ESAIM: M2AN 55 (2021) 2101–2139
https://doi.org/10.1051/m2an/2021047

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

UPSTREAM MOBILITY FINITE VOLUMES FOR THE RICHARDS EQUATION IN HETEROGENOUS DOMAINS

Sabrina Bassetto[1], Clément Cancès[2], Guillaume Enchéry[1,*] and Quang-Huy Tran[1]

Abstract. This paper is concerned with the Richards equation in a heterogeneous domain, each subdomain of which is homogeneous and represents a rocktype. Our first contribution is to rigorously prove convergence toward a weak solution of cell-centered finite-volume schemes with upstream mobility and without Kirchhoff's transform. Our second contribution is to numerically demonstrate the relevance of locally refining the grid at the interface between subregions, where discontinuities occur, in order to preserve an acceptable accuracy for the results computed with the schemes under consideration.

Mathematics Subject Classification. 65M08, 65M12, 76S05.

Received January 15, 2021. Accepted August 18, 2021.

1. Presentation of the continuous model

The Richards equation [46] is one of the most well-known simplified models for water filtration in unsaturated soils. While it has been extensively studied in the case of a homogeneous domain, the heterogeneous case seems to have received less attention in the literature, at least from the numerical perspective. The purpose of this paper is to investigate a class of discretization scheme for a special instance of heterogeneous domains, namely, those with piecewise-uniform physical properties.

Before stating our objectives in a precise manner, a few prerequisites must be introduced regarding the model in Sections 1.1, 1.2 and the scheme in Sections 2.1, 2.2. The goal of the paper is fully described in Section 1.3, in relation with other works. Practical aspects related to the numerical resolution are detailed in Section 5 and results on illustrative test cases are shown in Section 6. A summary of our main results is provided in Section 2.3, together with the outline of the paper.

1.1. Richards' equation in heterogeneous porous media

Let $\Omega \subset \mathbb{R}^d$, where $d \in \{2, 3\}$, be a connected open polyhedral domain with Lipschitz boundary $\partial\Omega$. A porous medium defined over the region Ω is characterized by

– the porosity $\phi : \Omega \to (0, 1]$;

Keywords and phrases. Richards' equation, heterogeneous domains, finite-volume schemes, mobility upwinding.

[1] IFP Energies nouvelles, 1 et 4 avenue de Bois Préau, 92852 Rueil-Malmaison Cedex, France.
[2] Inria, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, 59000 Lille, France.
*Corresponding author: guillaume.enchery@ifpen.fr

– the permeability $\lambda : \Omega \to \mathbb{R}_+^*$;
– the mobility function $\eta : [0,1] \times \Omega \to \mathbb{R}_+$;
– the saturation law $\mathcal{S} : \mathbb{R} \times \Omega \to [0,1]$ function of the water pressure and the space location.

The conditions to be satisfied by ϕ, λ, η and $\mathcal{S}$ will be elaborated on later. In a homogeneous medium, these physical properties are uniform over Ω, *i.e.*,

$$\phi(x) = \phi_0, \qquad \lambda(x) = \lambda_0, \qquad \eta(s,x) = \eta_0(s), \qquad \mathcal{S}(p,x) = \mathcal{S}_0(p)$$

for all $x \in \Omega$. In a heterogeneous medium, the dependence of ϕ, λ, η and $\mathcal{S}$ on x must naturally be taken into account. The quantity s, called saturation, measures the relative volumic presence of water in the medium. The quantity p is the water pressure, which in our case is the opposite of the capillary pressure.

Let $T > 0$ be a finite time horizon. We designate by $Q_T = (0,T) \times \Omega$ the space-time domain of interest. Our task is to find the saturation field $s : Q_T \to [0,1]$ and the pressure field $p : Q_T \to \mathbb{R}$ so as to satisfy

– the interior equations

$$\phi(x)\,\partial_t s + \operatorname{div} F = 0 \qquad \text{in } Q_T, \tag{1.1a}$$
$$F + \lambda(x)\,\eta(s,x)\,\nabla(p - \varrho g \cdot x) = 0 \qquad \text{in } Q_T, \tag{1.1b}$$
$$s - \mathcal{S}(p,x) = 0 \qquad \text{in } Q_T; \tag{1.1c}$$

– the boundary conditions

$$F \cdot n(x) = 0 \qquad \text{on } (0,T) \times \Gamma^{\mathrm{N}}, \tag{1.1d}$$
$$p(t,x) = p^{\mathrm{D}}(x) \qquad \text{on } (0,T) \times \Gamma^{\mathrm{D}}; \tag{1.1e}$$

– the initial data

$$s(0,x) = s^0(x) \qquad \text{in } \Omega. \tag{1.1f}$$

The partial differential equation (1.1a) expresses the water volume balance. The flux F involved in this balance is given by the Darcy–Muskat law (1.1b), in which g is the gravity vector and ϱ is the known constant density of water, assumed to be incompressible. It is convenient to introduce

$$\psi = -\varrho g \cdot x, \qquad \vartheta = p + \psi, \tag{1.2}$$

referred to respectively as gravity potential and hydraulic head. In this way, the Darcy–Muskat law (1.1b) can be rewritten as

$$F + \lambda(x)\,\eta(s,x)\,\nabla(p + \psi) = F + \lambda(x)\,\eta(s,x)\,\nabla\vartheta = 0.$$

Equation (1.1c) connecting the saturation s and the pressure p is the capillary pressure relation. The boundary $\partial\Omega$ is split into two non-overlapping parts, *viz.*,

$$\partial\Omega = \Gamma^{\mathrm{N}} \cup \Gamma^{\mathrm{D}}, \qquad \Gamma^{\mathrm{N}} \cap \Gamma^{\mathrm{D}} = \varnothing, \tag{1.3}$$

where Γ^{N} is open and Γ^{D} is closed, the latter having a positive $(d-1)$-dimensional Hausdorff measure $\nu^{d-1}(\Gamma^{\mathrm{D}}) > 0$. The no-flux Neumann condition (1.1d) is prescribed on $(0,T) \times \Gamma^{\mathrm{N}}$, where $n(x)$ is the outward normal unit vector at $x \in \Gamma^{\mathrm{N}}$. The Dirichlet condition (1.1e) with a known Lipschitz function $p^{\mathrm{D}} \in W^{1,\infty}(\Omega)$ is imposed on $(0,T) \times \Gamma^{\mathrm{D}}$. Note that, in our theoretical development, the function p^{D} is assumed to be defined over the whole domain Ω, which is stronger than a data $p^{\mathrm{D}} \in L^\infty(\Gamma^{\mathrm{D}})$ given only on the boundary. The assumption that p^{D} does not depend on time can be removed by following the lines of [16], but we prefer here not to deal with time-dependent boundary data in order to keep the presentation as simple as possible. Finally, the initial data $s^0 \in L^\infty(\Omega;[0,1])$ in (1.1f) is also a given data.

