the non-Lipschitz nonlinearities by Lipschitz ones, and then iterative methods like Newton, Picard, or the above mentioned L-scheme is applied. In this paper we show that the L-scheme can be applied for the Hölder continuous case as well. We prove this rigorously for the simplified case of the Richards equation, but the extension to two-phase flow does not present any special difficulties (up to extremely technical calculations).

Finally, we mention that the L-scheme can be interpreted as a non-linear preconditioner, because the linear systems to be solved within each iteration are much better conditioned than the corresponding systems in the case of modified Picard or Newton's method. We refer to [26] for illustrative examples concerning the Richards equation, which is a particular case of the more general model considered in the present work.

To summarize, the main new contributions of this paper are

- We present and analyze a MFEM based numerical scheme for two-phase flow in porous media. Order of convergence estimates are obtained.
- We show the existence and uniqueness of the considered variational formulation. This is based on the equivalence between the conformal and the mixed formulations, which is proved here for the continuous and the time discrete models.
- We present and analyze rigorously a robust, first-order convergent linearization method for MFEM based schemes for two-phase flow in porous media.
- The paper is the first to apply the L-scheme to the model involving non-Lipschitz nonlinearities.

The paper is structured as follows. In Section 2, we present the model equations for two-phase flow in porous media and we define the discretization scheme. In Section 3 we analyze the convergence of the discretization scheme based on *a priori* error estimates. We also prove existence and uniqueness for the problem involved, and give stability estimates. A new MFEM linearization scheme is presented and analyzed in Section 4. Section 5 provides numerical examples confirming the theoretical results. The paper is ending with concluding remarks in Section 6.

2 Mathematical model and discretization

In this section we introduce the notations used in this work, the mathematical model and its MFEM/Euler implicit discretization (Problems P, P^n and P_h^n). A linearization scheme (Problem $P_h^{n,i}$) is proposed to solve the non-linear systems appearing at each time step.

The model is defined in the d-dimensional bounded domain $\Omega \subset \mathbb{R}^d$ having a Lipschitz continuous boundary Γ. Further $T > 0$ is the final computational time. We use common notations from functional analysis, e.g. $L^\infty(\Omega)$ is the space of essential bounded functions on Ω, $L^2(\Omega)$ the space of square integrable functions on Ω, or $H^1(\Omega)$ is the subspace of $L^2(\Omega)$ containing functions which have also first order derivatives in $L^2(\Omega)$. We denote by $H_0^1(\Omega)$ the space of $H^1(\Omega)$ functions with a vanishing trace on Γ and by $H^{-1}(\Omega)$ its dual. $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(\Omega)$, or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Further, $\| \cdot \|$, $\| \cdot \|_1$ and $\| \cdot \|_\infty$ stand for the norms in $L^2(\Omega)$, $H^1(\Omega)$, respectively $L^\infty(\Omega)$. The functions in $H(\mathrm{div};\Omega)$ are vector valued having a L^2 divergence. The norm in $H(\mathrm{div};\Omega)$ is denoted by $\| \cdot \|_{div}$. $L^2(0,T;X)$ denotes the Bochner space of X-valued functions defined on $(0,T)$, where X is a Banach space. Similarly, $C(0,T;X)$ are X-valued functions continuous (w. r. t. X norm) on $[0,T]$. By C we mean a generic positive constant, not depending on the unknowns or the discretization parameters and we denote by L_f the Lipschitz constant of a (Lipschitz continuous) function $f(\cdot)$.

Further, we will denote by $N \geq 1$ an integer giving the time step $\tau = T/N$. For a given $n \in \{1, 2, \dots, N\}$, the nth time point is $t_n = n\tau$. We will also use the following notation for the mean over a time interval. Given the function $g \in L^2(0,T;X)$ (X being a Banach space like $L^2(\Omega)$, or $H^1(\Omega)$)), its time averaged over the interval $(t_{n-1}, t_n]$ is defined as

$$\bar{g}^n := \frac{1}{\tau} \int_{t_{n-1}}^{t_n} g(t)dt.$$

Clearly, this is an element in X as well.

The two-phase porous media flow model considered here assumes that the fluids are immiscible and incompressible, and that the solid matrix is non-deformable. By denoting with $\alpha = w, n$ the wetting and non-wetting phases, $s_\alpha, p_\alpha, \mathbf{q}_\alpha, \rho_\alpha$ the saturation, pressure, flux and density of phase α, respectively, the two-phase model under consideration reads (see e.g. [6, 11, 21, 31])

$$\begin{aligned} \frac{\partial(\phi\rho_\alpha s_\alpha)}{\partial t} + \nabla\cdot(\rho_\alpha \mathbf{q}_\alpha) &= 0, \qquad\qquad \alpha = w, n, && (1)\\ \mathbf{q}_\alpha &= -\frac{k_{r,\alpha}}{\mu_\alpha} k(\nabla p_\alpha - \rho_\alpha \mathbf{g}), \quad \alpha = w, n, && (2)\\ s_w + s_n &= 1, && (3)\\ p_n - p_w &= p^{cap}(s_w), && (4) \end{aligned}$$

where $\mathbf{g}$ denotes the constant gravitational vector. Equation (1) is a mass balance, (2) is the Darcy law, (3) is an algebraic evidence expressing that all pores in the medium are filled by a mixture of the two fluid phases and (4) is the capillary pressure relationship, with $p^{cap}(\cdot)$ supposed to be known. The porosity ϕ, permeability k, the viscosities μ_α are given constants and the relative permeabilities $k_{r,\alpha}(\cdot)$ are given functions of s_w. We consider here a scalar permeability, but the results can be easily extended to the case when the permeability is positive-definite tensor.

In this paper we adopt a global/complementary pressure formulation [3, 11, 12]. The global pressure (denoted by p) was introduced in [11] and the complementary pressure in [3]. They are defined by

$$p(\mathbf{x}, s_w) := p_n(\mathbf{x}) - \int_0^{s_w} f_w(\mathbf{x},\xi) \frac{\partial p^{cap}}{\partial \xi}(\mathbf{x},\xi) d\,\xi, \qquad (5)$$

$$\Theta(\mathbf{x}, s_w) := -\int_0^{s_w} f_w(\mathbf{x},\xi)\lambda_n(\mathbf{x},\xi) \frac{\partial p^{cap}}{\partial \xi}(\mathbf{x},\xi) d\,\xi, \qquad (6)$$

where we denoted by $\lambda_\alpha := \dfrac{k_{r,\alpha}}{\mu_\alpha}$, $\alpha = w, n$ the phase mobilities and by $f_w := \dfrac{\lambda_w}{\lambda_w + \lambda_n}$ the fractional flow function. We note the use of Kirchhoff transformation above. In the new unknowns, the resulting system consists of two coupled non-linear partial differential equations, a degenerate parabolic one and an elliptic one. For more details on the modelling we refer to [12], where the existence and uniqueness of a weak solution is proved for a Galerkin-MFEM formulation. In the new unknowns the system (1)-(4) becomes

$$\begin{aligned} \partial_t s(\Theta) + \nabla\cdot\mathbf{q} &= 0, && (7)\\ \mathbf{q} &= -\nabla\Theta + f_w(s)\mathbf{u} + \mathbf{f}_1(s), && (8)\\ \nabla\cdot\mathbf{u} &= f_2(s), && (9)\\ a(s)\mathbf{u} &= -\nabla p - \mathbf{f}_3(s). && (10) \end{aligned}$$

with $s := s_w$, $a(s) := \dfrac{1}{k\lambda(s)}$, $\mathbf{q}$ the (wetting) flux, and $\mathbf{u}$ the total flux. The equations hold true in $\Omega \times (0,T]$. The coefficient functions $s(\cdot), a(\cdot), f_w(\cdot), \mathbf{f}_1(\cdot), f_2(\cdot), \mathbf{f}_3(\cdot)$ are given and satisfy the assumptions listed below. The system is completed by initial conditions specified below, and by boundary conditions. For simplicity, we restrict our attention to homogeneous Dirichlet boundary conditions, but other kinds of conditions can be considered.

Also note that the results can be extended straightforwardly to the case when a source term f_s is present on the right hand side of (7). For the ease of presentation and in view of the analogy with the model considered in [40], the source terms are left out here. This simplifies the presentation of the convergence proof in Section 3.

Problem P: Continuous mixed variational formulation.
Find $\Theta, p \in L^2(0,T;L^2(\Omega))$, $\mathbf{q} \in L^2(0,T;(L^2(\Omega))^d)$, $\mathbf{u} \in L^2(0,T;H(\mathrm{div};\Omega))$ such that there holds $s(\Theta) \in L^\infty(\Omega \times (0,T))$, $\int_0^t \mathbf{q}(y)\,dy \in C(0,T;H(\mathrm{div};\Omega))$, and

$$\begin{aligned}
\langle s(\Theta(t)) - s(\Theta^0), w\rangle + \langle \nabla\cdot \int_0^t \mathbf{q}(y)\,dy, w\rangle &= 0, && (11)\\
\langle \int_0^t \mathbf{q}(y)\,dy, \mathbf{v}\rangle - \langle \int_0^t \Theta(y)\,dy, \nabla\cdot\mathbf{v}\rangle & && \\
-\langle \int_0^t f_w(s(\Theta(y)))\mathbf{u}(y)\,dy, \mathbf{v}\rangle &= \langle \int_0^t \mathbf{f}_1(s(\Theta(y)))\,dy, \mathbf{v}\rangle, && (12)\\
\langle \nabla\cdot\mathbf{u}(t), w\rangle &= \langle f_2(s(\Theta(t))), w\rangle, && (13)\\
\langle a(s(\Theta(t)))\mathbf{u}(t), \mathbf{v}\rangle - \langle p(t), \nabla\cdot\mathbf{v}\rangle + \langle \mathbf{f}_3(s(\Theta(t))), \mathbf{v}\rangle &= 0 && (14)
\end{aligned}$$

for all $t \in (0,T]$, $w \in L^2(\Omega)$ and $\mathbf{v} \in H(\mathrm{div};\Omega)$, with $\Theta(0) = \Theta_I \in L^2(\Omega)$.

The function Θ_I is given. For example, if the function $s(\cdot)$ is one to one, a natural initial condition is $\Theta_I = s^{-1}(s_I)$, where s_I is a given initial saturation. By (11), since $\int_0^t \mathbf{q}(y)dy$ is continuous in time, it follows that $s(\Theta) \in C(0,T;L^2(\Omega))$. Then, since $s(\cdot)$ and $f_2(\cdot)$ are assumed continuous (see (A1) and (A3) below), from (13) one obtains that $\mathbf{u} \in C(0.T;H(\mathrm{div};\Omega))$. Similarly, p is continuous in time as well, so (12)-(14) hold for all $t \in (0,T]$.

We now proceed with the time discretization for Problem P, which is achieved by the Euler implicit scheme. For a given $n \in \{1,2,\ldots,N\}$, we define the time discrete mixed variational problem at time t_n (the time step is denoted by τ):

Problem P^n: Semi-discrete variational formulation. Let Θ^{n-1} be given. Find $\Theta^n, p^n \in L^2(\Omega)$ and $\mathbf{u}^n, \mathbf{q}^n \in H(\mathrm{div};\Omega)$ such that

$$\begin{aligned}
\langle s^n - s^{n-1}, w\rangle + \tau\langle \nabla\cdot\mathbf{q}^n, w\rangle &= 0, && (15)\\
\langle \mathbf{q}^n, \mathbf{v}\rangle - \langle \Theta^n, \nabla\cdot\mathbf{v}\rangle - \langle f_w(s^n)\mathbf{u}^n, \mathbf{v}\rangle &= \langle \mathbf{f}_1(s^n), \mathbf{v}\rangle, && (16)\\
\langle \nabla\cdot\mathbf{u}^n, w\rangle &= \langle f_2(s^n), w\rangle, && (17)\\
\langle a(s^n)\mathbf{u}^n, \mathbf{v}\rangle - \langle p^n, \nabla\cdot\mathbf{v}\rangle + \langle \mathbf{f}_3(s^n), \mathbf{v}\rangle &= 0 && (18)
\end{aligned}$$

for all $w \in L^2(\Omega)$, and $\mathbf{v} \in H(\mathrm{div};\Omega)$. Initially we take $\Theta^0 = \Theta_I \in L^2(\Omega)$. Throughout this paper s^k stands for $s(\Theta^k)$, $k \in \mathbb{N}$, making the presentation easier.

We can now proceed with the spatial discretization. For this let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size (see [15]). Here we assume $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$, hence Ω is polygonal. Thus we neglect the errors caused by an approximation of a non-polygonal domain and avoid an excess of technicalities (a complete analysis in this sense can be found in [32]).

The discrete subspaces $W_h \times V_h \subset L^2(\Omega) \times H(\mathrm{div};\Omega)$ are defined as

$$\begin{aligned} W_h &:= \{p \in L^2(\Omega) |\, p \text{ is constant on each element } T \in \mathcal{T}_h\}, \\ V_h &:= \{\mathbf{q} \in H(\mathrm{div};\Omega) | \mathbf{q}_{|T}(\mathbf{x}) = \mathbf{a_T} + b_T\mathbf{x}, \mathbf{a_T} \in \mathbb{R}^2, b_T \in \mathbb{R} \text{ for all } T \in \mathcal{T}_h\}. \end{aligned} \tag{19}$$

So W_h denotes the space of piecewise constant functions, while V_h is the RT_0 space (see [7]).

We will use the following projectors (see [7] and [44], p. 237):

$$P_h : L^2(\Omega) \to W_h, \qquad \langle P_h w - w, w_h\rangle = 0, \tag{20}$$

and

$$\Pi_h : H(\mathrm{div};\Omega) \to V_h, \qquad \langle \nabla \cdot (\Pi_h \mathbf{v} - \mathbf{v}), w_h\rangle = 0, \tag{21}$$

for all $w \in L^2(\Omega)$,$\mathbf{v} \in H(\mathrm{div};\Omega)$ and $w_h \in W_h$. For these operators we have

$$\|w - P_h w\| \leq Ch\|w\|_1, \qquad \text{respectively} \qquad \|\mathbf{v} - \Pi_h \mathbf{v}\| \leq Ch\|\mathbf{v}\|_1 \tag{22}$$

for any $w \in H^1(\Omega)$ and $\mathbf{v} \in (H^1(\Omega))^d$.

The fully discrete (non-linear) scheme can now be given. To simplify notations we use in the following the notation $s_h^n := s(\Theta_h^n)$, $n \in \mathbb{N}$.

Problem P_h^n: Fully discrete (non-linear) variational formulation. Let $n \in \mathbb{N}, n \geq 1$, and assume Θ_h^{n-1} is known. Find $\Theta_h^n, p_h^n \in W_h$ and $\mathbf{q}_h^n, \mathbf{u}_h^n \in V_h$ such that there holds

$$\begin{aligned} \langle s_h^n - s_h^{n-1}, w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^n, w_h\rangle &= 0, & (23)\\ \langle \mathbf{q}_h^n, \mathbf{v}_h\rangle - \langle \Theta_h^n, \nabla \cdot \mathbf{v}_h\rangle - \langle f_w(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle &= \langle \mathbf{f}_1(s_h^n), \mathbf{v}_h\rangle, & (24)\\ \langle \nabla \cdot \mathbf{u}_h^n, w_h\rangle &= \langle f_2(s_h^n), w_h\rangle, & (25)\\ \langle a(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle - \langle p_h^n, \nabla \cdot \mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^n), \mathbf{v}_h\rangle &= 0 & (26) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$.

The fully discrete scheme (23) – (26) is non-linear, and iterative schemes are required for solving it. Moreover, in (A1) it is only required that $s(\cdot)$ is monotone and Lipschitz, so it may have a vanishing derivative, which makes (7) degenerate. This is, indeed, the situation encountered in two-phase porous media flow models. In such cases, usual schemes such as the Newton method may not converge without performing a regularization step. This may affect the mass balance. For the Richards equation, this is proved in [39]. Following the ideas in [37, 45, 49], we propose a robust, first order convergent linearization scheme for solving (23) – (26). The scheme is not requiring any regularization. Similar ideas can be applied in connection with any other spatial discretization method, see e.g. [42] for multipoint flux approximation. The analysis of the scheme is presented in Section 4.