In this work, we restrict ourselves to a specific type of heterogeneous media, defined as follows. We assume that the domain Ω can be partitioned into several connected polyhedral subdomains Ω_i, $1 \leqslant i \leqslant I$. Technically, this means that if $\Gamma_{i,j}$ denotes the interface between Ω_i and Ω_j (which can be empty for some particular choices of $\{i,j\}$), then

$$\Omega_i \cap \Omega_j = \varnothing, \quad \overline{\Omega}_i \cap \overline{\Omega}_j = \Gamma_{i,j}, \text{ if } i \neq j, \quad \Omega = \left(\bigcup_{1 \leqslant i \leqslant I} \Omega_i \right) \cup \Gamma, \tag{1.4}$$

with $\Gamma = \bigcup_{i \neq j} \Gamma_{i,j}$. Each of these subdomains corresponds to a distinctive rocktype. Inside each Ω_i, the physical properties are homogeneous. In other words,

$$\phi(x) = \phi_i, \qquad \lambda(x) = \lambda_i, \qquad \eta(s,x) = \eta_i(s), \qquad \mathcal{S}(p,x) = \mathcal{S}_i(p)$$

for all $x \in \Omega_i$. Therefore, system (1.1) is associated with

$$\phi(x) = \sum_{1 \leqslant i \leqslant I} \phi_i \, \mathbf{1}_{\Omega_i}(x), \qquad \eta(s,x) = \sum_{1 \leqslant i \leqslant I} \eta_i(s) \, \mathbf{1}_{\Omega_i}(x), \tag{1.5a}$$

$$\lambda(x) = \sum_{1 \leqslant i \leqslant I} \lambda_i \, \mathbf{1}_{\Omega_i}(x), \qquad \mathcal{S}(p,x) = \sum_{1 \leqslant i \leqslant I} \mathcal{S}_i(p) \, \mathbf{1}_{\Omega_i}(x), \tag{1.5b}$$

where $\mathbf{1}_{\Omega_i}$ stands for the characteristic function of Ω_i. For all $i \in \{1, \dots, I\}$, we assume that $\phi_i \in (0,1]$ and $\lambda_i > 0$. Furthermore, we require that

$$\eta_i \text{ is increasing on } [0,1], \qquad \eta_i(0) = 0, \qquad \eta_i(1) = \frac{1}{\mu}, \tag{1.6a}$$

where $\mu > 0$ is the (known) viscosity of water. In addition to the assumption that $\mathcal{S}(\cdot, x)$, defined in (1.5b), is absolutely continuous and nondecreasing, the functions $\mathcal{S}_i$ are also subject to some generic requirements commonly verified the models available in the literature: for each $i \in \{1, \dots, I\}$, there exists $\overline{p}_i \leqslant 0$ such that

$$\mathcal{S}_i \text{ is increasing on } (-\infty, \overline{p}_i], \qquad \lim_{p \to -\infty} \mathcal{S}_i(p) = 0, \qquad \mathcal{S}_i \equiv 1 \text{ on } [\overline{p}_i, +\infty). \tag{1.6b}$$

This allows us to define an inverse $\mathcal{S}_i^{-1} : (0,1] \to (-\infty, \overline{p}_i]$ such that $\mathcal{S}_i \circ \mathcal{S}_i^{-1}(s) = s$ for all $s \in (0,1]$. We further assume that for all $i \in \{1, \dots, I\}$ the function $\mathcal{S}_i$ is bounded in $L^1(\mathbb{R}_-)$, or equivalently, that $\mathcal{S}_i^{-1} \in L^1(0,1)$.

It thus makes sense to consider the capillary energy density functions $\mathfrak{E}_i : \mathbb{R} \times \Omega_i \to \mathbb{R}_+$ defined by

$$\mathfrak{E}_i(s,x) = \int_{\mathcal{S}_i(p^{\mathrm{D}}(x))}^{s} \phi_i \left(\mathcal{S}_i^{-1}(\varsigma) - p^{\mathrm{D}}(x) \right) \mathrm{d}\varsigma. \tag{1.7}$$

For all $x \in \Omega_i$, the function $\mathfrak{E}_i(\cdot, x)$ is nonnegative, convex since $\mathcal{S}_i^{-1}$ is monotone, and bounded on $[0,1]$ as a consequence of the integrability of $\mathcal{S}_i$. For technical reasons that will appear clearly later on, we further assume that

$$\sqrt{\eta_i \circ \mathcal{S}_i} \in L^1(\mathbb{R}_-), \qquad \forall i \in \{1, \dots, I\}. \tag{1.8}$$

Let $Q_{i,T} = (0,T) \times \Omega_i$ be the space-time subdomains for $1 \leqslant i \leqslant I$. The interior equations (1.1a)–(1.1c) then boil down to

$$\phi_i \, \partial_t s + \operatorname{div} F = 0 \qquad \text{in } Q_{i,T}, \tag{1.9a}$$

$$F + \lambda_i \, \eta_i \, \nabla(p + \psi) = 0 \qquad \text{in } Q_{i,T}, \tag{1.9b}$$

$$s - \mathcal{S}_i(p) = 0 \qquad \text{in } Q_{i,T}. \tag{1.9c}$$

At the interface $\Gamma_{i,j}$ between Ω_i and Ω_j, $i \neq j$, any solution of (1.1a)–(1.1c) satisfies the matching conditions

$$F_i \cdot n_i + F_j \cdot n_j = 0 \qquad \text{on } (0,T) \times \Gamma_{i,j}, \tag{1.10a}$$

$$p_i - p_j = 0 \qquad \text{on } (0,T)\times\Gamma_{i,j}. \tag{1.10b}$$

In the continuity of the normal fluxes (1.10a), which is enforced by the conservation of water volume, n_i denotes the outward normal to $\partial\Omega_i$ and $F_i\cdot n_i$ stands for the trace of the normal component of $F_{|_{Q_{i,T}}}$ on $(0,T)\times\partial\Omega_i$. In the continuity of pressure (1.10b), which also results from (1.1a)–(1.1c), p_i denotes the trace on $(0,T)\times\partial\Omega_i$ of the pressure $p_{|_{Q_{i,T}}}$ in the i-th domain.

1.2. Stability features and notion of weak solutions

We wish to give a proper sense to the notion of weak solution for problem (1.1). To achieve this purpose, we need a few mathematical transformations the definition of which crucially relies on a fundamental energy estimate at the continuous level. The calculations below are aimed at highlighting this energy estimate and will be carried out in a formal way, in constrast to those in the fully discrete setting.

Multiplying (1.9a) by $p-p^{\mathrm{D}}$, invoking (1.7), integrating over Ω_i and summing over i, we end up with

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{I}\int_{\Omega_i} \mathfrak{E}_i(s,x)\,\mathrm{d}x + \sum_{i=1}^{I}\int_{\Omega_i}\operatorname{div} F\left(p-p^{\mathrm{D}}\right)\,\mathrm{d}x = 0. \tag{1.11}$$

We now integrate by parts the second term. Thanks to the matching conditions (1.10) and the regularity of p^{D}, we obtain

$$\mathtt{A} := \sum_{i=1}^{I}\int_{\Omega_i}\operatorname{div} F\left(p-p^{\mathrm{D}}\right)\,\mathrm{d}x = -\sum_{i=1}^{I}\int_{\Omega_i} F\cdot\nabla\left(p-p^{\mathrm{D}}\right)\,\mathrm{d}x.$$

It follows from the flux value (1.9b) that

$$\begin{aligned}\mathtt{A} &= \sum_{i=1}^{I}\int_{\Omega_i}\lambda_i\eta_i(s)\nabla(p+\psi)\cdot\nabla\left(p-p^{\mathrm{D}}\right)\,\mathrm{d}x\\ &= \sum_{i=1}^{I}\int_{\Omega_i}\lambda_i\eta_i(s)|\nabla p|^2\,\mathrm{d}x - \sum_{i=1}^{I}\int_{\Omega_i}\lambda_i\eta_i(s)\nabla\psi\cdot\nabla p^{\mathrm{D}}\,\mathrm{d}x\\ &\quad + \sum_{i=1}^{I}\int_{\Omega_i}\lambda_i\eta_i(s)\nabla p\cdot\nabla\left(\psi-p^{\mathrm{D}}\right)\,\mathrm{d}x.\end{aligned}$$

Young's inequality, combined with the boundedness of ∇p^{D}, $\nabla\psi$, λ and η, yields

$$\mathtt{A} \geqslant \frac{1}{2}\sum_{i=1}^{I}\int_{\Omega_i}\lambda_i\eta_i(s)|\nabla p|^2\,\mathrm{d}x - C$$

for some $C\geqslant 0$ depending only on λ, η, ψ, μ, Ω and p^{D}.

Let us define the energy $\mathfrak{E}:[0,T]\to\mathbb{R}_+$ by

$$\mathfrak{E}(t) = \sum_{i=1}^{I}\int_{\Omega_i}\mathfrak{E}_i(s(t,x),x)\,\mathrm{d}x, \qquad 0\leqslant t\leqslant T.$$

Integrating (1.11) w.r.t. time results in

$$\mathfrak{E}(T) + \frac{1}{2}\sum_{i=1}^{I}\iint_{Q_{i,T}}\lambda_i\eta_i(s)|\nabla p|^2\,\mathrm{d}x\,\mathrm{d}t \leqslant \mathfrak{E}(0) + CT. \tag{1.12}$$

Estimate (1.12) is the core of our analysis. However, it is difficult to use in its present form since $\eta_i(s) = \eta_i(\mathcal{S}_i(p))$ vanishes as p tends to $-\infty$, so that the control of ∇p degenerates. To circumvent this difficulty, we resort to the nonlinear functions (customarily referred to as the Kirchhoff transforms) $\Theta_i : \mathbb{R} \to \mathbb{R}$, $\Phi_i : \mathbb{R} \to \mathbb{R}$, and $\Upsilon : \mathbb{R} \times \Omega \to \mathbb{R}$ respectively defined by

$$\Theta_i(p) = \int_0^p \sqrt{\lambda_i \eta_i \circ \mathcal{S}_i(\pi)}\, \mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{1.13a}$$

$$\Phi_i(p) = \int_0^p \lambda_i \eta_i \circ \mathcal{S}_i(\pi)\, \mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{1.13b}$$

$$\Upsilon(p) = \int_0^p \min_{1 \leqslant i \leqslant I} \sqrt{\lambda_i \eta_i \circ \mathcal{S}_i(\pi)}\, \mathrm{d}\pi, \qquad p \in \mathbb{R}, \tag{1.13c}$$

the notion of Υ being due to [25]. Bearing in mind that $\mathfrak{E}(T) \geqslant 0$, estimate (1.12) implies that

$$\sum_{i=1}^{I} \iint_{Q_{i,T}} |\nabla \Theta_i(p)|^2 \, \mathrm{d}x \, \mathrm{d}t \leqslant 2(\mathfrak{E}(0) + CT) < +\infty. \tag{1.14}$$

As $\Phi_i \circ \Theta_i^{-1}$ is Lipschitz continuous, this also gives rise to a $L^2(Q_{i,T})$-estimate on $\nabla \Phi_i(p)$. The functions $\sum_i \Theta_i(p) \mathbf{1}_{\Omega_i}$ and $\sum_i \Phi_i(p) \mathbf{1}_{\Omega_i}$ are in general discontinuous across the interfaces $\Gamma_{i,j}$, unlike $\Upsilon(p)$. Since the functions $\Upsilon \circ \Theta_i^{-1}$ are Lipschitz continuous, we can readily infer from (1.14) that

$$\iint_{Q_T} |\nabla \Upsilon(p)|^2 \, \mathrm{d}x \leqslant C \tag{1.15}$$

for some C depending on T, Ω, $\|\nabla p^{\mathrm{D}}\|_\infty$, the $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$'s and

$$\overline{\lambda} = \|\lambda\|_{L^\infty(\Omega)} = \max_{1 \leqslant i \leqslant I} \lambda_i, \qquad \overline{\eta} = \|\eta\|_{L^\infty(\Omega)} = \max_{1 \leqslant i \leqslant I} \|\eta_i\|_{L^\infty(\Omega)} = \frac{1}{\mu},$$

the last equality being due to (1.6a).

Moreover, $\Upsilon(p) - \Upsilon(p^{\mathrm{D}})$ vanishes on $(0,T) \times \Gamma^{\mathrm{D}}$. Poincaré's inequality provides a $L^2(Q_T)$-estimate on $\Upsilon(p)$ since Γ^{D} has positive measure and since $\Upsilon(p^{\mathrm{D}})$ is bounded in Ω. In view of assumption (1.8), the functions Θ_i and Υ are bounded on $\mathbb{R}_-$. Besides, for $p \geqslant 0$, $\eta_i \circ \mathcal{S}_i(p) = 1/\mu$, so that $\Theta_i(p) = p\sqrt{\lambda_i/\mu}$ and $\Upsilon(p) = \min_{1 \leqslant i \leqslant I} p\sqrt{\lambda_i/\mu}$. It finally comes that

$$\Theta_i(p) \leqslant C(1 + \Upsilon(p)), \qquad \forall p \in \mathbb{R},\ 1 \leqslant i \leqslant I, \tag{1.16}$$

from which we infer a $L^2(Q_{i,T})$-estimate on $\Theta_i(p)$. Putting

$$V = \left\{ u \in H^1(\Omega) \mid u_{|_{\Gamma^{\mathrm{D}}}} = 0 \right\},$$

the above estimates suggest the following notion of weak solution for our problem.

Definition 1.1. A measurable function $p : Q_T \to \mathbb{R}$ is said to be a weak solution to the problem (1.9a)–(1.9c) if

$$\Theta_i(p) \in L^2((0,T); H^1(\Omega_i)), \qquad \text{for } 1 \leqslant i \leqslant I, \tag{1.17a}$$

$$\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T); V) \tag{1.17b}$$

and if for all $\varphi \in C_c^\infty([0,T) \times (\Omega \cup \Gamma^{\mathrm{N}}))$, there holds

$$\iint_{Q_T} \phi\, \mathcal{S}(p,x) \partial_t \varphi \, \mathrm{d}x \, \mathrm{d}t + \int_\Omega \phi\, s^0 \varphi(\cdot, 0) \, \mathrm{d}x + \iint_{Q_T} F \cdot \nabla \varphi \, \mathrm{d}x \, \mathrm{d}t = 0, \tag{1.17c}$$

with

$$F = -\nabla \Phi_i(p) + \lambda_i \eta_i(\mathcal{S}_i(p))\, \varrho g \qquad \text{in } Q_{i,T},\ 1 \leqslant i \leqslant I. \tag{1.17d}$$

The expression (1.17d) is a reformulation of the original one (1.9b) in a quasilinear form which is suitable for analysis, even though the physical meaning of the Kirchhoff transform $\Phi_i(p)$ is unclear. While the formulation (1.17c) should be thought of as a weak form of (1.9a), (1.10a), (1.1f), and (1.1d), the condition $\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T);V)$ contains (1.10b) and (1.1e).