Let $n \in N, n \geq 1$ be fixed. Assuming that $s(\cdot)$ is Lipschitz continuous, the following iterative scheme can be used to solve the non-linear problem (23) – (26):

Problem $P_h^{n,i}$: Linearization scheme (L-scheme). Let $L \geq L_s$, $i \in \mathbb{N}, i \geq 1$ and let $\Theta_h^{n,i-1} \in W_h$ be given. Find $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ such that

$$\begin{aligned} \langle L(\Theta_h^{n,i} - \Theta_h^{n,i-1}) + s_h^{n,i-1}, w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,i}, w_h\rangle &= \langle s_h^{n-1}, w_h\rangle, & (27)\\ \langle \mathbf{q}_h^{n,i}, \mathbf{v}_h\rangle - \langle \Theta_h^{n,i}, \nabla \cdot \mathbf{v}_h\rangle - \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i}, \mathbf{v}_h\rangle &= \langle \mathbf{f}_1(s_h^{n,i-1}), \mathbf{v}_h\rangle, & (28)\\ \langle \nabla \cdot \mathbf{u}_h^{n,i}, w_h\rangle &= \langle f_2(s_h^{n,i-1}), w_h\rangle, & (29)\\ \langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i}, \mathbf{v}_h\rangle - \langle p_h^{n,i}, \nabla \cdot \mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^{n,i-1}), \mathbf{v}_h\rangle &= 0 & (30) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. We use the notation $s_h^{n,i} := s(\Theta_h^{n,i})$, $n \in \mathbb{N}$ and, as previously, $s_h^n := s(\Theta_h^n)$, $n \in \mathbb{N}$. For starting the iterations, a natural choice is $\Theta_h^{n,0} := \Theta_h^{n-1}$ and, correspondingly, $s_h^{n,0} := s_h^{n-1}$. Note that since we prove below that the iterative scheme is a contraction, this choice is not compulsory for the convergence.

Throughout this paper we make the following assumptions:

(A1) The function $s(\cdot) : \mathbb{R} \to [0, 1]$ is monotone increasing and Lipschitz continuous.

(A2) $a(\cdot)$ is Lipschitz continuous and there exists $a_\star, a^\star > 0$ such that for all $y \in \mathbb{R}$ one has

$$0 < a_\star \leq a(y) \leq a^\star < \infty. \tag{31}$$

(A3) $\mathbf{f}_1(\cdot), f_2(\cdot), \mathbf{f}_3(\cdot)$ and $f_w(\cdot)$ are Lipschitz continuous. Additionally, $f_w(\cdot)$ is uniformly bounded.

(A4) There exits a constant $M_\mathbf{u} < \infty$ such that $\|\mathbf{u}\|_\infty \leq M_\mathbf{u}$, $\|\mathbf{u}_h^n\|_\infty \leq M_\mathbf{u}$, and $\|\mathbf{u}_h^{n,i}\|_\infty \leq M_\mathbf{u}$ for all $n \in \mathbb{N}$, the last two being uniformly in h and i. Here $\mathbf{u}$, $\mathbf{u}_h^n$ and $\mathbf{u}_h^{n,i}$ are the solution components in Problems P, P_h^n and $\mathrm{P}_h^{n,i}$ respectively.

(A5) The function Θ_I is in $L^2(\Omega)$.

Remark 2.1. *The assumptions are satisfied in most situations of practical interest. Concerning (A4), for* $\mathbf{u}$ *this is practically the outcome of the assumptions (A1) and (A3), which guarantee that for every* $t \in [0, T]$ *one has* $f_2(s(\Theta(t))) \in L^\infty(\Omega)$, *and that the* L^∞ *norm is bounded uniformly w.r.t. time. Now, without being rigorous, we observe that by* (13) *one obtains* $\mathbf{u}(t) = -\nabla w(t)$, *where* w *satisfies* $-\Delta w(t) = f_2(s(\Theta(t)))$. *Classical regularity theory (see e.g. [29], Thm. 15.1 in Chapter 3) guarantees that* ∇w *is continuous on the compact* $\bar{\Omega}$, *and that the* L^∞ *norm can be bounded uniformly in time. For the approximation* $\mathbf{u}_h^n$, *one can reason in the same manner, and observe that* $\mathbf{u}_h^n$ *becomes the projection* $\Pi_h(-\nabla w(t_n))$. *Since* $\|\nabla w(t_n)\|_\infty$ *is bounded uniformly in time, the construction of the projector* Π_h *(see e. g. [44], Chapter 7.2) guarantees that* $\mathbf{u}_h^n$ *satisfies the same bounds as* ∇w. *Finally, case of* $\mathbf{u}_h^{n,i}$ *is similar. We also refer to [40] for a similar situation but in the case of a one phase flow model, where conditions ensuring the validity of (A4) are provided.*

Remark 2.2. *One may relax the Lipschitz continuity for the parameter functions to a more general context, where e. g.* f_w *satisfies a growth condition* $|f_w(s_1) - f_w(s_2)|^2 \leq C|\langle f_w(s_1) - f_w(s_2), s_1 - s_2\rangle|$. *We do not exploit this possibility here, but refer to [12, 40, 38] for the details on the procedure.*

The following two technical lemmas will be used in Sections 3 and 4. Their proofs can be found e.g. in [40] and [47], respectively.

Lemma 2.1. *Given a* $w \in L^2(\Omega)$, *there exists a* $\mathbf{v} \in H(\mathrm{div}; \Omega)$ *such that*

$$\nabla \cdot \mathbf{v} = w \text{ and } \|\mathbf{v}\| \leq C_\Omega \|w\|,$$

with $C_\Omega > 0$ *not depending on* w.

Lemma 2.2. *Given a* $w_h \in W_h$, *there exits a* $\mathbf{v}_h \in V_h$ *satisfying*

$$\nabla \cdot \mathbf{v}_h = w_h \text{ and } \|\mathbf{v}_h\| \leq C_{\Omega,d} \|w_h\|,$$

with $C_{\Omega,d} > 0$ *not depending on* w_h *or mesh size.*

Also, the following elementary results will be used

Proposition 2.1. *Let $\mathbf{a}_k \in \mathbb{R}^d$ $(k \in \{1, \dots, N\}, d \geq 1)$ be a set of N vectors. It holds*

$$\sum_{n=1}^{N} \langle \mathbf{a}_n, \sum_{k=1}^{n} \mathbf{a}_k \rangle \quad = \quad \frac{1}{2} \left\| \sum_{n=1}^{N} \mathbf{a}_n \right\|^2 + \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{a}_n\|^2 \, . \tag{32}$$

Proposition 2.2. *(Hölder's inequality) Let $a, b \in \mathbb{R}$, $\epsilon > 0$ and $p, q > 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

$$|ab| \leq \epsilon \frac{|a|^p}{p} + \epsilon^{-\frac{q}{p}} \frac{|b|^q}{q} \, . \tag{33}$$

3 Analysis of the discretization: existence and uniqueness and a priori error estimates

In this section we analyze the problems P, P_n and P_h^n introduced in Section 2. The existence and uniqueness of a solution will be discussed in Subsection 3.1. For the continuous and semi-discrete cases this will be done by showing an equivalence with conformal variational formulations. The convergence of the numerical scheme will be shown by deriving *a priori* error estimates. The main convergence result is given in Theorem 3.1. The convergence is established by assuming that the non-linear systems (23) - (26) are solved exactly. We refer to Section 4 for the analysis of the linearization scheme (27)-(30), which was proposed in the previous section to solve these non-linear algebraic systems numerically.

3.1 Existence and uniqueness for the variational problems

In this subsection we discuss the existence and uniqueness of the continuous, semi-discrete, and fully discrete variational formulations for the considered model (7)-(10). We establish an equivalence between the continuous mixed formulation and a conformal formulation, which will deliver the existence and uniqueness for the continuous case. The semi-discrete case can be treated analogously. For the fully discrete case we prove below the uniqueness. Existence can be proved by fixed point arguments, using e.g. Lemma 1.4, p. 164 in [46]. We omit the details here as the existence is also a direct consequence of the results in Section 4, where the convergence of a linear iterative scheme is proved. The limit of this iteration is exactly a solution for the fully discrete system.

A conformal variational formulation for the model (7)-(10) reads:

Problem PC: Continuous conformal variational formulation. Let $\Theta_I \in L^2(\Omega)$ be given. Find Θ_C, $p_C \in L^2(0, T; H_0^1(\Omega))$, such that $\partial_t s(\Theta_C) \in L^2(0, T; H^{-1}(\Omega))$, $\Theta_C(0) = \Theta_I$, and for all $v \in L^2(0, T; H_0^1(\Omega))$ and $w \in H_0^1(\Omega)$ one has

$$\begin{aligned} \int_0^T \langle \partial_t s(\Theta_C), v \rangle dt + \int_0^T \langle \nabla \Theta_C + \frac{f_w(s\Theta_C)}{a(s(\Theta_C))} \nabla p_C, \nabla v \rangle dt & \\ + \int_0^T \langle \frac{f_w(s\Theta_C)}{a(s(\Theta_C))} \mathbf{f}_3(s(\Theta_C)) - \mathbf{f}_1(s(\Theta_C)), \nabla v \rangle dt &= 0, \end{aligned} \tag{34}$$

$$\langle \frac{1}{a(s(\Theta_C))} \big(\nabla p_C + \mathbf{f}_3(s(\Theta_C)) \big), \nabla w \rangle = -\langle f_2(s), w \rangle. \tag{35}$$

The existence and uniqueness of a solution for Problem PC has been studied intensively in the past. Closest to the framework considered here is [12] (see also [14]). There the existence and uniqueness is proved, however, for the case that the inverse of $s(\cdot)$ is Lipschitz (the so-called slow diffusion case). Here we assume $s(\cdot)$ Lipschitz but not necessarily strictly increasing, which is a fast diffusion case.

Other relevant references for the existence and uniqueness are [1, 2, 20, 24]. Also, we refer to [8] for the existence of a solution in heterogeneous media, where the phase pressure differences may become discontinuous at the interface separating two homogeneous blocks.

Having this in mind, one can use the existence and uniqueness results for the conformal formulation to obtain the existence of a solution for Problem P (as by-product also establish its regularity). The equivalence is established in Proposition 3.1, whose proof follows the ideas in [38], Proposition 2.2.

Proposition 3.1. *Let $\Theta_C, p_C \in L^2(0,T;H_0^1(\Omega))$ be a solution to Problem PC, define $s_C = s(\Theta_C)$ and assume that (A1)-(A5) hold true. Then, a solution to Problem P is given by $\Theta = \Theta_C$, $p = p_C$, $\mathbf{q} = -\nabla\Theta_C - \frac{f_w(s_C)}{a(s_C)}(\nabla p_C + \mathbf{f}_3(s_C)) + \mathbf{f}_1(s_C)$ and $\mathbf{u} = -\frac{1}{a(s_C)}(\nabla p_C + \mathbf{f}_3(s_C))$. Conversely, if $(\Theta, \mathbf{q}) \in L^2(0,T;L^2(\Omega)) \times L^2(0,T;(L^2(\Omega))^d)$, $(p, \mathbf{u}) \in L^2(0,T;L^2(\Omega)) \times L^2(0,T;H(\mathrm{div};\Omega))$ are solving Problem P, then $\Theta, p \in L^2(0,T;H_0^1(\Omega))$ and (Θ, p) is a solution of Problem PC.*

Proof. " $\Rightarrow$ " Clearly, Θ and p defined above have the regularity required in Problem PC. Furthermore, $\mathbf{u}$ and $\mathbf{q}$ are elements of $L^2(0,T;L^2(\Omega)^d)$. Recalling that $s(\cdot)$ is Lipschitz continuous, one immediately obtains that $s_C \in L^2(0,T;H_0^1(\Omega))$. Since $\partial_t s_C \in L^2(0,T;H^{-1}(\Omega))$ this shows that $s_C \in C(0,T;L^2(\Omega))$. With $t \in (0,T]$ and $\phi \in H_0^1(\Omega)$ arbitrary chosen, taking now $v = \chi_{(0,t]}\phi$ in (34) and using the definition of $\mathbf{q}$ gives

$$\langle s(\Theta_C) - s(\Theta^0), \phi\rangle - \langle \int_0^t \mathbf{q}(y)\, dy, \nabla\phi\rangle = 0, \tag{36}$$

for all $\phi \in H_0^1(\Omega)$. In other words, $\nabla \cdot \int_0^t \mathbf{q}(y)\, dy = s(\Theta^0) - s(\Theta_C(t))$ in distributional sense. The regularity of s_C mentioned above implies that, actually, $\int_0^t \mathbf{q}(y)dy$ lies in $H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$ as well, and that $\int_0^t \mathbf{q} \in C(0,T;H(\mathrm{div};\Omega))$. Moreover, by density arguments (11) holds for any $w \in L^2(\Omega)$. Also, from (36) one gets $s_C \in C(0,T;L^2(\Omega))$.

In a similar way, using (35) and the definition of $\mathbf{u}$ one obtains

$$\big\langle -\mathbf{u}, \nabla w\big\rangle = \langle f_2(s_C), w\rangle,$$

for all $w \in H_0^1(\Omega)$, so $\nabla \cdot \mathbf{u} = f_2(s_C)$ for a. e. $t \in [0,T]$. In view of the regularity of s_C and of the assumptions on f_2, this means that $\mathbf{u} \in C(0,T;H(\mathrm{div};\Omega))$ and that (13) holds for every $t \in [0,T]$.

To obtain (14) one uses (35), the definition of $\mathbf{u}$ and that p_C has a vanishing trace on Γ. This gives

$$\langle a(s_C)\mathbf{u}, \mathbf{v}\rangle = -\langle \nabla p_C + \mathbf{f}_3(s_C), \mathbf{v}\rangle = \langle p_C, \nabla \cdot \mathbf{v}\rangle - \langle \mathbf{f}_3(s_C), \mathbf{v}\rangle,$$

for a. e. $t \in [0,T]$ and for all $\mathbf{v} \in H(\mathrm{div};\Omega)$. Recalling the continuity in time for s_C and $\mathbf{u}$, it follows that (14) is valid for every $t \in [0,T]$. Finally, one can use similar ideas to show that (12) holds true as well.

Q. E. D." $\Rightarrow$ "

" $\Leftarrow$ " Let now $(\Theta, \mathbf{q}, p, \mathbf{u})$ be the solution of Problem P. We have to show that $\Theta, p \in L^2((0,T)\times\Omega)$ is the solution of PC, i.e. to show that the functions are actually in $L^2(0,T;H_0^1(\Omega))$ and that they satisfy (34)-(35). Further, since in (A1) $s(\cdot)$ is assumed Lipschitz, it follows that $s(\Theta) \in L^2(0,T;H^1(\Omega))$ as well. Clearly, (11) gives $\partial_t s(\Theta) = -\nabla \cdot \mathbf{q}$ for a.e. $t \in [0,T]$. Since $\mathbf{q} \in L^2(0,T;(L^2(\Omega))^d)$ one gets $\partial_t s(\Theta) \in L^2(0,T;H^{-1}(\Omega))$.

Taking $v \in (C_0^\infty(\Omega))^d \subset H(\mathrm{div}, \Omega)$ arbitrary as test function in (14) we get

$$\langle a(s(\Theta(y)))\mathbf{u}, \mathbf{v}\rangle + \langle \nabla p, \mathbf{v}\rangle + \langle \mathbf{f}_3(s(\Theta(y))), \mathbf{v}\rangle = 0, \tag{37}$$

which implies

$$\nabla p = -a(s(\Theta(y)))\mathbf{u} - \mathbf{f}_3(s(\Theta(y))) \tag{38}$$

first in a distributional sense, and then by using the regularity of $\mathbf{u}$ and $s(\cdot)$ in L^2 sense. It follows that $p \in L^2(0, T; H^1(\Omega))$. It remains to verify that p has a vanishing trace on the boundary of Ω. Taking now $\mathbf{v} \in H(\mathrm{div}; \Omega)$ and using the regularity of p and (14) one obtains

$$\langle \nabla p, \mathbf{v} \rangle = -\langle a(s(\Theta(y)))\mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{f}_3(s(\Theta(y))), \mathbf{v} \rangle = -\langle p, \nabla \cdot \mathbf{v} \rangle, \tag{39}$$

for all $v \in H(\mathrm{div}; \Omega)$. By using now the Green theorem for $\mathbf{v} \in H(\mathrm{div}; \Omega)$, see [7], pg. 91

$$\int_\Gamma p\mathbf{v} \cdot \mathbf{n} ds = (\nabla p, \mathbf{v}) + (p, \nabla \cdot \mathbf{v}) = 0.$$

It follows immediately that $p \in L^2(0, T; H_0^1(\Omega))$. From (38) and (13) one gets that p satisfies (35). In a similar manner one can show that $\Theta \in L^2(0, T; H_0^1(\Omega))$ and it satisfies (34).

Q. E. D." $\Leftarrow$ "

We can now state an analogous result for the semi-discrete case.

Problem PC^n: Semi-discrete conformal variational formulation. Let $n \in \mathbb{N}, n \geq 1$ and Θ^{n-1} be given. Find $\Theta_C^n, p_C^n \in H_0^1(\Omega)$, such that for all $v, w \in H_0^1(\Omega)$ one has

$$\begin{aligned} \langle s(\Theta_C^n) - s(\Theta_C^{n-1}), v \rangle + \langle \nabla \Theta_C^n + \frac{f_w(s\Theta_C^n)}{a(s(\Theta_C^n))} \nabla p_C^n, \nabla v \rangle & \\ + \langle \frac{f_w(s\Theta_C^n)}{a(s(\Theta_C^n))} \mathbf{f}_3(s(\Theta_C^n)) - \mathbf{f}_1(s(\Theta_C^n)), \nabla v \rangle &= 0, \qquad (40) \\ \langle \frac{1}{a(s(\Theta_C^n))} (\nabla p_C^n + \mathbf{f}_3(s(\Theta_C^n))), \nabla w \rangle &= -\langle f_2(s(\Theta_C^n)), w \rangle. \qquad (41) \end{aligned}$$

Proposition 3.2. *Let $n \in \mathbb{N}, n \geq 1$, Θ^{n-1} be given and assume that (A1)-(A5) hold true. If $\Theta^n, p^n \in H_0^1(\Omega)$ is a solution to Problem PC^n then, a solution to Problem P^n is given by $\Theta^n = \Theta_C^n$, $p^n = p_C^n$, $\mathbf{q}^n = -\nabla \Theta_C^n - \frac{f_w(s_C^n)}{a(s_C^n)} (\nabla p_C^n + \mathbf{f}_3(s_C^n)) + \mathbf{f}_1(s_C^n)$ and $\mathbf{u}^n = -\frac{1}{a(s_C^n)} (\nabla p_C^n + \mathbf{f}_3(s_C^n))$. Conversely, if $(\Theta^n, \mathbf{q}^n) \in L^2(\Omega) \times H(\mathrm{div}; \Omega)$, $(p, \mathbf{u}) \in L^2(\Omega) \times H(\mathrm{div}; \Omega)$ are solving Problem P^n, then $\Theta^n, p^n \in H_0^1(\Omega)$ and (Θ^n, p^n) is a solution of Problem PC^n.*

The proof of Proposition 3.2 is similar with the one of Proposition 3.1 (also see [38], Prop. 2.3) and it will be skipped here. Then the existence and uniqueness of a solution for the semi-discrete variational formulation (15)–(18) is a direct consequence of the equivalence result.

The following proposition establish the uniqueness for the fully-discrete variational problem (23)-(26). As explained in the introduction of Sec. 3.1, the existence of a solution will follow from the convergence of the linear iteration scheme.

Proposition 3.3. *Let $n \in \mathbb{N}, n \geq 1$ be fixed. If (A1)-(A5) hold true and the time step τ is sufficiently small, then the problem* (23)-(26) *has at most one solution.*

Proof. Let us assume that there exists two solutions of (23)-(26), $(\Theta_{h,i}^n, \mathbf{q}_{h,i}^n, p_{h,i}^n, \mathbf{u}_{h,i}^n) \in W_h \times V_h \times W_h \times V_h$ with $i = 1, 2$. Further, we let $s_{h,i}^n := s(\Theta_{h,i}^n)$ stand for the two saturations. These solutions are then satisfying

$$\begin{aligned} \langle s_{h,i}^n - s_h^{n-1}, w_h \rangle + \tau \langle \nabla \cdot \mathbf{q}_{h,i}^n, w_h \rangle &= 0, \qquad (42) \\ \langle \mathbf{q}_{h,i}^n, \mathbf{v}_h \rangle - \langle \Theta_{h,i}^n, \nabla \cdot \mathbf{v}_h \rangle - \langle f_w(s_{h,i}^n) \mathbf{u}_{h,i}^n, \mathbf{v}_h \rangle &= \langle \mathbf{f}_1(s_{h,i}^n), \mathbf{v}_h \rangle, \qquad (43) \\ \langle \nabla \cdot \mathbf{u}_{h,i}^n, w_h \rangle &= \langle f_2(s_{h,i}^n), w_h \rangle, \qquad (44) \\ \langle a(s_{h,i}^n) \mathbf{u}_{h,i}^n, \mathbf{v}_h \rangle - \langle p_{h,i}^n, \nabla \cdot \mathbf{v}_h \rangle + \langle \mathbf{f}_3(s_{h,i}^n), \mathbf{v}_h \rangle &= 0 \qquad (45) \end{aligned}$$

for $i = 1, 2$ and for all $w_h \in W_h$, $\mathbf{v}_h \in V_h$. By subtracting (44) and (45) for $i = 2$ from the same equations for $i = 1$ we get for all $w_h \in W_h$, $\mathbf{v}_h \in V_h$

$$\begin{aligned} \langle \nabla \cdot (\mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n), w_h \rangle &= \langle f_2(s_{h,2}^n) - f_2(s_{h,1}^n), w_h \rangle, & (46) \\ \langle a(s_{h,2}^n)\mathbf{u}_{h,2}^n - a(s_{h,1}^n)\mathbf{u}_{h,1}^n, \mathbf{v}_h \rangle - \langle p_{h,2}^n - p_{h,1}^n, \nabla \cdot \mathbf{v}_h \rangle &= \langle \mathbf{f}_3(s_{h,1}^n) - \mathbf{f}_3(s_{h,2}^n), \mathbf{v}_h \rangle. & (47) \end{aligned}$$

We test now (46) with $w_h = p_{h,2}^n - p_{h,1}^n \in W_h$ and (47) with $\mathbf{v}_h = \mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n \in V_h$, and add the resulting equations to obtain

$$\begin{aligned} \langle a(s_{h,2}^n)\mathbf{u}_{h,2}^n - a(s_{h,1}^n)\mathbf{u}_{h,1}^n, \mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n \rangle = \langle f_2(s_{h,2}^n) - f_2(s_{h,1}^n), p_{h,2}^n - p_{h,1}^n \rangle \\ + \langle \mathbf{f}_3(s_{h,1}^n) - \mathbf{f}_3(s_{h,2}^n), \mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n \rangle. \end{aligned} \tag{48}$$

After some algebraic manipulations, using (A2)-(A4) and Cauchy-Schwarz and Young inequalities one gets from (48) above

$$\frac{a_\star}{4} \|\mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n\|^2 \le \left(\frac{L_{f_2}^2}{2\delta} + \frac{L_{\mathbf{f}_3}^2}{2a_\star} + \frac{L_a^2 M_u^2}{a_\star}\right) \|s_{h,2}^n - s_{h,1}^n\|^2 + \frac{\delta}{2} \|p_{h,2}^n - p_{h,1}^n\|^2, \tag{49}$$

for all $\delta > 0$. Using Lemma 2.2 there exists $\mathbf{v}_h \in V_h$ such that $\nabla \cdot \mathbf{v}_h = p_{h,2}^n - p_{h,1}^n$ and $\|\mathbf{v}_h\| \le C_{\Omega,d} \|p_{h,2}^n - p_{h,1}^n\|$. Testing with this $\mathbf{v}_h$ in (47) one gets

$$\|p_{h,2}^n - p_{h,1}^n\|^2 = \langle a(s_{h,2}^n)\mathbf{u}_{h,2}^n - a(s_{h,1}^n)\mathbf{u}_{h,1}^n, \mathbf{v}_h \rangle + \langle \mathbf{f}_3(s_{h,1}^n) - \mathbf{f}_3(s_{h,2}^n), \mathbf{v}_h \rangle. \tag{50}$$

Using the properties of $\mathbf{v}_h$, (A2)-(A4) and Cauchy-Schwarz's inequality, (50) implies

$$\|p_{h,2}^n - p_{h,1}^n\| \le C(\|\mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n\| + \|s_{h,1}^n - s_{h,2}^n\|), \tag{51}$$

with C not depending on the solutions or discretization parameters. From (49) and (51) follows, by properly chosen δ

$$\|\mathbf{u}_{h,2}^n - \mathbf{u}_{h,1}^n\| \le C \|s_{h,1}^n - s_{h,2}^n\|, \tag{52}$$

with C not depending on the solutions or discretization parameters. We proceed by subtracting (42) and (43) for $i = 2$ from the same equations for $i = 1$ to obtain

$$\begin{aligned} \langle s_{h,2}^n - s_{h,1}^n, w_h \rangle + \tau \langle \nabla \cdot (\mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n), w_h \rangle &= 0, & (53) \\ \langle \mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n, \mathbf{v}_h \rangle - \langle \Theta_{h,2}^n - \Theta_{h,1}^n, \nabla \cdot \mathbf{v}_h \rangle &= \langle f_w(s_{h,2}^n)\mathbf{u}_{h,2}^n - f_w(s_{h,1}^n)\mathbf{u}_{h,1}^n, \mathbf{v}_h \rangle \\ &\quad + \langle \mathbf{f}_1(s_{h,2}^n) - \mathbf{f}_1(s_{h,1}^n), \mathbf{v}_h \rangle, & (54) \end{aligned}$$

for all $w_h \in W_h, \mathbf{v}_h \in V_h$. Testing (53) with $w_h = \Theta_{h,2}^n - \Theta_{h,1}^n \in W_h$ and (54) with $\mathbf{v}_h = \tau(\mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n) \in V_h$ and then adding the results gives

$$\begin{aligned} &\langle s_{h,2}^n - s_{h,1}^n, \Theta_{h,2}^n - \Theta_{h,1}^n \rangle + \tau \|\mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n\|^2 = \\ &\quad \tau \langle f_w(s_{h,2}^n)\mathbf{u}_{h,2}^n - f_w(s_{h,1}^n)\mathbf{u}_{h,1}^n, \mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n \rangle + \tau \langle \mathbf{f}_1(s_{h,2}^n) - \mathbf{f}_1(s_{h,1}^n), \mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n \rangle. \end{aligned} \tag{55}$$

By some algebraic manipulations, using (A1)-(A4), the Cauchy-Schwarz and Young inequalities and the result (52), we obtain from (55) above

$$\langle s_{h,2}^n - s_{h,1}^n, \Theta_{h,2}^n - \Theta_{h,1}^n \rangle + \frac{\tau}{4} \|\mathbf{q}_{h,2}^n - \mathbf{q}_{h,1}^n\|^2 \le C\tau \|s_{h,1}^n - s_{h,2}^n\|^2, \tag{56}$$

with C not depending on the solutions or discretization parameters. Due to (A1), there holds $\langle s_{h,2}^n - s_{h,1}^n, \Theta_{h,2}^n - \Theta_{h,1}^n \rangle \geq \frac{1}{L_s} \| s_{h,1}^n - s_{h,2}^n \|^2$, which together with (56) immediately implies (for sufficiently small τ) that $\mathbf{q}_{h,2}^n = \mathbf{q}_{h,1}^n$ and $s_{h,1}^n = s_{h,2}^n$. Using these and (52), we obtain that $\mathbf{u}_{h,2}^n = \mathbf{u}_{h,1}^n$. From (51) results also $p_{h,2}^n = p_{h,1}^n$. Finally, equation (54) will furnish $\Theta_{h,2}^n = \Theta_{h,1}^n$, which completes the proof of uniqueness.

Q. E. D.

Remark 3.1. *The uniqueness of a solution for the L-scheme introduced in* (27)–(30) *can be proved similarly. Since this is a linear algebraic system, uniqueness also gives the existence of a solution.*

3.2 Stability estimates

As in [40] and, actually, as in Proposition 3.2 one can obtain some stability estimates for the Problems P^n. Moreover, the same holds for Problem P_h^n, but these estimates are not needed for proving the convergence of the scheme and are therefore skipped here. Following [40], Lemma 3.2, pg. 293 there holds:

Proposition 3.4. *Assume that (A1)-(A5) hold true. Let* $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$*,* $n \in \mathbb{N}, n \geq 1$ *be the solution of Problem* P^n*. Then there holds*

$$\tau \sum_{n=1}^{N} \|\Theta^n\|_1^2 + \tau \sum_{n=1}^{N} \|\mathbf{q}^n\|_{div}^2 + \tau \sum_{n=1}^{N} \|p^n\|_1^2 + \tau \sum_{n=1}^{N} \|\mathbf{u}^n\|_{div}^2 \quad \leq \quad C, \tag{57}$$

with C not depending on discretization parameters.

3.3 A priori error estimates

Having established the existence and uniqueness for Problems P, P^n and P_h^n, we can focus now on the convergence of the scheme (23)-(26). This will be done by deriving *a priori* error estimates. We assume that the fully discrete non-linear problem (23)-(26) is solved exactly. The proofs of this section follow the lines in [40] and [14]. The following two propositions will quantify the error between the continuous and the semi-discrete formulations, and between the semi-discrete and discrete ones, respectively. Finally the two propositions will be put together to obtain the main convergence result given in Theorem 3.1.

Recalling the definition $\bar{g}^n := \frac{1}{\tau} \int_{t_{n-1}}^{t_n} g(t) dt \in X$, for any $g \in L^2(0,T;X)$ and X a Banach space, we have

Proposition 3.5. *Let* $(\Theta, \mathbf{q}, p, \mathbf{u})$ *be the solution of Problem P and* $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$ *be the solution of* $P^n, n \in \mathbb{N}, n \geq 1$*. Assuming (A1)-(A5) and that the time step is small enough there holds:*

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta^n), \Theta(t) - \Theta^n \rangle \, dt + \| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \mathbf{q} - \mathbf{q}^n \, dt \|^2$$
$$+ \sum_{n=1}^{N} \| \int_{t_{n-1}}^{t_n} (\mathbf{q} - \mathbf{q}^n) \, dt \|^2 \quad \leq \quad C\tau, \tag{58}$$

$$\tau \|\bar{\mathbf{u}}^n - \mathbf{u}^n\|^2 + \tau \|\bar{p}^n - p^n\|^2 \quad \leq \quad C \int_{t_{n-1}}^{t_n} \|s(\Theta(t)) - s(\Theta^n)\|^2 dt, \tag{59}$$

$$\| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\Theta(t) - \Theta^n) \, dt \|^2 \quad \leq \quad C\tau, \tag{60}$$

with the constants C not depending on the discretization parameters.