1.3. Goal and positioning of the paper

We are now in a position to clearly state the two objectives of this paper.

The first objective is to put forward a rigorous proof that, for problem (1.1) with heterogeneous data (1.5), cell-centered finite-volume schemes with upstream mobility such as described in Section 2.2, do converge towards a weak solution (in the sense of Def. 1.1) as the discretization parameters tend to 0. Such mathematically assessed convergence results are often dedicated to homogeneous cases: see for instance [4, 28, 45] for schemes involving the Kirchhoff transforms for Richards' equation, Ait Hammou Oulhaj *et al.* [1] for a upstream mobility CVFE approximation of Richards' equation in anisotropic domains, Chavent and Jaffré [19] and Chen and Ewing [20,21] for schemes for two-phase flows involving the Kirchhoff transform, and [31,36] for upstream mobility schemes for two-phase porous media flows. For flows in highly heterogeneous porous media, rigorous mathematical results have been obtained for schemes involving the introduction of additional interface unknowns and Kirchhoff's transforms (see for instance [8,13,14,25]), or under the non-physical assumption that the mobilities are strictly positive [30,33]. We also refer the reader to [3,44] where the assumption of the non-degeneracy of the mobility has been made. It was established very recently in [10] that cell-centered finite-volumes with (hybrid) upwinding also converge for two-phase flows in heterogeneous domains, but with a specific treatment of the interfaces located at the heterogeneities. Here, the novelty lies in the fact that we do not consider any specific treatment of the interface in the design of the scheme.

The second objective is of more practical nature. Even though our analysis still holds without any specific treatment of the interface, it is well-known that cell-centered upstream mobility finite-volumes can be inaccurate in the presence of heterogeneities. This observation motivated several contributions (see for instance [25,26,33,37]) where skeletal (*i.e.*, edge or vertex) unknowns where introduced in order to enforce the continuity of the pressures at the interfaces $\Gamma_{i,j}$. By means of extensive numerical simulations in Section 6, we will show that without local refinement of the grid at the interface, the method still converges, but with a degraded order. Our ultimate motivation is to propose an approach which consists in adding very thin cells on both sides of the interface before using the cell centered scheme under study. Then the scheme appears to behave better, with first-order accuracy. Moreover, one can still make use of the parametrized cut-Newton method proposed in [5] to compute the solution to the nonlinear system corresponding to the scheme. This method appears to be very efficient, while it avoids the possibly difficult construction of compatible parametrizations at the interfaces as in [9–11]. An involved comparative study on the robustness of the Newton solver is presented in [6], where other strategies to capture the discontinuities related to rock changes are also addressed.

2. Finite-volume discretization

The scheme we consider in this paper is based on two-point flux approximation (TPFA) finite-volumes. Hence, it is subject to some restrictions on the mesh [32,35]. We first review the requirements on the mesh in Section 2.1. Next, we construct the upstream mobility finite-volume scheme for Richards' equation in Section 2.2. The main mathematical results of the paper, which are the well-posedness of the nonlinear system corresponding to the scheme and the convergence of the scheme, are then summarized in Section 2.3.

2.1. Admissible discretization of Q_T

Let us start by discretizing w.r.t. space.

Definition 2.1. An *admissible mesh of* Ω is a triplet $(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}})$ such that the following conditions are fulfilled:

(i) Each control volume (or cell) $K \in \mathscr{T}$ is non-empty, open, polyhedral and convex, with positive d-dimensional Lebesgue measure $m_K > 0$. We assume that

$$K \cap L = \varnothing \quad \text{if } K, L \in \mathscr{T} \text{ with } K \neq L, \qquad \text{while} \quad \bigcup_{K \in \mathscr{T}} \overline{K} = \overline{\Omega}.$$

Moreover, we assume that the mesh is adapted to the heterogeneities of Ω, in the sense that for all $K \in \mathscr{T}$, there exists $i \in \{1, \dots, I\}$ such that $K \subset \Omega_i$.

(ii) Each face $\sigma \in \mathscr{E}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff measure $\nu^{d-1}(\sigma) = m_\sigma > 0$. We assume that $\nu^{d-1}(\sigma \cap \sigma') = 0$ for $\sigma, \sigma' \in \mathscr{E}$ unless $\sigma' = \sigma$. For all $K \in \mathscr{T}$, we assume that there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \bigcup_{\sigma \in \mathscr{E}_K} \sigma$. Moreover, we suppose that $\bigcup_{K \in \mathscr{T}} \mathscr{E}_K = \mathscr{E}$. Given two distinct control volumes $K, L \in \mathscr{T}$, the intersection $\overline{K} \cap \overline{L}$ either reduces to a single face $\sigma \in \mathscr{E}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

(iii) The cell-centers $(x_K)_{K \in \mathscr{T}}$ are pairwise distinct with $x_K \in K$, and are such that, if $K, L \in \mathscr{T}$ share a face $K|L$, then the vector $x_L - x_K$ is orthogonal to $K|L$.

(iv) For the boundary faces $\sigma \subset \partial\Omega$, we assume that either $\sigma \subset \Gamma^{\mathrm{D}}$ or $\sigma \subset \overline{\Gamma}^{\mathrm{N}}$. For $\sigma \subset \partial\Omega$ with $\sigma \in \mathscr{E}_K$ for some $K \in \mathscr{T}$, we assume additionally that there exists $x_\sigma \in \sigma$ such that $x_\sigma - x_K$ is orthogonal to σ.

In our problem, the standard Definition 2.1 must be supplemented by a compatibility property between the mesh and the subdomains. By "compatbility" we mean that each cell must lie entirely inside a single subregion. Put another way,

$$\forall K \in \mathscr{T}, \quad \exists!\, i(K) \in \{1, \dots, I\} \;\mid\; K \subset \Omega_{i(K)}. \tag{2.1}$$

This has two consequences. The first one is that, if we define

$$\mathscr{T}_i = \{K \in \mathscr{T} \mid K \subset \Omega_i\}, \qquad 1 \leqslant i \leqslant I, \tag{2.2}$$

then $\mathscr{T} = \bigcup_{i=1}^{I} \mathscr{T}_i$. The second one is that the subdomain interfaces $\Gamma_{i,j}$ for $i \neq j$ coincide necessarily with some edges $\sigma \in \mathscr{E}$. To express this more accurately, let $\mathscr{E}_\Gamma = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma\}$ be the set of the interface edges, $\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma^{\mathrm{D}}\}$ be the set of Dirichlet boundary edges, and $\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}} = \{\sigma \in \mathscr{E} \mid \sigma \subset \overline{\Gamma}^{\mathrm{N}}\}$ be the set of Neumann boundary edges. Then, $\Gamma = \bigcup_{\sigma \in \mathscr{E}_\Gamma} \sigma$, while $\Gamma^{\mathrm{D}} = \bigcup_{\sigma \in \mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}} \sigma$ and $\overline{\Gamma}^{\mathrm{N}} = \bigcup_{\sigma \in \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}} \sigma$. For later use, it is also convenient to introduce the subset $\mathscr{E}_i \subset \mathscr{E}$ consisting of those edges that correspond to cells in $\mathscr{T}_i$ only, *i.e.*,

$$\mathscr{E}_i = \left(\bigcup_{K \in \mathscr{T}_i} \mathscr{E}_K \right) \backslash \mathscr{E}_\Gamma, \qquad 1 \leqslant i \leqslant I, \tag{2.3a}$$

and the subset $\mathscr{E}_{\mathrm{int}}$ of the internal edges, *i.e.*,

$$\mathscr{E}_{\mathrm{int}} = \mathscr{E} \backslash (\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}) = \bigcup_{K, L \in \mathscr{T}} \{\sigma = K|L\}. \tag{2.3b}$$

Note that $\mathscr{E}_\Gamma \subset \mathscr{E}_{\mathrm{int}}$.

To each edge $\sigma \in \mathscr{E}$, we associate a distance d_σ by setting

$$d_\sigma = \begin{cases} |x_K - x_L| & \text{if } \sigma = K|L \in \mathscr{E}_{\mathrm{int}}, \\ |x_K - x_\sigma| & \text{if } \sigma \in \mathscr{E}_K \cap \left(\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}} \cup \mathscr{E}_{\mathrm{ext}}^{\mathrm{N}}\right). \end{cases} \tag{2.4}$$

We also define $d_{K\sigma} = \mathrm{dist}(x_K, \sigma)$ for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$. The transmissivity of the edge $\sigma \in \mathscr{E}$ is defined by

$$a_\sigma = \frac{m_\sigma}{d_\sigma}. \tag{2.5}$$

Throughout the paper, many discrete quantities $\boldsymbol{u}$ will be defined either in cells $K \in \mathscr{T}$ or on Dirichlet boundary edges $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$, *i.e.* $\boldsymbol{u} = ((u_K)_{K\in\mathscr{T}}, (u_\sigma)_{\sigma\in\mathscr{E}_{\rm ext}^{\rm D}}) \in \mathbb{X}^{\mathscr{T}\cup\mathscr{E}_{\rm ext}^{\rm D}}$, where $\mathbb{X}$ can be either $\mathbb{R}^\ell$, $\ell \geqslant 1$, or a space of functions. Then for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$, we define the mirror value $u_{K\sigma}$ by

$$u_{K\sigma} = \begin{cases} u_L & \text{if } \sigma = K|L \in \mathscr{E}_{\rm int}, \\ u_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm N}, \\ u_\sigma & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm D}. \end{cases} \tag{2.6}$$

The diamond cell Δ_σ corresponding to the edge σ is defined as the convex hull of $\{x_K, x_{K\sigma}, \sigma\}$ for K such that $\sigma \in \mathscr{E}_K$, while the half-diamond cell $\Delta_{K\sigma}$ is defined as the convex hull of $\{x_K, \sigma\}$. Denoting by m_{Δ_σ} the Lebesgue measure of Δ_σ, the elementary geometrical relation $m_{\Delta_\sigma} = d\, m_\sigma d_\sigma$ where d stands for the dimension will be used many times in what follows.

Another notational shorthand is worth introducing now, since it will come in handy in the sequel. Let

$$f(\cdot, x) = \sum_{1\leqslant i\leqslant I} f_i(\cdot)\mathbf{1}_{\Omega_i}(x) \tag{2.7a}$$

be a scalar quantity or a function whose dependence of $x \in \Omega$ is of the type (1.5). Then, for $K \in \mathscr{T}$, we slightly abuse the notations in writing

$$f_K(\cdot) := f(\cdot, x_K) = f_{i(K)}(\cdot), \tag{2.7b}$$

where the index $i(K)$ is defined in (2.1). The last equality in the above equation holds by virtue of the compatibility property. For example, we will have not only $\phi_K = \phi(x_K)$, $\lambda_K = \lambda(x_K)$, $\eta_K(s) = \eta(s, x_K)$, $\mathcal{S}_K(p) = \mathcal{S}(p, x_K)$ but also $\mathbb{\epsilon}_K(s) = \mathbb{\epsilon}(s, x_K)$. Likewise, we shall be writing $f_{K\sigma}(\cdot) = f(\cdot, x_{K\sigma})$ for the mirror cell without any ambiguity: if $\sigma \in \mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}^{\rm N}$, then $x_{K\sigma}$ is a cell-center; if $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$, then $x_{K\sigma}$ lies on the boundary but does not belong to an interface between subdomains.

The size $h_\mathscr{T}$ and the regularity $\zeta_\mathscr{T}$ of the mesh are respectively defined by

$$h_\mathscr{T} = \max_{K\in\mathscr{T}} \operatorname{diam}(K), \qquad \zeta_\mathscr{T} = \min_{K\in\mathscr{T}} \left(\frac{1}{\operatorname{Card} \mathscr{E}_K} \min_{\sigma\in\mathscr{E}_K} \frac{d_{K\sigma}}{\operatorname{diam}(K)} \right). \tag{2.8}$$

The time discretization is given by $(t^n)_{0\leqslant 1\leqslant N}$ with $0 = t^0 < t^1 < \cdots < t^N = T$. We denote by $\Delta t^n = t^n - t^{n-1}$ for all $n \in \{1, \ldots, N\}$ and by $\boldsymbol{\Delta t} = (\Delta t^n)_{1\leqslant n\leqslant N}$.

2.2. Upstream mobility TPFA finite volume scheme

Given a discrete saturation profile $\left(s_K^{n-1}\right)_{K\in\mathscr{T}} \in [0,1]^\mathscr{T}$ at time t^{n-1}, $n \in \{1, \ldots, N\}$, we seek for a discrete pressure profile $(p_K^n)_{K\in\mathscr{T}} \in \mathbb{R}^\mathscr{T}$ at time t^n solution to the following nonlinear system of equations. Taking advantage of the notational shorthand (2.7b), we define

$$s_K^n = \mathcal{S}_K(p_K^n), \qquad K \in \mathscr{T},\ n \geqslant 1. \tag{2.9}$$

The volume balance (1.9a) is then discretized into

$$m_K \phi_K \frac{s_K^n - s_K^{n-1}}{\Delta t^n} + \sum_{\sigma\in\mathscr{E}_K} m_\sigma F_{K\sigma}^n = 0, \qquad K \in \mathscr{T},\ n \geqslant 1, \tag{2.10}$$

using the approximation

$$F_{K\sigma}^n = \frac{1}{d_\sigma} \lambda_\sigma \eta_\sigma^n \left(\vartheta_K^n - \vartheta_{K\sigma}^n\right), \quad \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ n \geqslant 1, \tag{2.11a}$$

for the flux (1.1b), with

$$\vartheta_K^n = p_K^n + \psi_K, \qquad \vartheta_{K\sigma}^n = p_{K\sigma}^n + \psi_{K\sigma}, \tag{2.11b}$$

where the mirror values $p^n_{K\sigma}$ and $\psi_{K\sigma}$ are given by (2.6). In the numerical flux (2.11a), the edge permeabilities $(\lambda_\sigma)_{\sigma\in\mathscr{E}}$ are set to

$$\lambda_\sigma = \begin{cases} \dfrac{\lambda_K\lambda_L d_\sigma}{\lambda_K d_{L,\sigma} + \lambda_L d_{K,\sigma}} & \text{if } \sigma = K|L \in \mathscr{E}_{\rm int}, \\ \lambda_K & \text{if } \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\rm ext}, \end{cases}$$

while the edge mobilities are upwinded according to

$$\eta^n_\sigma = \begin{cases} \eta_K\left(s^n_K\right) & \text{if } \vartheta^n_K > \vartheta^n_{K\sigma}, \\ \frac{1}{2}\left(\eta_K(s^n_K) + \eta_{K\sigma}\left(s^n_{K\sigma}\right)\right) & \text{if } \vartheta^n_K = \vartheta^n_{K\sigma}, \\ \eta_{K\sigma}\left(s^n_{K\sigma}\right) & \text{if } \vartheta^n_K < \vartheta^n_{K\sigma}. \end{cases} \tag{2.11c}$$

In practice, the definition of η^n_σ when $\vartheta^n_K = \vartheta^n_{K\sigma}$ has no influence on the scheme. We choose here to give a symmetric definition that does not depend on the orientation of the edge σ in order to avoid ambiguities.