Proof. We start with proving (59). By integrating (13), (14) from t_{n-1} to t_n one obtains

$$\begin{aligned}\langle \nabla \cdot \overline{\mathbf{u}}^n, w\rangle &= \langle \overline{f_2(s)}^n, w\rangle, & (61)\\ \langle \overline{a(s)\mathbf{u}}^n, \mathbf{v}\rangle - \langle \overline{p}^n, \nabla\cdot\mathbf{v}\rangle + \langle \overline{\mathbf{f}_3(s)}^n, \mathbf{v}\rangle &= 0, & (62)\end{aligned}$$

for all $w \in L^2(\Omega)$ and $\mathbf{v} \in H(\mathrm{div};\Omega)$. By subtracting now (17) and (18) from (61) and (62), respectively, we get

$$\begin{aligned}\langle \nabla \cdot (\overline{\mathbf{u}}^n - \mathbf{u}^n), w\rangle &= \langle \overline{f_2(s)}^n - f_2(s^n), w\rangle, & (63)\\ \langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \mathbf{v}\rangle - \langle \overline{p}^n - p^n, \nabla\cdot\mathbf{v}\rangle &= -\langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \mathbf{v}\rangle, & (64)\end{aligned}$$

for all $w \in L^2(\Omega)$ and $\mathbf{v} \in H(\mathrm{div};\Omega)$. Taking $w = \overline{p}^n - p^n \in L^2(\Omega)$ in (63) and $\mathbf{v} = \overline{\mathbf{u}}^n - \mathbf{u}^n \in H(\mathrm{div};\Omega)$ in (64), and summing the results we obtain

$$\langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle = \langle \overline{f_2(s)}^n - f_2(s^n), \overline{p}^n - p^n\rangle - \langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle. \quad (65)$$

By Young's inequality, this further implies

$$\begin{aligned}\langle \frac{1}{\tau}\int_{t^{n-1}}^{t_n} a(s)\mathbf{u} - a(s^n)\mathbf{u}^n\,dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle &\le \frac{1}{2\delta}\|\overline{f_2(s)}^n - f_2(s^n)\|^2 + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2\\ &+\frac{1}{a_\star}\|\overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n)\|^2 + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2.\end{aligned}$$

The above is further equivalent to

$$\begin{aligned}&\langle \frac{1}{\tau}\int_{t^{n-1}}^{t_n}(a(s) - a(s^n))\mathbf{u}\,dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle + \langle \frac{a(s^n)}{\tau}\int_{t^{n-1}}^{t_n}\mathbf{u} - \mathbf{u}^n\,dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle\\ &\quad\le \frac{1}{2\delta}\|\overline{f_2(s)}^n - f_2(s^n)\|^2 + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2 + \frac{1}{a_\star}\|\overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n)\|^2 + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2,\end{aligned}$$

which by using (A2)-(A3) leads to

$$\frac{3a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 \le (\frac{L_{f_2}^2}{2\tau\delta} + \frac{L_{\mathbf{f}_3}^2}{2\tau a_\star})\int_{t_{n-1}}^{t_n}\|s - s^n\|^2\,dt + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2 - T_1, \quad (66)$$

for all $\delta > 0$, where $T_1 = \langle \frac{1}{\tau}\int_{t^{n-1}}^{t_n}(a(s) - a(s^n))\mathbf{u}\,dt, \overline{\mathbf{u}}^n - \mathbf{u}^n\rangle$. Now, to estimate T_1 one uses (A2), (A4) and the Young inequality to obtain

$$\begin{aligned}|T_1| &\le \|\frac{1}{\tau}\int_{t^{n-1}}^{t_n}(a(s) - a(s^n))\mathbf{u}\,dt\|\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|\\ &\le \frac{1}{\tau^2 a_\star}\int_\Omega\left(\int_{t^{n-1}}^{t_n}(a(s) - a(s^n))\mathbf{u}\,dt\right)^2 dx + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2\\ &\le \frac{M_{\mathbf{u}}^2}{\tau a_\star}\int_\Omega\int_{t^{n-1}}^{t_n}((a(s) - a(s^n))^2\,dt\,dx + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2\\ &\le \frac{M_{\mathbf{u}}^2 L_a^2}{\tau a_\star}\int_{t^{n-1}}^{t_n}\|s - s^n\|^2\,dt + \frac{a_\star}{4}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2. & (67)\end{aligned}$$

From (66) and (67) it follows immediately

$$\frac{a_\star}{2}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 \leq \frac{C}{\tau}\int_{t_{n-1}}^{t_n} \|s(\Theta(t)) - s(\Theta^n)\|^2 dt + \frac{\delta}{2}\|\overline{p}^n - p^n\|^2, \tag{68}$$

with C not depending on the discretization parameters.

To estimate the last term above, one uses Lemma 2.1, ensuring the existence of a $\mathbf{v} \in H(\mathrm{div};\Omega)$ such that $\nabla \cdot \mathbf{v} = \overline{p}^n - p^n$ and $\|\mathbf{v}\| \leq C_\Omega\|\overline{p}^n - p^n\|$. Using this as test function in (64) gives

$$\begin{aligned} \|\overline{p}^n - p^n\|^2 &= \langle \overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n, \mathbf{v}\rangle + \langle \overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n), \mathbf{v}\rangle \\ &\leq C_\Omega^2\|\overline{a(s)\mathbf{u}}^n - a(s^n)\mathbf{u}^n\|^2 + C_\Omega^2\|\overline{\mathbf{f}_3(s)}^n - \mathbf{f}_3(s^n)\|^2 + \frac{1}{2}\|\overline{p}^n - p^n\|^2. \end{aligned} \tag{69}$$

Proceeding as for (68), for (69) we get:

$$\|\overline{p}^n - p^n\|^2 \leq \frac{C}{\tau}\int_{t_{n-1}}^{t_n} \|s(\Theta(t)) - s(\Theta^n)\|^2 dt + C\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2, \tag{70}$$

with the constant C not depending on the discretization parameters. Using (70) and (68), and choosing δ properly, one obtains (59).

To prove (58) we follow the steps in the proof of Lemma 3.3, pg. 296 in [40]. By summing up (15) for $k = 1$ to n and subtracting (11) from the resulting we get for all $w \in L^2(\Omega)$

$$\langle s(\Theta(t_n)) - s^n, w\rangle + \tau\sum_{k=1}^{n}\langle \nabla \cdot (\overline{\mathbf{q}}^n - \mathbf{q}^k), w\rangle = 0. \tag{71}$$

Further, subtracting (12) at $t = t_{k-1}$ from (12) at $t = t_k$, dividing by the time step size τ and subtracting from the result (16) we obtain for all $\mathbf{v} \in H(\mathrm{div};\Omega)$

$$\langle \overline{\mathbf{q}}^n - \mathbf{q}^n, \mathbf{v}\rangle - \langle \overline{\Theta}^n - \Theta^n, \nabla \cdot \mathbf{v}\rangle - \langle \overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \mathbf{v}\rangle = \langle \overline{\mathbf{f}_1(s)}^n - \mathbf{f}_1(s^n), \mathbf{v}\rangle. \tag{72}$$

By testing (71) with $w = \overline{\Theta}^n - \Theta^n \in L^2(\Omega)$ and (72) with $\mathbf{v} = \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k) \in H(\mathrm{div};\Omega)$, adding the results and summing up from $n = 1$ to N we get

$$\begin{aligned} &\sum_{n=1}^{N}\langle s(\Theta(t_n)) - s^n, \overline{\Theta}^n - \Theta^n\rangle + \sum_{n=1}^{N}\tau\langle \overline{\mathbf{q}}^n - \mathbf{q}^n, \sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle \\ &\qquad -\sum_{n=1}^{N}\langle \overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle = \sum_{n=1}^{N}\langle \overline{\mathbf{f}_1(s)}^n - \mathbf{f}_1(s^n), \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle. \end{aligned} \tag{73}$$

We estimate separately each term in (73), which are denoted by T_1, T_2, T_3 and T_4. For T_1 we proceed as in the proof of Lemma 3.3, pg. 296 in [40], as the term here is identical to the one there and obtain

$$T_1 = \frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} \langle s(\Theta) - s^n, \Theta - \Theta^n\rangle\, dt + \frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} \langle s(\Theta(t_n)) - s(\Theta), \Theta - \Theta^n\rangle\, dt \tag{74}$$

with the first term above being positive to remain on the left hand side of (58). Using the regularity of the solutions (both continuous and semi-discrete) and the stability estimates in Proposition 3.4) one can follow the steps in estimating T_{11} in [40] to obtain for the second term above

$$|\frac{1}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} \langle s(\Theta(t_n)) - s(\Theta), \Theta - \Theta^n\rangle| \leq C, \tag{75}$$

with C not depending on the discretization parameters. Moreover, if the data is such that both phases present at any time and everywhere in the system, the estimate in (75) can be improved to $C\tau$, as discussed in Remark 3.3 below. Such estimates are optimal.

For the second term in (73) one uses the algebraic identity (32) to obtain

$$T_2 = \frac{\tau}{2}\|\sum_{n=1}^{N}(\overline{\mathbf{q}}^n - \mathbf{q}^n)\|^2 + \sum_{n=1}^{N}\frac{\tau}{2}\|\overline{\mathbf{q}}^n - \mathbf{q}^n\|^2. \tag{76}$$

The two terms above will remain on the left hand side of (58). We proceed by estimating T_3 in (73). By the Young inequality there holds

$$\begin{aligned}
|T_3| &= |\sum_{n=1}^{N}\langle\overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n, \tau\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\rangle| \\
&\leq \frac{\delta}{2}\sum_{n=1}^{N}\|\overline{f_w(s)\mathbf{u}}^n - f_w(s^n)\mathbf{u}^n\|^2 + \frac{\tau^2}{2\delta}\sum_{n=1}^{N}\|\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\|^2 \\
&\leq T_{31} + \frac{\tau^2}{2\delta}\sum_{n=1}^{N}\|\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\|^2. \qquad (77)
\end{aligned}$$

The second term on the right is estimated by using the Gronwall lemma. For the first one, one uses (A3), (A4) and (59) to obtain

$$\begin{aligned}
|T_{31}| &= \frac{\delta}{2}\sum_{n=1}^{N}\int_\Omega\left(\frac{1}{\tau}\int_{t_{n-1}}^{t_n} f_w(s)\mathbf{u} - f_w(s^n)\mathbf{u}^n\, dt\right)^2 dx \\
&\leq \frac{\delta}{\tau^2}\sum_{n=1}^{N}\int_\Omega\left(\int_{t_{n-1}}^{t_n}(f_w(s) - f_w(s^n))\mathbf{u}\, dt\right)^2 dx + \frac{\delta}{\tau^2}\sum_{n=1}^{N}\int_\Omega f_w^2(s^n)\left(\int_{t_{n-1}}^{t_n}(\mathbf{u} - \mathbf{u}^n)\, dt\right)^2 dx \\
&\leq \frac{\delta M_{\mathbf{u}}^2 L_{f_w}^2}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2 + \delta M_{f_w}^2\sum_{n=1}^{N}\|\overline{\mathbf{u}}^n - \mathbf{u}^n\|^2 \\
&\leq \frac{\delta M_{\mathbf{u}}^2 L_{f_w}^2}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2 + \frac{\delta M_{f_w}^2 C}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2 \\
&\leq \frac{C\delta}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2\, dt, \qquad (78)
\end{aligned}$$

for all $\delta > 0$ and with a constant C not depending on the discretization parameters. In the same manner one can bound the last term in (73). Using again Young's inequality and (A5), one gets

$$|T_4| \leq \frac{C\delta'}{\tau}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2\, dt + \frac{\tau^2}{2\delta'}\sum_{n=1}^{N}\|\sum_{k=1}^{n}(\overline{\mathbf{q}}^k - \mathbf{q}^k)\|^2 \tag{79}$$

for all $\delta' > 0$ and with a constant C not depending on the discretization parameters. We observe that due to (A1) there holds

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|s - s^n\|^2 \leq \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle s - s^n, \Theta - \Theta^n\rangle\, dt. \tag{80}$$

By using (80), choosing δ and δ' properly, the first terms in the right hand sides of (78) and (79) are absorbed in the left hand side of (58). Putting now together (73) - (79) and applying the discrete Gronwall lemma gives (58).

Finally, to prove (60) one follows the step in Lemma 3.9, pg. 300 in [40]. By Lemma 2.1, there exists a function $\mathbf{v} \in H(\mathrm{div};\Omega)$ which satisfies $\nabla \cdot \mathbf{v} = \sum\limits_{n=1}^{N}(\overline{\Theta}^n - \Theta^n)$ and $\|\mathbf{v}\| \leq C_\Omega \|\sum\limits_{n=1}^{N}(\overline{\Theta}^n - \Theta^n)\|$. We use this as test function in (72). Now (60) follows from (58).

Q. E. D.

The next proposition quantifies the error between the semi-discrete solution and the fully discrete one. Recall the notations $s^k = s(\Theta^k)$ and $s_h^k = s(\Theta_h^k)$, $k \in \mathbb{N}$.

Proposition 3.6. *Let $n \in \mathbb{N}, n \geq 1$ and let $(\Theta^n, \mathbf{q}^n, p^n, \mathbf{u}^n)$ be the solution of P^n, and $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n)$ be the solution of P_h^n. Assuming (A1)-(A5) and that the time step is small enough, there holds*

$$\begin{aligned}\sum_{n=1}^{N}\left(\langle s^n - s_h^n, \Theta^n - \Theta_h^n\rangle + \|s^n - s_h^n\|^2\right) + \tau\|\sum_{n=1}^{N}(\Pi_h\mathbf{q}^n - \mathbf{q}_h^n)\|^2 \\ \leq C\sum_{n=1}^{N}(\|\mathbf{q}^n - \Pi_h\mathbf{q}^n\|^2 + \|\Theta^n - P_h\Theta^n\|^2 + \|\mathbf{u}^n - \Pi_h\mathbf{u}^n\|^2 + \|p^n - P_hp^n\|^2)\end{aligned} \tag{81}$$

and

$$\begin{aligned}\|\mathbf{u}^n - \mathbf{u}_h^n\|^2 + \|\nabla\cdot(\mathbf{u}^n - \mathbf{u}_h^n)\|^2 + \|p^n - p_h^n\|^2 \\ \leq C(\|\mathbf{u}^n - \Pi_h\mathbf{u}^n\|^2 + \|s^n - s_h^n\|^2 + \|p^n - P_hp^n\|^2),\end{aligned} \tag{82}$$

with the constants C above not depending on the discretization parameters.

Proof. The proof of (82) can be found in [14], where a MFEM was applied for the discretization of the pressure equation, but the Galerkin FEM for the saturation equation. Therefore we give here only the proof of (81). By subtracting (23) and (24) from (15) and (16), summing up from $k = 1$ to n and using the properties of the projectors, one gets

$$\langle s^n - s_h^n, w_h\rangle + \tau\sum_{k=1}^{n}\langle\nabla\cdot(\Pi_h\mathbf{q}^k - \mathbf{q}_h^k), w_h\rangle = 0, \tag{83}$$

$$\langle\mathbf{q}^n - \mathbf{q}_h^n, \mathbf{v}_h\rangle - \langle P_h\Theta^n - \Theta_h^n, \nabla\cdot\mathbf{v}_h\rangle - \langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \mathbf{v}_h\rangle = \langle\mathbf{f}_1(s^n) - \mathbf{f}_1(s_h^n), \mathbf{v}_h\rangle \tag{84}$$

for all $w_h \in W_h$ and $\mathbf{v}_h \in V_h$. Taking $w_h = P_h\Theta^n - \Theta_h^n \in W_h$ and $\mathbf{v}_h = \tau\sum_{k=1}^{n}(\Pi_h\mathbf{q}^k - \mathbf{q}_h^k) \in V_h$ in (83) and (84), respectively, adding the results and summing up from $n = 1$ to N we obtain

$$\begin{aligned}\sum_{n=1}^{N}\langle s^n - s_h^n, P_h\Theta^n - \Theta_h^n\rangle + \tau\sum_{n=1}^{N}\langle\mathbf{q}^n - \mathbf{q}_h^n, \sum_{k=1}^{n}(\Pi_h\mathbf{q}^k - \mathbf{q}_h^k)\rangle \\ -\sum_{n=1}^{N}\langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \tau\sum_{k=1}^{n}(\Pi_h\mathbf{q}^k - \mathbf{q}_h^k)\rangle = \sum_{n=1}^{N}\langle\mathbf{f}_1(s^n) - \mathbf{f}_1(s_h^n), \tau\sum_{k=1}^{n}(\Pi_h\mathbf{q}^k - \mathbf{q}_h^k)\rangle.\end{aligned} \tag{85}$$

Denoting the terms above by $\hat{T}_1$, $\hat{T}_2$, $\hat{T}_3$ and $\hat{T}_4$, we proceed by estimating them separately. For $\hat{T}_1$ there holds

$$\hat{T}_1 = \sum_{n=1}^{N}\langle s^n - s_h^n, \Theta^n - \Theta_h^n\rangle + \sum_{n=1}^{N}\langle s^n - s_h^n, P_h\Theta^n - \Theta^n\rangle \tag{86}$$

with the first part above being positive due to the monotonicity of $s(\cdot)$. By Young's inequality, for the second term in (86) one gets

$$\sum_{n=1}^{N} \langle s^n - s_h^n, P_h\Theta^n - \Theta^n \rangle \leq \frac{\delta_1}{2} \sum_{n=1}^{N} \|s^n - s_h^n\|^2 + \frac{1}{2\delta_1} \sum_{n=1}^{N} \|P_h\Theta^n - \Theta^n\|^2, \tag{87}$$

for all $\delta_1 > 0$. Note that due to (A1) there holds

$$\sum_{n=1}^{N} \langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle \geq \sum_{n=1}^{N} \frac{1}{L_s} \|s^n - s_h^n\|^2. \tag{88}$$

After properly choosing δ_1, the second term on the right in (87) can be absorbed by $\frac{1}{2}\sum_{n=1}^{N}\langle s^n - s_h^n, \Theta^n - \Theta_h^n \rangle$.