The boundary condition $p^{\rm D}$ is discretized into

$$\begin{cases} p^{\rm D}_K = \frac{1}{m_K}\int_K p^{\rm D}(x)\,{\rm d}x & \text{for } K \in \mathscr{T}, \\ p^{\rm D}_\sigma = \frac{1}{m_\sigma}\int_\sigma p^{\rm D}(x)\,{\rm d}\nu^{d-1}(x) & \text{for } \sigma \in \mathscr{E}^{\rm D}_{\rm ext}, \end{cases} \tag{2.12}$$

whereas the initial condition is discretized into

$$s^0_K = \frac{1}{m_K}\int_K s^0(x)\,{\rm d}x, \qquad \text{for } K \in \mathscr{T}. \tag{2.13}$$

The Dirichlet boundary condition is encoded in the fluxes (2.11a) by setting

$$p^n_\sigma = p^{\rm D}_\sigma, \qquad \forall \sigma \in \mathscr{E}^{\rm D}_{\rm ext},\ n \geqslant 1. \tag{2.14}$$

Bearing in mind the definition (2.6) of the mirror values for $\sigma \in \mathscr{E}^{\rm N}_{\rm ext}$, the no-flux boundary condition across $\sigma \in \mathscr{E}^{\rm N}_{\rm ext}$ is automatically encoded, *i.e.*, $F^n_{K\sigma} = 0$ for all $\sigma \in \mathscr{E}_K \cap \mathscr{E}^{\rm N}_{\rm ext}$, $K \in \mathscr{T}$ and $n \geqslant 1$.

In what follows, we denote by $\boldsymbol{p}^n = (p^n_K)_{K\in\mathscr{T}}$ for $1 \leqslant n \leqslant N$, and by $\boldsymbol{s}^n = (s^n_K)_{K\in\mathscr{T}}$ for $0 \leqslant n \leqslant N$. Besides, we set $\boldsymbol{p}^{\rm D} = ((p^{\rm D}_K)_{K\in\mathscr{T}}, (p^{\rm D}_\sigma)_{\sigma\in\mathscr{E}^{\rm D}})$.

2.3. Main results and organization of the paper

The theoretical part of this paper includes two main results. The first one, which emerges from the analysis at fixed grid, states that the schemes admits a unique solution $(\boldsymbol{p}^n)_{1\leqslant n\leqslant N}$.

Theorem 2.2. *For all $n \in \{1,\dots,N\}$, there exists a unique solution $\boldsymbol{p}^n$ to the scheme (2.9)–(2.11c).*

With Theorem 2.2 at hand, we define the approximate pressure $p_{\mathscr{T},\Delta t}$ by

$$p_{\mathscr{T},\boldsymbol{\Delta t}}(t,x) = p^n_K \qquad \text{for } (t,x) \in (t^{n-1},t^n]\times K. \tag{2.15a}$$

We also define the approximate saturation as

$$s_{\mathscr{T},\boldsymbol{\Delta t}} = \mathcal{S}(p_{\mathscr{T},\boldsymbol{\Delta t}},x). \tag{2.15b}$$

The second main result guarantees the convergence towards a weak solution of the sequence of approximate solutions as the mesh size and the time steps tend to 0. Let $(\mathscr{T}_m,\mathscr{E}_m,(x_K)_{K\in\mathscr{T}_m})_{m\geqslant 1}$ be a sequence of admissible discretizations of the domain Ω in the sense of Definition 2.1 such that

$$h_{\mathscr{T}_m} \underset{m\to\infty}{\longrightarrow} 0, \qquad \sup_{m\geqslant 1}\ \zeta_{\mathscr{T}_m} =: \zeta < +\infty, \tag{2.16}$$

where the size $h_{\mathscr{T}_m}$ and the regularity $\zeta_{\mathscr{T}_m}$ are defined in (2.8). Let $(\boldsymbol{\Delta t}_m)_{m\geqslant 1}$ be time discretizations of $(0,T)$ such that

$$\lim_{m\to\infty}\ \max_{1\leqslant n\leqslant N_m} \Delta t^n_m = 0. \tag{2.17}$$

Theorem 2.3. *There exists a weak solution $p : Q_T \to \mathbb{R}$ in the sense of Definition 1.1 such that, up to a subsequence,*

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \mathcal{S}(p,x) \qquad \text{a.e. in } Q_T, \tag{2.18a}$$

$$\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to\infty}{\longrightarrow} \Upsilon(p) \qquad \text{weakly in } L^2(Q_T). \tag{2.18b}$$

The rest of this paper is outlined as follows. Section 3 is devoted to the numerical analysis at fixed grid. This encompasses the existence and uniqueness result stated in Theorem 2.2 as well as *a priori* estimates that will help proving Theorem 2.3. The convergence of the scheme, which is taken up in Section 4, relies on compactness arguments, which require *a priori* estimates that are uniform w.r.t. the grid. These estimates are mainly adaptations to the discrete setting of their continuous counterparts that arised in the stability analysis sketched out in Section 1.2. These estimates are shown in Section 4.1 to provide some compactness on the sequence of approximate solutions. In Section 4.2, we show that these compactness properties together with the *a priori* estimates are sufficient to identify any limit of an approximate solution as a weak solution to the problem.

In Section 5, we provide some details about the practical numerical resolution by laying emphasis on the switch of variable for selecting the primary unknown and on the mesh refinement at an interface in order to better enforce pressure continuity. Finally, in Section 6, numerical experiments on two configurations (drying and filling cases) for two capillary pressure models (Brooks–Corey and van Genuchten–Mualem) testify to the relevance of the local refinement strategy as a simple technique to preserve accuracy.

Remark 2.4. Theorem 2.3 only states the convergence of the scheme up to a subsequence. In the case where the weak solution is unique, then the whole sequence of approximate solutions would converge towards this solution. As far as we know, uniqueness of the weak solutions to Richards' equation is in general an open problem for heterogeneous media where $x \mapsto \mathcal{S}(p,x)$ is discontinuous. Uniqueness results are however available in the one-dimensional setting for a slightly more restrictive notion of solutions, *cf.*, [14], or under additional assumptions on the nonlinearities $\eta_i, \mathcal{S}_i$, *cf.*, [13].

3. Analysis at fixed grid

3.1. Some uniform *a priori* estimates

In this section, our aim is to derive *a priori* estimates on the solutions to the scheme (2.9)–(2.13). These estimates will be at the core of the existence proof of a solution to the scheme. They will also play a key role in proving the convergence of the scheme.

The main estimate on which our analysis relies is a discrete counterpart of (1.12). We recall that a_σ is the transmissivity introduced in (2.5).