Using the algebraic identity (32), for $\hat{T}_2$ it holds

$$\begin{aligned} \hat{T}_2 &= \tau \sum_{n=1}^{N} \langle \mathbf{q}^n - \Pi_h \mathbf{q}^n, \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \rangle + \tau \sum_{n=1}^{N} \langle \Pi_h \mathbf{q}^n - \mathbf{q}_h^n, \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \rangle \\ &= \hat{T}_{21} + \frac{\tau}{2} \| \sum_{n=1}^{N} (\Pi_h \mathbf{q}^n - \mathbf{q}_h^n) \|^2 + \frac{\tau}{2} \sum_{n=1}^{N} \|\Pi_h \mathbf{q}^n - \mathbf{q}_h^n\|^2. \end{aligned} \tag{89}$$

The only term remaining to be estimated is $\hat{T}_{21}$. This is done by using Young's inequality

$$|\hat{T}_{21}| \leq \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{q}^n - \Pi_h \mathbf{q}^n\|^2 + \frac{\tau^2}{2} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \|^2. \tag{90}$$

In estimating $\hat{T}_3$ we use (A2)-(A4), Young's inequality and (82). There holds

$$\begin{aligned} |\hat{T}_3| &= | \sum_{n=1}^{N} \langle f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n, \tau \sum_{k=1}^{n} \Pi_h \mathbf{q}^k - \mathbf{q}_h^k \rangle | \\ &\leq \frac{\delta_3}{2} \sum_{n=1}^{N} \|f_w(s^n)\mathbf{u}^n - f_w(s_h^n)\mathbf{u}_h^n\|^2 + \frac{\tau^2}{2\delta_3} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \|^2 \\ &\leq C\delta_3 \sum_{n=1}^{N} \|s^n - s_h^n\|^2 + \delta_3 M_{f_w}^2 \sum_{n=1}^{N} \|\mathbf{u}^n - \mathbf{u}_h^n\|^2 + \frac{\tau^2}{2\delta_3} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \|^2 \\ &\leq C\delta_3 \sum_{n=1}^{N} \|s^n - s_h^n\|^2 + C \sum_{n=1}^{N} \|\mathbf{u}^n - \mathbf{u}_h^n\|^2 + \frac{\tau^2}{2\delta_3} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \|^2 \end{aligned} \tag{91}$$

for all $\delta_3 > 0$ and with the constants C not depending on the discretization parameters. In a similar manner, by using (A5) we can bound also the last term $\hat{T}_4$. There holds for all $\delta_4 > 0$

$$|\hat{T}_4| \leq C\delta_4 \sum_{n=1}^{N} \|s^n - s_h^n\|^2 + \frac{\tau^2}{\delta_4} \sum_{n=1}^{N} \| \sum_{k=1}^{n} (\Pi_h \mathbf{q}^k - \mathbf{q}_h^k) \|^2. \tag{92}$$

Finally, putting together (85) - (92), choosing $\delta_1 - \delta_4$ properly, and using the discrete Gronwall lemma we obtain the result (81).

Q. E. D.

The main result below is a straightforward consequence of Proposition 3.5 and Proposition 3.6, the properties of the projectors and the regularity of the solution.

Theorem 3.1. *Let $(\Theta, \mathbf{q}, p, \mathbf{u})$ be the solution of Problem P and let $(\Theta_h^n, \mathbf{q}_h^n, p_h^n, \mathbf{u}_h^n)$ be the solution of P_h^n, $n \in \{1, \dots, N\}$. Assuming (A1)-(A5) and that the time step is small enough, there holds*

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta_h^n), \Theta(t) - \Theta_h^n \rangle \, dt + \| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \mathbf{q} - \mathbf{q}_h^n \, dt \|^2 \leq C(\tau + h^2), \tag{93}$$

$$\sum_{n=1}^{N} \tau \|\bar{\mathbf{u}}^n - \mathbf{u}_h^n\|^2 + \sum_{n=1}^{N} \tau \|\bar{p}^n - p_h^n\|^2 \leq C(\tau + h^2), \tag{94}$$

with the constant C not depending on the discretization parameters.

Remark 3.2. *The error estimates presented above can be extended to the case of $s(\cdot)$ being only Hölder continuous (instead Lipschitz continuous). Following [40] one would get*

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta_h^n), \Theta(t) - \Theta_h^n \rangle \, dt + \| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \mathbf{q} - \mathbf{q}_h^n \, dt \|^2 \leq C(\tau + h^2 \tau^{\frac{2(\alpha-1)}{1+\alpha}}), \tag{95}$$

$$\sum_{n=1}^{N} \tau \|\bar{\mathbf{u}}^n - \mathbf{u}_h^n\|^2 + \sum_{n=1}^{N} \tau \|\bar{p}^n - \mathbf{u}_h^n\|^2 \leq C(\tau + h^2 \tau^{\frac{2(\alpha-1)}{1+\alpha}}), \tag{96}$$

with α being the Hölder exponent of $s(\cdot)$.

Remark 3.3. *In the non-degenerate case, when the disappearance of phases is not allowed, one can obtain the optimal error estimates (similar to Corollary 3.6, pg. 299 in [40])*

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta_h^n), \Theta(t) - \Theta_h^n \rangle \, dt + \| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \mathbf{q} - \mathbf{q}_h^n \, dt \|^2 \leq C(\tau^2 + h^2), \tag{97}$$

$$\sum_{n=1}^{N} \tau \|\bar{\mathbf{u}}^n - \mathbf{u}_h^n\|^2 + \sum_{n=1}^{N} \tau \|\bar{p}^n - p_h^n\|^2 \leq C(\tau^2 + h^2). \tag{98}$$

4 Linearization scheme

In this section we analyze the convergence of the (fully discrete) linearization scheme (27)–(30) proposed to solve the non-linear system (23) – (26). We show that the scheme is robust it converges linearly. The scheme does not involve any regularization step. As the scheme is used to solve the nonlinear systems in one time step, throughout this section $n \in N, n \geq 1$ is fixed. For the ease of the presentation we recall the scheme (27)–(30):

Problem $P_h^{n,i}$: Linearization scheme (L-scheme). Let $L \geq L_s$, $i \in \mathbb{N}, i \geq 1$ and let $\Theta_h^{n,i-1} \in W_h$ be given. Find $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ such that

$$\begin{aligned}
\langle L(\Theta_h^{n,i} - \Theta_h^{n,i-1}) + s_h^{n,i-1}, w_h \rangle + \tau \langle \nabla \cdot \mathbf{q}_h^{n,i}, w_h \rangle &= \langle s_h^{n-1}, w_h \rangle, \\
\langle \mathbf{q}_h^{n,i}, \mathbf{v}_h \rangle - \langle \Theta_h^{n,i}, \nabla \cdot \mathbf{v}_h \rangle - \langle f_w(s_h^{n,i-1}) \mathbf{u}_h^{n,i}, \mathbf{v}_h \rangle &= \langle \mathbf{f}_1(s_h^{n,i-1}), \mathbf{v}_h \rangle, \\
\langle \nabla \cdot \mathbf{u}_h^{n,i}, w_h \rangle &= \langle f_2(s_h^{n,i-1}), w_h \rangle, \\
\langle a(s_h^{n,i-1}) \mathbf{u}_h^{n,i}, \mathbf{v}_h \rangle - \langle p_h^{n,i}, \nabla \cdot \mathbf{v}_h \rangle + \langle \mathbf{f}_3(s_h^{n,i-1}), \mathbf{v}_h \rangle &= 0
\end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. Here we use the notation $s_h^{n,i} := s(\Theta_h^{n,i})$ and, as in the previous section, $s_h^n := s(\Theta_h^n)$ and $s_h^{n-1} := s(\Theta_h^{n-1})$. Also, the iteration starts with $\Theta_h^{n,0} := \Theta_h^{n-1}$ and $s_h^{n,0} := s(\Theta_h^{n-1})$.

We introduce now the errors between two consecutive iterations i and $i-1$:

$$\begin{array}{lllllll} e_\Theta^{n,i} &=& \Theta_h^{n,i} - \Theta_h^{n,i-1}, & & e_{\mathbf{q}}^{n,i} &=& \mathbf{q}_h^{n,i} - \mathbf{q}_h^{n,i-1}, \\ e_p^{n,i} &=& p_h^{n,i} - p_h^{n,i-1}, & & e_{\mathbf{u}}^{n,i} &=& \mathbf{u}_h^{n,i} - \mathbf{u}_h^{n,i-1}, \\ e_s^{n,i} &=& s_h^{n,i} - s_h^{n,i-1} := s(\Theta_h^{n,i}) - s(\Theta_h^{n,i-1}). & & & & \end{array}$$

In order to show the convergence of the scheme (27) - (30) we prove that the iteration is a contraction in e_Θ and $e_{\mathbf{q}}$, together with estimates for e_p and $e_{\mathbf{u}}$. The convergence follows by applying the Banach fixed point theorem for e_Θ and $e_{\mathbf{q}}$ and by a similar argument for e_p and $e_{\mathbf{u}}$. Note that, in this case, the term involving the factor L will vanish and the limit of the iteration is a solution of Problem P_h^n.

The main result, the convergence of the scheme (27) – (30), is stated in the following

Theorem 4.1. *Assuming (A1)-(A5), if the time step τ is sufficiently small, the linearization scheme* (27) -(30) *converges linearly.*

In fact, the time step has to satisfy a mild restriction to guarantee the convergence of the scheme. This restriction is stated in (119) below. In particular, the convergence is robust w.r.t. the mesh size h. Moreover, the errors are reduced faster for larger values of τ than for smaller values, however still satisfying (119).

The proof of Theorem 4.1 follows directly from Lemma 4.1 and Lemma 4.2 below.

Lemma 4.1. *Let $n \in \mathbb{N}$ be fixed, and $\Theta_h^{n-1}, p_h^{n-1} \in W_h$ and $\mathbf{q}_h^{n-1}, \mathbf{u}_h^{n-1} \in V_h$ be given, solving P_h^{n-1}. Further, let $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ solving $P_h^{n,i}$ for any $i \geq 1, i \in \mathbb{N}$. Assuming (A2)–(A4), there holds*

$$\|e_{\mathbf{u}}^{n,i}\|^2 \leq C_1 \|e_s^{n,i-1}\|^2, \tag{99}$$

$$\|\nabla \cdot e_{\mathbf{u}}^{n,i}\|^2 \leq L_{f_2}^2 \|e_s^{n,i-1}\|^2, \tag{100}$$

$$\|e_p^{n,i}\|^2 \leq C_2 \|e_s^{n,i-1}\|^2, \tag{101}$$

where C_1 and C_2 are two constants not depending on the discretization parameters or the iteration index.

Proof. The constants C_1 and C_2 are determined in the proof. With $i > 1$, subtracting (29) and (30) for i and $i-1$ respectively, one obtains

$$\begin{aligned} \langle \nabla \cdot e_{\mathbf{u}}^{n,i}, w_h \rangle &= \langle f_2(s_h^{n,i-1}) - f_2(s_h^{n,i-2}), w_h \rangle, & (102) \\ \langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i} - a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h \rangle - \langle e_p^{n,i}, \nabla \cdot \mathbf{v}_h \rangle &= \langle \mathbf{f}_3(s_h^{n,i-2}) - \mathbf{f}_3(s_h^{n,i-1}), \mathbf{v}_h \rangle & (103) \end{aligned}$$

for all $w_h \in W_h$, $\mathbf{v}_h \in V_h$. Taking now $w_h = e_p^{n,i} \in W_h$ in (102) and $\mathbf{v}_h = e_{\mathbf{u}}^{n,i} \in V_h$ in (102), and adding the results one gets

$$\langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i} - a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, e_{\mathbf{u}}^{n,i} \rangle = \langle f_2(s_h^{n,i-1}) - f_2(s_h^{n,i-2}), e_p^{n,i} \rangle + \langle \mathbf{f}_3(s_h^{n,i-2}) - \mathbf{f}_3(s_h^{n,i-1}), e_{\mathbf{u}}^{n,i} \rangle. \tag{104}$$

Using (A2) - (A4), together with Young's inequality, from (104) and for any $\epsilon_1 > 0$ one gets

$$\frac{a_\star}{2} \|e_{\mathbf{u}}^{n,i}\|^2 \leq \left(\frac{M_{\mathbf{u}}^2 L_a^2 + L_{\mathbf{f}_3}^2}{a_\star} + \frac{L_{f_2}^2}{2\epsilon_1} \right) \|e_s^{n,i-1}\|^2 + \frac{\epsilon_1}{2} \|e_p^{n,i}\|^2. \tag{105}$$

Recalling Lemma 2.2, a $\mathbf{v}_h \in V_h$ exists such that $\nabla \cdot \mathbf{v}_h = e_p^{n,i}$ and $\|\mathbf{v}_h\| \leq C_{\Omega,d}\|e_p^{n,i}\|$. Taking this $\mathbf{v}_h$ as test function in (103), using (A2)-(A4) gives

$$\begin{aligned}\|e_p^{n,i}\|^2 &= \langle a(s_h^{n,i-1})\mathbf{u}_h^{n,i} - a(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h\rangle + \langle \mathbf{f}_3(s_h^{n,i-2}) - \mathbf{f}_3(s_h^{n,i-1}), \mathbf{v}_h\rangle \\ &\leq C_{\Omega,d}\left(a^\star\|e_{\mathbf{u}}^{n,i}\| + (M_{\mathbf{u}}L_a + L_{\mathbf{f}_3})\|e_s^{n,i-1}\|\right)\|e_p^{n,i}\|.\end{aligned}$$

This allows estimating $e_p^{n,i}$ in terms of $e_s^{n,i}$ and $e_{\mathbf{u}}^{n,i}$,

$$\|e_p^{n,i}\|^2 \leq 2C_{\Omega,d}^2(a^\star)^2\|e_{\mathbf{u}}^{n,i}\|^2 + 2C_{\Omega,d}^2(M_{\mathbf{u}}L_a + L_{\mathbf{f}_3})^2\|e_s^{n,i-1}\|^2. \tag{106}$$

With $\epsilon_1 = \dfrac{a_\star}{4(a^\star)^2C_{\Omega,d}^2}$, from (105) and (106) one obtains

$$\frac{a_\star}{4}\|e_{\mathbf{u}}^{n,i}\|^2 \leq \left(\frac{M_{\mathbf{u}}^2L_a^2 + L_{\mathbf{f}_3}^2}{a_\star} + \frac{2(a^\star)^2C_{\Omega,d}^2L_{f_2}^2}{a_\star} + \frac{a_\star(M_{\mathbf{u}}L_a + L_{\mathbf{f}_3})^2}{4(a^\star)^2}\right)\|e_s^{n,i-1}\|^2, \tag{107}$$

which is, in fact (99) with $C_1 = 4\dfrac{M_{\mathbf{u}}^2L_a^2 + L_{\mathbf{f}_3}^2 + 2(a^\star)^2C_{\Omega,d}^2L_{f_2}^2}{a_\star^2} + \dfrac{(M_{\mathbf{u}}L_a + L_{\mathbf{f}_3})^2}{(a^\star)^2}$. Further, (101) follows immediately from (106) and (99), with $C_2 = 2C_{\Omega,d}^2(a^\star)^2C_1 + 2C_{\Omega,d}^2(M_{\mathbf{u}}L_a + L_{\mathbf{f}_3})^2$. Finally, (100) is a straightforward consequence of (102) and (A3). **Q. E. D.**

Lemma 4.2. *Let $n \in \mathbb{N}$ be fixed, $\Theta_h^{n-1}, p_h^{n-1} \in W_h$ and $\mathbf{q}_h^{n-1}, \mathbf{u}_h^{n-1} \in V_h$ solve $\mathbf{P_h^{n-1}}$ and $\Theta_h^{n,i}, p_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i}, \mathbf{u}_h^{n,i} \in V_h$ solve $\mathbf{P_h^{n,i}}$ for any $i \geq 1, i \in \mathbb{N}$. Assuming (A1)–(A4), it holds*

$$\|e_{\mathbf{q}}^{n,i}\|^2 \geq \frac{1}{3C_\Omega^2}\|e_\Theta^{n,i}\|^2 - C_3\|e_s^{n,i-1}\|^2, \tag{108}$$

and

$$\begin{aligned}&\|e_\Theta^{n,i}\|^2 + \frac{3C_{\Omega,d}^2\tau}{2(3C_{\Omega,d}^2L + \tau)}\|e_{\mathbf{q}}^{n,i}\|^2 \\ &\quad + \frac{3C_{\Omega,d}^2\left(1 - \tau L(C_3 + 4L_{\mathbf{f}_1}^2 + 8L_{f_w}^2M_{\mathbf{u}}^2 + 8C_1M_{f_w}^2)\right)}{L(3C_{\Omega,d}^2L + \tau)}\|e_s^{n,i-1}\|^2 \leq \frac{3C_{\Omega,d}^2L}{3C_{\Omega,d}^2L + \tau}\|e_\Theta^{n,i-i}\|^2,\end{aligned} \tag{109}$$

where $C_3 = M_{f_w}^2C_1 + (M_{f_w}L_{f_w} + L_{\mathbf{f}_1})^2$ is not depending on the discretization parameters.

Proof. Subtracting (27) and (28) for i and $i-1$ respectively gives

$$\langle L(e_\Theta^{n,i} - e_\Theta^{n,i-1}) + e_s^{n,i-1}, w_h\rangle + \tau\langle\nabla\cdot e_{\mathbf{q}}^{n,i}, w_h\rangle = 0, \tag{110}$$

$$\langle e_{\mathbf{q}}^{n,i}, \mathbf{v}_h\rangle - \langle e_\Theta^{n,i}, \nabla\cdot\mathbf{v}_h\rangle - \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h\rangle = \langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), \mathbf{v}_h\rangle. \tag{111}$$

By Lemma 2.2, there exists a $\mathbf{v}_h \in V_h$ such that $\nabla\cdot\mathbf{v}_h = e_\Theta^{n,i}$ and $\|\mathbf{v}_h\| \leq C_{\Omega,d}\|e_\Theta^{n,i}\|$. Taking this $\mathbf{v}_h$ as test function in (111) and using (A3) and (A4) gives

$$\begin{aligned}\|e_\Theta^{n,i}\|^2 &= \langle e_{\mathbf{q}}^{n,i}, \mathbf{v}_h\rangle + \langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-1}, \mathbf{v}_h\rangle + \langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), \mathbf{v}_h\rangle \\ &\leq \left(\|e_{\mathbf{q}}^{n,i}\| + \|f_w(s_h^{n,i-1})e_{\mathbf{u}}^{n,i}\| + \|(f_w(s_h^{n,i-1}) - f_w(s_h^{n,i-2}))\mathbf{u}_h^{n,i-1}\| + L_{\mathbf{f}_1}\|e_s^{n,i-1}\|\right)\|\mathbf{v}_h\| \\ &\leq \left(\|e_{\mathbf{q}}^{n,i}\| + M_{f_w}\|e_{\mathbf{u}}^{n,i}\| + (M_{\mathbf{u}}L_{f_w} + L_{\mathbf{f}_1})\|e_s^{n,i-1}\|\right)C_{\Omega,d}\|e_\Theta^{n,i}\|.\end{aligned} \tag{112}$$

Combining (112) with (99) further implies

$$\|e_\Theta^{n,i}\|^2 \leq 3C_{\Omega,d}^2\|e_\mathbf{q}^{n,i}\|^2 + 3C_{\Omega,d}^2 C_3\|e_s^{n,i-1}\|^2. \tag{113}$$

where $C_3 = M_{f_w}^2 C_1 + (M_\mathbf{u} L_{f_w} + L_{\mathbf{f}_1})^2$. From (113), (108) follows immediately.

To prove (109), one takes $w_h = e_\Theta^{n,i} \in W_h$ in (110) and $\mathbf{v}_h = \tau e_\mathbf{q}^{n,i}$ in (111), add the resulting and obtains

$$\begin{aligned} &L\langle (e_\Theta^{n,i} - e_\Theta^{n,i-1}) + e_s^{n,i-1}, e_\Theta^{n,i}\rangle + \tau\|e_\mathbf{q}^{n,i}\|^2 = \\ &\qquad\qquad \tau\langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-2}, e_\mathbf{q}^{n,i}\rangle + \tau\langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), e_\mathbf{q}^{n,i}\rangle. \end{aligned}$$

This further implies

$$\begin{aligned} \frac{L}{2}\|e_\Theta^{n,i}\|^2 + \frac{L}{2}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle + \tau\|e_\mathbf{q}^{n,i}\|^2 = \frac{L}{2}\|e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i-1} - e_\Theta^{n,i}\rangle \\ +\tau\langle f_w(s_h^{n,i-1})\mathbf{u}_h^{n,i} - f_w(s_h^{n,i-2})\mathbf{u}_h^{n,i-2}, e_\mathbf{q}^{n,i}\rangle + \tau\langle \mathbf{f}_1(s_h^{n,i-1}) - \mathbf{f}_1(s_h^{n,i-2}), e_\mathbf{q}^{n,i}\rangle. \end{aligned} \tag{114}$$

By the monotonicity and the Lipschitz continuity of $s(\cdot)$ as stated in (A1) there holds

$$\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle \geq \frac{1}{L_s}\|e_s^{n,i-1}\|^2 \geq \frac{1}{L}\|e_s^{n,i-1}\|^2. \tag{115}$$

From (108) and (115), by (A1) - (A4) and Young's inequality, (114) implies

$$\begin{aligned} &(\frac{L}{2} + \frac{\tau}{6C_{\Omega,d}^2})\|e_\Theta^{n,i}\|^2 + \frac{L}{2}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|^2 + (\frac{1}{L} - \tau\frac{C_3}{2})\|e_s^{n,i-1}\|^2 + \frac{\tau}{2}\|e_\mathbf{q}^{n,i}\|^2 \\ &\quad \leq \frac{L}{2}\|e_\Theta^{n,i-1}\|^2 + \frac{1}{2L}\|e_s^{n,i-1}\|^2 + \frac{L}{2}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|^2 + 2\tau L_{\mathbf{f}_1}^2\|e_s^{n,i-1}\|^2 \\ &\quad +\frac{\tau}{8}\|e_\mathbf{q}^{n,i}\|^2 + 4\tau L_{f_w}^2 M_\mathbf{u}^2\|e_s^{n,i-1}\|^2 + 4\tau M_{f_w}^2\|e_\mathbf{u}^{n,i}\|^2 + \frac{\tau}{8}\|e_\mathbf{q}^{n,i}\|^2. \end{aligned} \tag{116}$$

This rewrites as

$$\begin{aligned} &(\frac{L}{2} + \frac{\tau}{6C_{\Omega,d}^2})\|e_\Theta^{n,i}\|^2 + (\frac{1}{L} - \tau\frac{C_3}{2})\|e_s^{n,i-1}\|^2 + \frac{\tau}{4}\|e_\mathbf{q}^{n,i}\|^2 \\ &\quad \leq \frac{L}{2}\|e_\Theta^{n,i-1}\|^2 + (\frac{1}{2L} + 2\tau L_{\mathbf{f}_1}^2 + 4\tau L_{f_w}^2 M_\mathbf{u}^2)\|e_s^{n,i-1}\|^2 + 4\tau M_{f_w}^2\|e_\mathbf{u}^{n,i}\|^2. \end{aligned} \tag{117}$$

Using now (99) in (117) and rearranging the terms leads to

$$\begin{aligned} &(\frac{L}{2} + \frac{\tau}{6C_{\Omega,d}^2})\|e_\Theta^{n,i}\|^2 + \left(\frac{1}{2L} - \tau(\frac{C_3}{2} + 2L_{\mathbf{f}_1}^2 + 4L_{f_w}^2 M_\mathbf{u}^2 + 4C_1 M_{f_w}^2)\right)\|e_s^{n,i-1}\|^2 \\ &\quad +\frac{\tau}{4}\|e_\mathbf{q}^{n,i}\|^2 \leq \frac{L}{2}\|e_\Theta^{n,i-1}\|^2, \end{aligned} \tag{118}$$

which is nothing else as the result (109). **Q. E. D.**

Remark 4.1. *The estimate* (109) *is not practical unless the factor multiplying the last term on the left is positive. This gives a restriction on the time step,*

$$\tau \leq \frac{1}{L\left(C_3 + 4L_{\mathbf{f}_1}^2 + 8L_{f_w}^2 M_\mathbf{u}^2 + 8C_1 M_{f_w}^2\right)}, \tag{119}$$

where $C_1 = 4\frac{M_\mathbf{u}^2 L_a^2 + L_{\mathbf{f}_3}^2 + (a^\star)^2 C_{\Omega,d}^2 L_{f_2}^2}{a_\star^2} + \frac{2(M_\mathbf{u} L_a + L_{\mathbf{f}_3})^2}{(a^\star)^2}$ *and* $C_3 = M_{f_w}^2 C_1 + (M_\mathbf{u} L_{f_w} + L_{\mathbf{f}_1})^2$*. This is a mild condition because it does not depend on the grid size. In this sense, it is superior to the conditions guaranteeing the stability of an explicit scheme in time, or to the typical conditions guaranteeing the convergence of the Newton method for degenerate parabolic problems (see e.g. [39, 41]).*

Remark 4.2. *The constant L is independent on the discretization parameters, but must be chosen greater than the Lipschitz constant of the function $s(\cdot)$, i.e. L_s.*

Remark 4.3. *One can use the L-scheme* (27)–(30) *also in combination with the Newton method. The goal is to combine the robustness of the L-scheme, which converges regardless of the starting point, with the quadratic convergence of the Newton Method, which requires instead a starting point close to the solution. Specifically, at each time step one can perform a few L-scheme iterations, followed by Newton iterations. In this way one enhances the robustness of the Newton method and reduces the severe restriction on the time step that guarantees its convergence. We refer to [26], where this strategy is studied for solving the Richards equation. Still for the Richards equation, A similar idea was also proposed in [25], but there the modified Picard method was used to improve the robustness of the Newton method.*

4.1 The Hölder continuous case

The convergence of the L-scheme (27)–(30) is proved rigorously for the case of Lipschitz continuous $s(\cdot)$. Since commonly used parametrizations, e.g. van Genuchten-Mualem for porous media flow models do not necessary lead to a Lipschitz continuous $s(\cdot)$ but Hölder continuous, one question appearing naturally is whether the L-scheme can also be used for such cases too. Note that then the derivative $s'(\cdot)$ becomes unbounded, so Newton's method cannot be applied without regularization, i.e. approximating the function $s(\cdot)$ by a function $s_\epsilon(\cdot)$ that is Lipschitz continuous, but with a Lipschitz constant that goes to infinity as the small regularization parameter ϵ approaches 0. Since the Newton scheme converges conditionally, this would also make the restriction on the time step even more severe, as it would depend on ϵ as well. Clearly, in this case alternative schemes are needed.

We note that the case when $s(\cdot)$ is only the Hölder continuous has not been discussed so far. Previous papers dealing with the L-scheme (see [49, 45, 37, 26] for the Richards equation and [42] for a multipoint flux approximation method for a two-phase flow model) only assume Lipschitz continuous functions. Here we show how to use the L-scheme for the Hölder continuous case as well. however, the convergence will be in general slower.

In this section, assumption (A1) is replaced by

(A1H) The function $s : \mathbb{R} \to [0,1]$ is monotone increasing and Hölder continuous, i.e. there exist $L_s > 0$ and $\alpha \in (0,1)$ such that

$$|s(x) - s(y)| \leq L_s |x-y|^\alpha \qquad \text{for all} \quad x, y \in \mathbb{R}. \tag{120}$$

For the ease of the presentation we consider a case corresponding to the Richards equation without gravity. However, the analysis can be extended to the general situation proceeding as in the proof of Theorem 4.1 and Lemmata 4.1 and 4.2. In this section we let $\|\cdot\|_p$ denote the norm in $L^p(\Omega)$. In the context described previously, we omit the original model and give directly its Euler implicit MFEM discretization

Problem PH_h^n: Fully discrete scheme for the Richards equation without gravity. Let $\Theta_h^{n-1} \in W_h$ be given. Find $\Theta_h^n \in W_h$ and $\mathbf{q}_h^n \in V_h$ such that

$$\begin{aligned} \langle s_h^n - s_h^{n-1}, w_h\rangle + \tau\langle \nabla\cdot \mathbf{q}_h^n, w_h\rangle &= 0\rangle, & (121)\\ \langle \mathbf{q}_h^n, \mathbf{v}_h\rangle - \langle \Theta_h^n, \nabla\cdot\mathbf{v}_h\rangle &= 0, & (122)\end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$.

We refer to [38, 40] for the existence and uniqueness of a solution to the algebraic systems above.

Observe that this problem is nonlinear. As for the two-phase model, a linear iteration scheme is introduced in

Problem $PH_h^{n,i}$: Linearization scheme (L-scheme) for the Hölder continuous case. Let $\epsilon > 0$, $L = \dfrac{1}{\epsilon}$, $i \in \mathbb{N}, i \geq 1$ and let $\Theta_h^{n,i-1} \in W_h$ be given. Find $\Theta_h^{n,i} \in W_h$ and $\mathbf{q}_h^{n,i} \in V_h$ such that

$$\begin{aligned} \langle L(\Theta_h^{n,i} - \Theta_h^{n,i-1}) + s_h^{n,i-1}, w_h\rangle + \tau\langle \nabla \cdot \mathbf{q}_h^{n,i}, w_h\rangle &= \langle s_h^{n-1}, w_h\rangle, & (123) \\ \langle \mathbf{q}_h^{n,i}, \mathbf{v}_h\rangle - \langle \Theta_h^{n,i}, \nabla \cdot \mathbf{v}_h\rangle &= 0, & (124) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$.
As before, we use the notations $s_h^{n,i} := s(\Theta_h^{n,i})$, $s_h^n := s(\Theta_h^n)$, $n \in \mathbb{N}$. Since now the existence of a solution in of the non-linear problem PH_h^n is established, the errors are defined as $e_\Theta^{n,i} = \Theta_h^{n,i} - \Theta_h^n$, $e_\mathbf{q}^{n,i} = \mathbf{q}_h^{n,i} - \mathbf{q}_h^n$, $e_s^{n,i} = s_h^{n,i} - s_h^n$, where Θ_h^n, $\mathbf{q}_h^n$ are the solution of
For proving the convergence of the L-scheme, we need

Theorem 4.2. *Let $\epsilon > 0$, $L = \dfrac{1}{\epsilon}$, $n, i \in \mathbb{N}$ and $\Theta_h^{n-1}, \Theta_h^{n,i-1} \in W_h$ be given. Let $(\Theta_h^n, \mathbf{q}_h^n), (\Theta_h^{n,i}, \mathbf{q}_h^{n,i}) \in W_h \times V_h$ the solutions of Problem PH_h^n and $PH_h^{n,i}$, respectively. Assuming (AH1), there holds:*

$$\|e_\Theta^{n,i}\|^2 + \tau\epsilon R(\epsilon,\tau)\|e_\mathbf{q}^{n,i}\|^2 \leq R(\epsilon,\tau)\|e_\Theta^{n,i-1}\|^2 + 2C(\alpha)R(\epsilon,\tau)\epsilon^{\frac{2}{1-\alpha}}, \tag{125}$$

where $R(\epsilon,\tau) = (1 + \dfrac{\tau\epsilon}{C_{\Omega_d}^2})^{-1}, C(\alpha) = (L_s(2\alpha)^\alpha)^{\frac{2}{1-\alpha}} \frac{(1-\alpha)(1+\alpha)^{\frac{1+\alpha}{\alpha-1}}}{2}$.