Proposition 3.1. *There exist two constants C_1, C_2 depending only on λ, μ, p^{D}, ψ, ζ, Ω, T, ϕ, and $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$ such that*

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(p_K^n - p_{K\sigma}^n\right)^2 \leqslant C_1, \tag{3.1a}$$

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\vartheta_K^n - \vartheta_{K\sigma}^n\right)^2 \leqslant C_2. \tag{3.1b}$$

In (3.1), the relationship between σ and K is to be understood as follows. For an inner edge $\sigma \in \mathscr{E}_{\mathrm{int}}$, although it can be written as $\sigma = K|L$ or $L|K$, only one of these contributes to the sum. For a boundary edge $\sigma \in \mathscr{E}_{\mathrm{ext}}$, there is only one cell K such that $\sigma \in \mathscr{E}_K$, so there is no ambiguity in the sum.

Proof. Multiplying (2.10) by $\Delta t^n(p_K^n - p_K^{\mathrm{D}})$, summing over $K \in \mathscr{T}$ and $n \in \{1, \dots, N\}$, and carrying out discrete integration by parts yield

$$\mathtt{A} + \mathtt{B} = 0, \tag{3.2}$$

where we have set

$$\mathtt{A} = \sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \phi_K \left(s_K^n - s_K^{n-1}\right)\left(p_K^n - p_K^{\mathrm{D}}\right), \tag{3.3a}$$

$$\mathtt{B} = \sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\vartheta_K^n - \vartheta_{K\sigma}^n\right)\left(p_K^n - p_K^{\mathrm{D}} - p_{K\sigma}^n + p_{K\sigma}^{\mathrm{D}}\right). \tag{3.3b}$$

The discrete energy density function $\mathscr{E}_K : [0,1] \to \mathbb{R}_+$, defined by means of the notation (2.7) from the functions $f_i = \mathscr{E}_i$ introduced in (1.7), is convex by construction. Consequently,

$$\mathscr{E}_K\left(s_K^{n-1}\right) - \mathscr{E}_K\left(s_K^n\right) \geqslant \mathscr{E}_K'\left(s_K^n\right)\left(s_K^{n-1} - s_K^n\right) = \phi_K\left(p_K^n - p_K^{\mathrm{D}}\right)\left(s_K^{n-1} - s_K^n\right).$$

Therefore, the quantity $\mathtt{A}$ of (3.3a) can be bounded below by

$$\mathtt{A} \geqslant \sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \left(\mathscr{E}_K(s_K^n) - \mathscr{E}_K\left(s_K^{n-1}\right)\right) = \sum_{K \in \mathscr{T}} m_K \left(\mathscr{E}_K\left(s_K^N\right) - \mathscr{E}_K\left(s_K^0\right)\right) \geqslant -C_{\mathtt{A}}, \tag{3.4}$$

the last inequality being a consequence of the boundedness of $\mathscr{E}_K$ on $[0,1]$.

Writing $\vartheta = p + \psi$ and expanding each summand of (3.3b), we can split $\mathtt{B}$ into

$$\mathtt{B} = \mathtt{B}_1 + \mathtt{B}_2 + \mathtt{B}_3,$$

with

$$\mathtt{B}_1 = \sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(p_K^n - p_{K\sigma}^n\right)^2,$$

$$\mathtt{B}_2 = \sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(p_K^n - p_{K\sigma}^n\right)\left(\psi_K - \psi_{K\sigma} - p_K^{\mathrm{D}} + p_{K\sigma}^{\mathrm{D}}\right),$$

$$\mathtt{B}_3 = -\sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\psi_K - \psi_{K\sigma}\right)\left(p_K^{\mathrm{D}} - p_{K\sigma}^{\mathrm{D}}\right).$$

It follows from Lemma 9.4 of [29] and from the boundedness of η that there exists a constant C depending only on λ, μ, $\zeta_{\mathscr{T}}$ and Ω such that

$$\sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(p_K^{\mathrm{D}} - p_{K\sigma}^{\mathrm{D}}\right)^2 \leqslant C \left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d}^2, \tag{3.5a}$$

$$\sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\psi_K - \psi_{K\sigma}\right)^2 \leqslant C \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}^2. \tag{3.5b}$$

Thanks to these estimates and to the Cauchy–Schwarz inequality, we have

$$\mathtt{B}_3 \geqslant -CT \left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d} \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}.$$

On the other hand, Young's inequality provides

$$\mathtt{B}_2 \geqslant -\frac{1}{2}\mathtt{B}_1 - CT\left(\left\|\nabla p^{\mathrm{D}}\right\|_{L^2(\Omega)^d}^2 + \left\|\nabla \psi\right\|_{L^\infty(\Omega)^d}^2\right).$$

Hence,

$$\mathtt{B} \geqslant \frac{1}{2}\sum_{n=1}^{N}\Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(p_K^n - p_{K\sigma}^n\right)^2 - C_{\mathtt{B}}, \tag{3.6}$$

by setting $C_{\mathtt{B}} = CT\left(\|\nabla p^{\mathrm{D}}\|_{L^2(\Omega)^d}^2 + \|\nabla\psi\|_{L^\infty(\Omega)^d}^2 + \|\nabla p^{\mathrm{D}}\|_{L^2(\Omega)^d}\|\nabla\psi\|_{L^\infty(\Omega)^d}\right)$. Inserting (3.4) and (3.6) into (3.2), we recover (3.1a) with $C_1 = 2(C_{\mathtt{A}} + C_{\mathtt{B}})$.

From (3.1a), we can deduce (3.1b) by elementary manipulations. □

So far, we have not used the upwind choice (2.11c) for the mobilities η_σ^n. This will be done in the next lemma, where we derive a more useful variant of estimate (3.1a), in which η_σ^n is replaced by $\overline{\eta}_\sigma^n$ defined below. In a homogeneous medium, $\overline{\eta}_\sigma^n \geqslant \eta_\sigma^n$ so that the new estimate (3.8) seems to be stronger than (3.1a).

We begin by introducing the functions $\breve{\eta}_\sigma : \mathbb{R} \to (0, 1/\mu]$ defined for $\sigma \in \mathscr{E}$ by

$$\breve{\eta}_\sigma(p) = \min\left\{\eta_K \circ \mathcal{S}_K(p), \eta_{K\sigma} \circ \mathcal{S}_{K\sigma}(p)\right\}, \qquad \forall p \in \mathbb{R}. \tag{3.7a}$$

By virtue of assumptions (1.6), each argument of the minimum function is nondecreasing and positive function of $p \in \mathbb{R}$. As a result, $\breve{\eta}_\sigma$ is also a nondecreasing and positive function of $p \in \mathbb{R}$. Note that $\breve{\eta}_\sigma = \eta_i \circ \mathcal{S}_i$ for all $\sigma \in \mathscr{E}_i$, while for interface edges $\sigma \subset \Gamma_{i,j}$, the mere inequality $\breve{\eta}_\sigma \leqslant \eta_i \circ \mathcal{S}_i$ holds. Next, we consider the intervals

$$\mathfrak{J}_\sigma^n = [p_K^n \perp p_{K\sigma}^n,\, p_K^n \top p_{K\sigma}^n], \qquad \text{for } \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ 1 \leqslant n \leqslant N, \tag{3.7b}$$

with the notations $a\perp b = \min(a,b)$ and $a\top b = \max(a,b)$. At last, we set

$$\overline{\eta}_\sigma^n = \max_{p\in\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p), \qquad \text{for } \sigma \in \mathscr{E},\ 1 \leqslant n \leqslant N. \tag{3.7c}$$