Observe that since $R(\epsilon,\tau) < 1$ and ϵ has a positive power in the last term on the right of (125), this gives a theoretical convergence of the scheme. A more detailed discussion is in Remark 4.4.
Proof. Subtracting (121) and (122) from (123) and (124), respectively, one get the error equations

$$\begin{aligned} \langle L(e_\Theta^{n,i} - e_\Theta^{n,i-1}) + e_s^{n,i-1}, w_h\rangle + \tau\langle \nabla \cdot e_\mathbf{q}^{n,i}, w_h\rangle &= 0, & (126) \\ \langle e_\mathbf{q}^{n,i}, \mathbf{v}_h\rangle - \langle e_\Theta^{n,i}, \nabla \cdot \mathbf{v}_h\rangle &= 0, & (127) \end{aligned}$$

for all $w_h \in W_h$ and all $\mathbf{v}_h \in V_h$. By testing (126) with $w_h = e_\Theta^{n,i} \in W_h$ and (127) with $\mathbf{v}_h = \tau e_\mathbf{q}^{n,i} \in V_h$, and adding the results one obtains

$$\langle L(e_\Theta^{n,i} - e_\Theta^{n,i-1}), e_\Theta^{n,i}\rangle + \langle e_s^{n,i-1}, e_\Theta^{n,i}\rangle + \tau\|e_\mathbf{q}^{n,i}\|^2 = 0, \tag{128}$$

which is further equivalent to

$$\frac{L}{2}\|e_\Theta^{n,i}\|^2 + \frac{L}{2}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle + \tau\|e_\mathbf{q}^{n,i}\|^2 = \frac{L}{2}\|e_\Theta^{n,i-1}\|^2 + \langle e_s^{n,i-1}, e_\Theta^{n,i} - e_\Theta^{n,i-1}\rangle. \tag{129}$$

By using the monotonicity and Hölder continuity of $s(\cdot)$, we have

$$\langle e_s^{n,i-1}, e_\Theta^{n,i-1}\rangle \geq \frac{1}{L_s^{1/\alpha}}\|e_s^{n,i-1}\|_{\frac{1+\alpha}{\alpha}}^{\frac{1+\alpha}{\alpha}}. \tag{130}$$

With $p = \frac{1+\alpha}{\alpha}$, $q = 1 + \alpha$, $a = \frac{|e_s^{n,i-1}|}{L_s^{\frac{1}{1+\alpha}}(\frac{2\alpha}{1+\alpha})^{\frac{\alpha}{1+\alpha}}}$ and $b = L_s^{\frac{1}{1+\alpha}}(\frac{2\alpha}{1+\alpha})^{\frac{\alpha}{1+\alpha}}|e_\Theta^{n,i} - e_\Theta^{n,i-1}|$, by (33) one gets

$$|\langle e_s^{n,i-1}, e_\Theta^{n,i} - e_\Theta^{n,i-1}\rangle| \leq \frac{\|e_s^{n,i-1}\|_{\frac{1+\alpha}{\alpha}}^{\frac{1+\alpha}{\alpha}}}{2L_s^{1/\alpha}} + \frac{2^\alpha L_s\alpha^\alpha}{(\alpha+1)^{(\alpha+1)}}\|e_\Theta^{n,i} - e_\Theta^{n,i-1}\|_{1+\alpha}^{1+\alpha}. \tag{131}$$

Combining (131) with (129) and (130) one immediately obtains

$$\frac{L}{2}\|e_\Theta^{n,i}\|^2+\frac{L}{2}\|e_\Theta^{n,i}-e_\Theta^{n,i-1}\|^2+\frac{1}{2}\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\tau\|e_{\mathbf{q}}^{n,i}\|^2 \leq \frac{L}{2}\|e_\Theta^{n,i-1}\|^2+\frac{2^\alpha L_s\alpha^\alpha}{(\alpha+1)^{(\alpha+1)}}\|e_\Theta^{n,i}-e_\Theta^{n,i-1}\|_{1+\alpha}^{1+\alpha}. \tag{132}$$

Similarly, with $p=\frac{2}{1+\alpha}$, $q=\frac{2}{1-\alpha}$, $a=|e_\Theta^{n,i}-e_\Theta^{n,i-1}|^{1+\alpha}(\frac{L}{1+\alpha})^{\frac{1+\alpha}{2}}$ and $b=\frac{(2\alpha)^\alpha L_s}{(\alpha+1)^{(\alpha+1)}}(\frac{1+\alpha}{L})^{\frac{1+\alpha}{2}}$, (33) gives

$$\frac{2^\alpha L_s\alpha^\alpha}{(\alpha+1)^{(\alpha+1)}}\|e_\Theta^{n,i}-e_\Theta^{n,i-1}\|_{1+\alpha}^{1+\alpha} \leq \frac{L}{2}\|e_\Theta^{n,i}-e_\Theta^{n,i-1}\|^2+C(\alpha)L^{\frac{1+\alpha}{1-\alpha}}, \tag{133}$$

where

$$C(\alpha)=(L_s(2\alpha)^\alpha)^{\frac{2}{1-\alpha}}\frac{(1-\alpha)(1+\alpha)^{\frac{1+\alpha}{\alpha-1}}}{2}. \tag{134}$$

The inequalities (132) and (133) imply

$$\frac{L}{2}\|e_\Theta^{n,i}\|^2+\frac{1}{2}\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\tau\|e_{\mathbf{q}}^{n,i}\|^2 \leq \frac{L}{2}\|e_\Theta^{n,i-1}\|^2+C(\alpha)L^{\frac{1+\alpha}{1-\alpha}}, \tag{135}$$

with $C(\alpha)$ defined in (134). Using now Lemma 2.2, and multiplying by 2 one gets from (135)

$$(L+\frac{\tau}{C_{\Omega,d}^2})\|e_\Theta^{n,i}\|^2+\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle+\tau\|e_{\mathbf{q}}^{n,i}\|^2 \leq L\|e_\Theta^{n,i-1}\|^2+2C(\alpha)L^{\frac{1+\alpha}{1-\alpha}}. \tag{136}$$

Recalling that $L=\frac{1}{\epsilon}$, (125) follows immediately from (136). **Q. E. D.**

Remark 4.4. *Practically, the term* $\frac{2C(\alpha)}{1+\frac{\tau\epsilon}{C_{\Omega_d}^2}}\epsilon^{\frac{2}{1-\alpha}}$ *is very small. For* $\alpha=\frac{3}{4}$, $\alpha=\frac{1}{2}$ *or* $\alpha=\frac{1}{4}$ *the power of* ϵ *in this term, i.e.* $\frac{2}{1-\alpha}$ *being* $8, 4$ *and* $\frac{8}{3}$. *This means that, for any reasonable* α, $\epsilon<1$ *needs not to be very small, but moderate. For example,* $\alpha=\frac{3}{4}$ *and* $\epsilon=0.2$ *gives* $\epsilon^{\frac{2}{1-\alpha}}=2.56\cdot10^{-6}$. *Additionally, the number* $C(\alpha)$ *is small too. In the situation above, if* $L_s=0.5$ *it is of order* 10^{-4}.

Remark 4.5. *The convergence rate* $R(\epsilon,\tau)=(1+\frac{\tau\epsilon}{C_{\Omega_d}^2})^{-1}$ *of the linearization scheme* (123)-(124) *is now greater as in the Lipschitz continuous case (in that case one has* $R(\tau)=(1+\frac{\tau}{C_{\Omega_d}^2})^{-1}$*), so the convergence is slower. Nevertheless, as mentioned above, in most of the practical relevant cases* ϵ *can be chosen* 0.1 *or even bigger, so we still get a good convergence for the Hölder case as well.*

Remark 4.6. *The convergence of the linearization scheme for the Hölder continuous case was rigorously established in Theorem 4.2 only for the Richards equation. In the general case, one follows the lines of Theorem 4.1 and Theorem 4.2, and uses the positive term* $\langle e_s^{n,i-1},e_\Theta^{n,i-1}\rangle$ *in* (136) *to absorb the remaining terms. This will lead to an additional constraint on the time step size (similar to* (119)*). We point out that the Lipschitz continuity in assumptions (A2)-(A3) should be now replaced by inequalities of the type mentioned in Remark 2.2.*

5 Numerical results

In this section we present two numerical studies, one concentrating on the convergence of the backward Euler/MFEM discretization and one on the convergence of the linearization scheme. For more numerical examples we refer to [40, 38] (for convergence of the discretization error) and [26] (for linearization schemes). These papers are considering Richards' equation (which is just a particular case of the two-phase flow considered in this paper). We further refer to [42] for an example concerning the linearization method for two-phase flow and MPFA.

We consider here two problems. The first is defined in a two-dimensional domain $\Omega = (0,1) \times (0,1)$ and has the analytical solution given in (137). For this, a source term was added to (7):

$$\begin{aligned} f(t,x,y) = 2tx^2(1-x)^2y^2(1-y)^2 + 2tx(1-x) + 2ty(1-y) \\ +t^2y^3(1-y)^3\left(10x^4 - 20x^3 + 12x^2 - 2x\right) \\ +t^2x^3(1-x)^3\left(10y^4 - 20y^3 + 12y^2 - 2y\right), \end{aligned}$$

and we choose appropriate initial and (Dirichlet) boundary conditions. For the spatial discretization, we use a rectangular and uniform mesh, whereas for the time discretization we choose a uniform time step with final time $T = 1$. In accordance with the estimate in Theorem 3.1, we will consider a sequence of discretizations with halving the spatial mesh size h and reducing the time step τ one-fourth.

Recalling the system of equations (7) - (10), the solution and coefficient functions are given by

$$\begin{array}{llll} p = x(1-x)y(1-y), & \Theta = tx(1-x)y(1-y), & s(\Theta) = \Theta^2, \ \lambda_w = s, & \\ \lambda_o = 1 - s, & \mathbf{f}_1 = 0, & f_2 = 2x(1-x) + 2y(1-y), & \mathbf{f}_3 = 0. \end{array} \tag{137}$$

h	τ	E_p	conv rate	$E_{s\Theta}$	conv rate	E_Θ	conv rate	E_s	conv rate
$\frac{1}{4}$	$\frac{1}{5}$	$1.96E-4$	–	3.35E-6	–	9.14 E-5	–	$1.82E-7$	–
$\frac{1}{8}$	$\frac{1}{20}$	$4.48E-5$	2.13	5.53E-7	2.59	1.66E-5	2.46	$2.88E-8$	2.66
$\frac{1}{16}$	$\frac{1}{80}$	$1.11E-5$	2.01	1.28 E-7	2.11	3.9 E-6	2.09	$6.67E-9$	2.11
$\frac{1}{32}$	$\frac{1}{320}$	$2.72E-6$	2.00	3.12E -8	2.04	9.53 E-7	2.03	$1.61E-9$	2.05

Table 1: Convergence rates for the manufactured solution (137).

The results are presented in Table 1, with the errors given by:

$$E_p = \sum_{n=1}^{N} \tau \|\overline{p}^n - p_h^n\|^2, \qquad E_\Theta = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \|\Theta(t) - \Theta_h^n\|^2\, dt,$$

$$E_s = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \|s(t) - s_h^n\|^2\, dt, \quad E_{s\Theta} = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \langle s(\Theta(t)) - s(\Theta_h^n), \Theta(t) - \Theta_h^n\rangle\, dt.$$

The convergence rate is computed by

$$\text{conv rate}(i) = \frac{\log E_z(i+1) - \log E_z(i)}{\log h(i+1) - \log h(i)},$$

where i is the array index and z stands for either p, Θ, s, or $s\Theta$, as shown in Table 1. We see that the convergence rate is 2 which is as expected.

Next, we discuss another example where we consider 3D rectangular grids of different sizes and study the convergence of linearization scheme. The computational domain is now the unit cube. We use the following constitutive relationships

$$k_{rw} = s^2, \ k_{ro} = (1-s)^2, \ \Theta = \sqrt{s},$$

and use the following parameters

$$T = 50 \text{ days} \ , \ \tau = 0.5 \text{ day} \ , \ L = 2, k = 10^{-6} \text{ m}^2, \ \mu_w = 1 \text{ cP} \ , \mu_o = 10 \text{ cP}.$$

In terms of the coefficients in the model equations (7) - (10), these choices correspond to:

$$a(s) = 10^{-6}\frac{1}{s^2 + (1-s)^2}, \ f_w(s) = \frac{s^2}{s^2 + (1-s)^2}, \ \mathbf{f_1} = \mathbf{0}, \ \mathbf{f_3} = \mathbf{0}.$$

For the pressure, we use Dirichlet boundary conditions at the left ($p = 0$ at $x = 0$) and right sides ($p = 10$ at $x = 1$) and homogeneous Neumann at the rest of the boundaries. For the saturation, we use no flow boundary conditions and consider an injection at the center of the cells $f_2 = 10^{-5}$ m^3/s. For the grid of size $nx = 20, ny = 20, nz = 20$, the saturation plot at $T = 20$ days is shown in Figure 1. In Figure 2, we show the convergence of linear iteration in one time step (at $T = 20$ days) for different grid sizes. We see that the convergence is rather independent of the problem size. Moreover, as shown in Figure 3 we show that the number of linear iterations is not very sensitive (5-9 iterations) for the given problem at any time step.

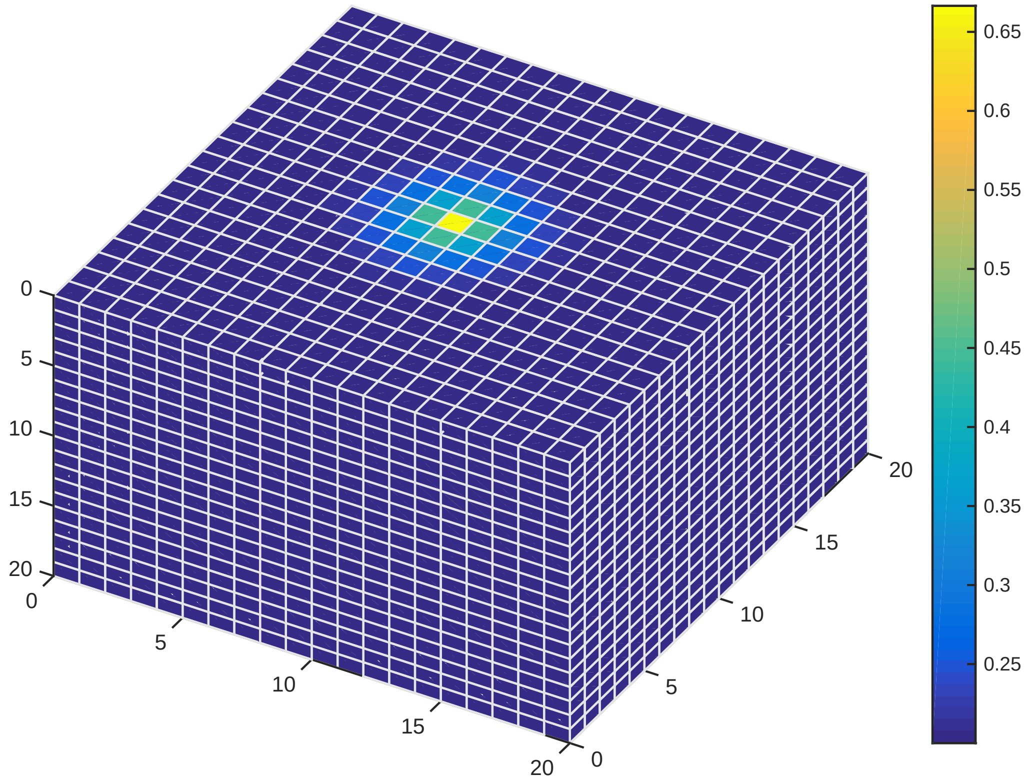

Figure 1: Plot of saturation at time $T = 20$ days. In the middle of the grid, water is injected at a constant rate.

The Hölder continuous case. We repeat now the numerical example above but for the case of a Hölder continuous saturation function $s(\cdot)$. The initial saturation is taken $s = 0.05$. We performed

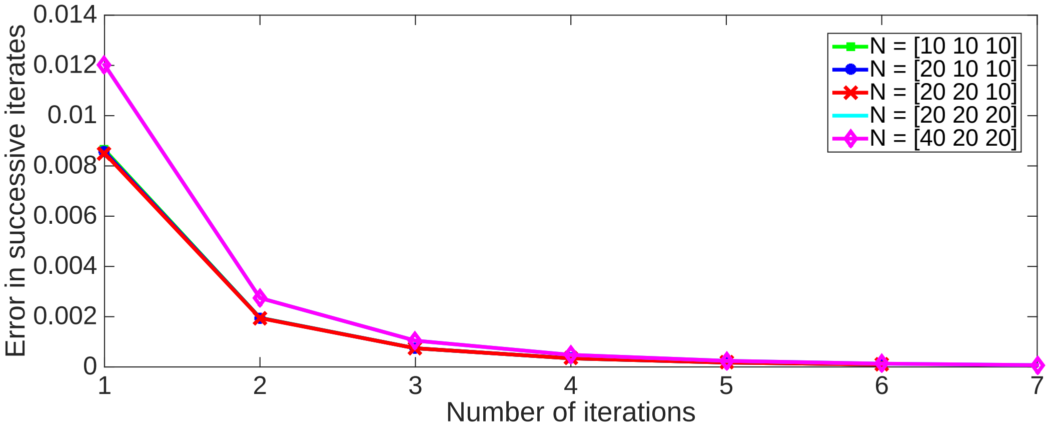

Figure 2: Convergence of the L-scheme for different grids and a Lipschitz continuous $s(\cdot)$. The legend in the figure shows the grid sizes which have been taken.

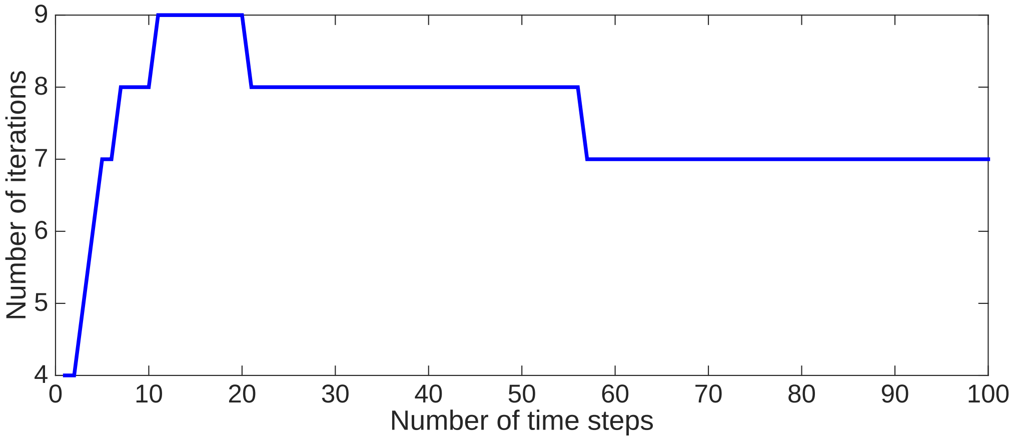

Figure 3: Number of iterations versus time step.

computations for four Hölder exponents (α) : $0.9, 0.7, 0, 5$ and 0.3, with the stabilization constant L being $1.2, 2, 8$ and 225 respectively. The time step was $\tau = 0.25$ (days) and the final time $T = 0.5$ (days), and the mesh size $nx = 20, ny = 20, nz = 20$. The convergence results are presented in Figure 4, the normalized error (i.e. the error obtained by dividing to the error after the first iteration) being plotted. As predicted by the theory, the L-scheme converges relatively fast as long as the Hölder coefficient does not become to small. The stabilization constant L increases as α decreases.

6 Conclusions

We considered a mathematical model for two-phase flow in porous media. The model was formulated in terms of a global and a complementary pressure. A fully implicit, mass conservative numerical scheme was proposed for solving it numerically. The scheme is based on backward Euler for the discretization in time and mixed finite elements (lowest order Raviart - Thomas elements) for the spatial discretization. The scheme was shown to be convergent. Moreover, order of convergence estimates are provided. For solving the non-linear systems at each time step we considered a robust, first order convergent lin-

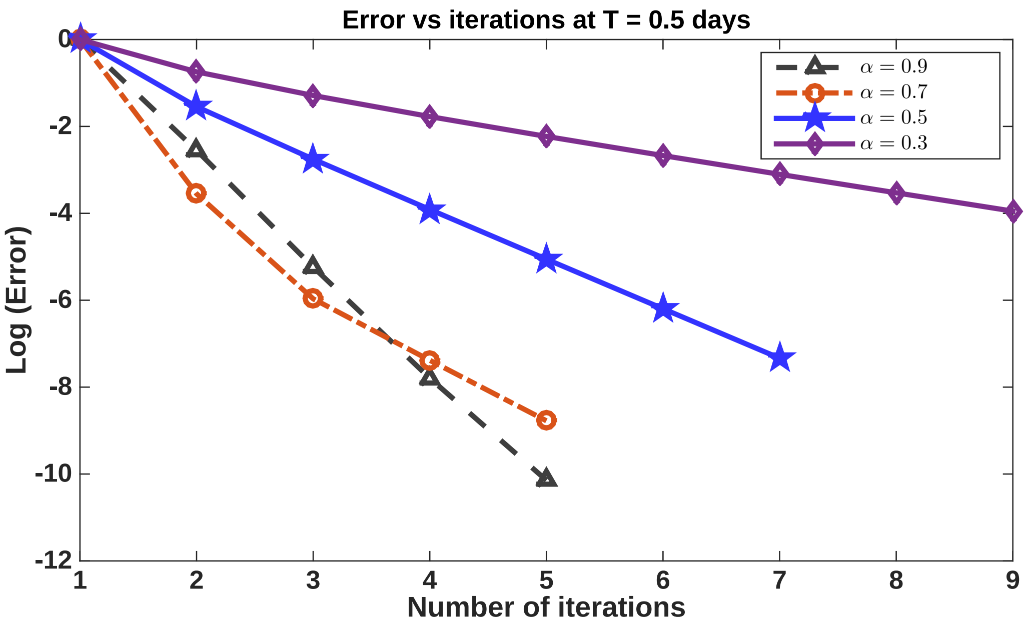

Figure 4: Convergence of the L-scheme for the Hölder continuous case.

earization method, called the L-method. The convergence of the linearization scheme was rigorously shown. We have furthermore demonstrated (rigorously for the case of Richards' equation and numerically for the general case) that the L-scheme can be also used for only Hölder continuous saturations. This method does not involve the computation of any derivatives, and does not require any regularization step. The L-method can be used also to enhance the robustness of Newton's method. The convergence rate of the method does not depend on the mesh diameter. Numerical examples have been shown to sustain the theoretical results.

Acknowledgments. The authors are members of the International Research Training Group NUPUS funded by the German Research Foundation DFG (GRK 1398), the Netherlands Organisation for Scientific Research NWO (DN 81-754) and by the Research Council of Norway (215627). F. A. Radu and I. S. Pop acknowledge the support of Statoil through the Akademia agreement. J.M. Nordbotten acknoweldges the NWO support through the Visitors Grant 040.11.351. Part of this work was done during F.A. Radu's visits in Eindhoven, support by the Meltzer foundation of the University of Bergen and by the Dutch Research Council (NWO) through the Visitors Grant 040.11.499.

References

[1] H. W. Alt and E. Di Benedetto, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scu. Norm. Sup. Pisa Cl. Sci. 12 (1985), pp. 335-392.

[2] B. Amaziane, M. Jurak and Žgaljic Keko, *An existence result for a coupled system modeling a fully equivalent global pressure formulation for immiscible compressible two-phase flow in porous media*, J. Differ. Equ. 250 (2011), pp. 1685-1718.

[3] T. ARBOGAST, *The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow*, J. non-linear Analysis: Theory, Methods & Applications 19 (1992), pp. 1009-1031.

[4] T. ARBOGAST, M. F. WHEELER AND N. Y. ZHANG, *A non-linear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal. 33 (1996), pp. 1669-1687.

[5] L. BERGAMASCHI AND M. PUTTI, *Mixed finite elements and Newton-type linearizations for the solution of Richards' equation*, Int. J. Num. Meth. Engng. 45 (1999), pp. 1025-1046.

[6] J. BEAR AND Y. BACHMAT, *Introduction to Modelling of Transport Phenomena in Porous Media*, Kluwer Academic, Dordrecht, 1991.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] C. CANCÈS, M. PIERRE, *An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field*, SIAM J. Math. Anal. 44 (2012), pp. 966-992.

[9] C. CANCÈS, I. S. POP AND M. VOHRALK, *An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*, Math. Comp. 83 (2014), pp. 153-188.

[10] M. CELIA, E. BOULOUTAS AND R. ZARBA, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res. 26 (1990), pp. 1483–1496.

[11] G. CHAVENT AND J. JAFFRE, *Mathematical models and finite elements for reservoir simulation*, Elsevier, 1991.

[12] Z. CHEN, *Degenerate two-phase incompressible flow. Existence, uniqueness and regularity of a weak solution*, J. Diff. Eqs. 171 (2001), pp. 203-232.

[13] Z. CHEN AND R. EWING, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology*, SIAM J. Numer. Anal. (1997), pp. 2228-2253.

[14] Z. CHEN AND R. EWING, *Degenerate two-phase incompressible flow III. Sharp error estimates*, Numer. Mat. 90 (2001), pp. 215-240.

[15] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[16] L.J. DURLOFSKY, *A triangle based mixed finite element-finite volume technique for modelling two-phase flow through porous media*, J. of Comput. Phys. Appl. Math. 105 (1993), pp. 252-266.

[17] Y. EPSHTEYN AND B. RIVIERE, *Analysis of hp discontinuous Galerkin methods for incompressible two-phase flow*, J. of Comput. and Appl. Math. 225 (2009), pp. 487-509.

[18] R. EYMARD, D. HILHORST AND M. VOHRALIK, *A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems*, Numer. Math. 105 (2006), pp. 73-131.

[19] R. EYMARD, R. HERBIN AND A. MICHEL, *Mathematical study of a petroleum-engineering scheme*, Math. Modell. and Numer. Anal. 37 (2003), pp. 937-962.

[20] K. B. Fadimba, *On existence and uniqueness for a coupled system modeling immiscible flow through a porous medium*, J. Math. Anal. Appl. 328 (2007), pp. 1034-1056.

[21] R. Helmig, *Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems*, Springer Verlag, 1997.

[22] J. Kou and S. Sun, *A new treatment of capillarity to improve the stability of IMPES two-phase ow formulation*, Computers and Fluids 39 (2010), pp. 1923-1931.

[23] J. Kou and S. Sun, *On iterative IMPES formulation for two phase ow with capillarity in heterogeneous porous media*, Inter. J. of Numer. Anal. and Modeling, Series B 1 (2010), pp. 20-40.

[24] D. Kröner and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differ. Equ. 55 (1984), pp. 276-288.

[25] F. Lehmann and Ph. Ackerer, *Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media*, Transport in porous media 31 (1998), pp. 275–292.

[26] F. List and F.A. Radu, *A study on iterative methods for Richards' equation*, Computational Geosciences (2015), to appear (also arXiv:1507.07837).

[27] S. Kräutle, *The semismooth Newton method for multicomponent reactive transport with minerals*, Adv. Water Resources 34 (2011), pp. 137-151.

[28] R. Klausen, F.A. Radu and G. Eigestad, *Convergence of MPFA on triangulations and for Richards' equation*, Intern. J. for Numerical Methods in Fluids (2008), pp. 1327-1351.

[29] O. A Ladyzhenskaya and N. N. Uraltseva, *Linear and quasilinear elliptic equations*, Academic Press, New York-London, 1968.

[30] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal. 41 (2003), pp. 1301-1317.

[31] J. M. Nordbotten and M. A. Celia, *Geological Storage of CO2. Modeling Approaches for Large-Scale Simulation*, John Wiley & Sons, 2012.

[32] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal. 25 (1988), pp. 784–814.

[33] R. Neumann, P. Bastian and O. Ippisch, *Modeling and simulation of two-phase two-component flow with disappearing nonwetting phase*, Computational geosciences 17 (2013), pp. 139-149.

[34] M. Ohlberger, *Convergence of a mixed finite element- finite volume method for two phase flow in porous media*, East-West J- Numer. Math. 5 (1997), pp. 183-210.

[35] E. J. Park, *Mixed finite elements for non-linear second-order elliptic problems*, SIAM J. Numer. Anal. 32 (1995), pp. 865-885.

[36] I. S. Pop, *Error estimates for a time discretization method for the Richards' equation*, Computational geosciences 6 (2002), pp. 141-160.

[37] I. S. POP, F.A. RADU AND P. KNABNER, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. and Appl. Math. 168 (2004), pp. 365-373.

[38] F. A. RADU, I. S. POP AND P. KNABNER, *Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation*, SIAM J. Numer. Anal. 42 (2004), pp. 1452-1478.

[39] F.A. RADU, I.S. POP AND P. KNABNER, *On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation*, In Numerical Mathematics and Advanced Applications. A. Bermudez de Castro et al. (editors), Springer, 1194-1200, 2006.

[40] F. A. RADU, I. S. POP AND P. KNABNER, *Error estimates for a mixed finite element discretization of some degenerate parabolic equations*, Numer. Math. 109 (2008), pp. 285-311.

[41] F.A. RADU AND I. S. POP, *Mixed finite element discretization and Newton iteration for a reactive contaminant transport model with nonequilibrium sorption: convergence analysis and error estimates*, Computational Geosciences 15 (2011), pp. 431-450.

[42] F. A. RADU, J. M. NORDBOTTEN, I. S. POP AND K. KUMAR, *A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media*, J. Comput. and Appl. Math. 289 (2015), pp. 134-141.

[43] B. RIVIERE AND N. WALKINGTON, *Convergence of a discontinuous Galerkin method for the miscible displacement equation under low regularity*, SIAM J. Numer. Anal. 49 (2011), pp. 1085-1110.

[44] A. QUARTERONI AND A. VALLI, *Numerical approximations of partial differential equations*, Springer-Verlag, 1994.

[45] M. SLODICKA, *A robust and efficient linearization scheme for doubly non-linear and degenerate parabolic problems arising in flow in porous media*, SIAM J. Sci. Comput. 23 (2002), pp.1593-1614.

[46] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, Vol. 343. American Mathematical Soc., 2001.

[47] J.M. THOMAS, *Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes*, These d'Etat, University Pierre et Marie Curie (Paris 6), 1977.

[48] C. WOODWARD AND C. DAWSON, *Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media*, SIAM J. Numer. Anal. 37 (2000), pp. 701-724.

[49] W.A. YONG AND I.S. POP, *A numerical approach to porous medium equations*, Preprint 95-50, SFB 359, IWR, University of Heidelberg, 1996.

[50] I. YOTOV, *A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arizing in porous media flow*, EastWest J. Numer. Math. 5 (1997), pp. 211-230.

PREVIOUS PUBLICATIONS IN THIS SERIES:

Number	Author(s)	Title	Month
15-35	V.A. Khoa A. Muntean	Asymptotic analysis of a semi-linear elliptic system in perforated domains: well-posedness and correctors for the homogenization limit	Dec. '15
15-36	K. Urbanowicz A.S. Tijsseling	Work and life of Piotr Szymański	Dec. '15
15-37	N. Kumar J.H.M. ten Thije Boonkkamp B. Koren	Flux approximation scheme for the incompressible Navier-Stokes equations using local boundary value problems	Dec. '15
15-38	C. Filosa J.H.M. ten Thije Boonkkamp W.J. IJzerman	A new ray tracing method in phase space using α-shapes	Dec. '15
15-39	F.A. Radu K. Kumar J.M. Nordbotten I.S. Pop	A convergent mass conservative numerical scheme based on mixed finite elements for two-phase flow in porous media	Dec. '15

Ontwerp: de Tantes,
Tobias Baanders, CWI