Lemma 3.2. *There exists a constant C_3 depending on the same data as C_1 such that*

$$\sum_{n=1}^{N}\Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \overline{\eta}_\sigma^n \left(p_K^n - p_{K\sigma}^n\right)^2 \leqslant C_3. \tag{3.8}$$

Proof. We partition the set $\mathscr{E}$ of edges into three subsets, namely,

$$\mathscr{E}_+^n = \left\{\sigma \mid \vartheta_K^n > \vartheta_{K\sigma}^n\right\}, \quad \mathscr{E}_-^n = \left\{\sigma \mid \vartheta_K^n < \vartheta_{K\sigma}^n\right\}, \quad \mathscr{E}_0^n = \left\{\sigma \mid \vartheta_K^n = \vartheta_{K\sigma}^n\right\}.$$

Invoking $\breve{\eta}_\sigma = \min(\eta_K \circ \mathcal{S}_K,\, \eta_{K\sigma} \circ \mathcal{S}_{K\sigma})$, we can minorize the left-hand side of (3.1a) to obtain

$$\sum_{n=1}^{N}\Delta t^n \left[\sum_{\sigma\in\mathscr{E}_+^n} a_\sigma\lambda_\sigma\breve{\eta}_\sigma\left(p_K^n\right)\left(p_K^n - p_{K\sigma}^n\right)^2 + \sum_{\sigma\in\mathscr{E}_-^n} a_\sigma\lambda_\sigma\breve{\eta}_\sigma\left(p_{K\sigma}^n\right)\left(p_K^n - p_{K\sigma}^n\right)^2 \right.$$
$$\left. + \sum_{\sigma\in\mathscr{E}_0^n} a_\sigma\lambda_\sigma\tfrac{1}{2}\left(\breve{\eta}_\sigma(p_K^n) + \breve{\eta}_\sigma\left(p_{K\sigma}^n\right)\right)\left(p_K^n - p_{K\sigma}^n\right)^2\right] \leqslant C_1.$$

Starting from this inequality and using the boundedness of η_i and ψ, we can readily show that there exists a constant C depending on the same data as C_1 such that

$$\mathtt{D}_1 := \sum_{n=1}^{N}\Delta t^n \left[\sum_{\sigma\in\mathscr{E}_+^n} a_\sigma\lambda_\sigma\breve{\eta}_\sigma(p_K^n)\left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) + \sum_{\sigma\in\mathscr{E}_-^n} a_\sigma\lambda_\sigma\breve{\eta}_\sigma\left(p_{K\sigma}^n\right)\left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right)\right] \leqslant C,$$

in which the sum over $\mathscr{E}_0^n$ was omitted because all of its summands vanish. Simlarly to what was pointed out in equation (2.9) in [1], we notice that since η_σ is nondecreasing w.r.t. p, it is straightforward to check that the definition

$$\breve{\eta}_\sigma^n := \begin{cases} \breve{\eta}_\sigma\left(p_K^n\right) & \text{if } \vartheta_K^n > \vartheta_{K\sigma}^n, \\ \frac{1}{2}\left(\breve{\eta}_\sigma\left(p_K^n\right) + \breve{\eta}_\sigma\left(p_{K\sigma}^n\right)\right) & \text{if } \vartheta_K^n = \vartheta_{K\sigma}^n, \\ \breve{\eta}_\sigma\left(p_{K\sigma}^n\right) & \text{if } \vartheta_K^n < \vartheta_{K\sigma}^n \end{cases} \tag{3.9}$$

exactly amounts to

$$\breve{\eta}_\sigma^n = \begin{cases} \max_{p\in\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p) & \text{if } \left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) > 0, \\ \frac{1}{2}\left(\breve{\eta}_\sigma\left(p_K^n\right) + \breve{\eta}_\sigma(p_{K\sigma}^n)\right) & \text{if } \left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) = 0, \\ \min_{p\in\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma(p) & \text{if } \left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) < 0. \end{cases} \tag{3.10}$$

Taking advantage of this equivalence, we can transform $\mathtt{D}_1$ into

$$\mathtt{D}_1 = \sum_{n=1}^N \Delta t^n \left[\sum_{\sigma\in\mathscr{E}_>^n} a_\sigma\lambda_\sigma \max_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma\left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) + \sum_{\sigma\in\mathscr{E}_<^n} a_\sigma\lambda_\sigma \min_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma\left(p_K^n - p_{K\sigma}^n\right)\left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) \right] \leqslant C, \tag{3.11}$$

where $\mathscr{E}_>^n = \{\sigma \,|\, (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) > 0\}$ and $\mathscr{E}_<^n = \{\sigma \,|\, (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) < 0\}$. The second sum over $\mathscr{E}_<^n$ contains only negative summands and can be further minorized if $\min_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma$ is replaced by $\max_{\mathfrak{J}_\sigma^n} \breve{\eta}_\sigma$. In other words,

$$\mathtt{D}_2 := \sum_{n=1}^N \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma\lambda_\sigma \overline{\eta}_\sigma^n (p_K^n - p_{K\sigma}^n)(\vartheta_K^n - \vartheta_{K\sigma}^n) \leqslant \mathtt{D}_1 \leqslant C.$$

Writing $\vartheta = p + \psi$, expanding each summand of $\mathtt{D}_2$ and applying Young's inequality, we end up with

$$\frac{1}{2}\sum_{n=1}^N \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma\lambda_\sigma \overline{\eta}_\sigma^n \left[(p_K^n - p_{K\sigma}^n)^2 - (\psi_K^n - \psi_{K\sigma}^n)^2\right] \leqslant \mathtt{D}_2 \leqslant C.$$

Estimate (3.8) finally follows from the boundedness of η, $1/\lambda$ and ψ. □

The above lemma has several important consequences for the analysis. Let us start with discrete counterparts to estimations (1.14) and (1.15).

Corollary 3.3. *Let C_3 be the constant in Lemma 3.2. Then,*

$$\sum_{n=1}^N \Delta t^n \sum_{i=1}^I \sum_{\sigma\in\mathscr{E}_i} a_\sigma \left(\Theta_i\left(p_K^n\right) - \Theta_i\left(p_{K\sigma}^n\right)\right)^2 \leqslant C_3, \tag{3.12a}$$

$$\sum_{n=1}^N \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma \left(\Upsilon\left(p_K^n\right) - \Upsilon\left(p_{K\sigma}^n\right)\right)^2 \leqslant C_3. \tag{3.12b}$$

Moreover, there exists two constants C_4, C_5 depending on the same data as C_1 and additionnally on $\|\sqrt{\eta_i \circ \mathcal{S}_i}\|_{L^1(\mathbb{R}_-)}$, $1 \leqslant i \leqslant I$, such that

$$\sum_{n=1}^N \Delta t^n \sum_{K\in\mathscr{T}} m_K |\Upsilon(p_K^n)|^2 \leqslant C_4, \tag{3.13a}$$

$$\sum_{n=1}^N \Delta t^n \sum_{i=1}^I \sum_{K\in\mathscr{T}_i} m_K |\Theta_i(p_K^n)|^2 \leqslant C_5. \tag{3.13b}$$

Proof. Consider those edges $\sigma \in \mathscr{E}_i$ – defined in (2.3a) – corresponding to some fixed $i \in \{1, \dots, I\}$, for which $\breve{\eta}_\sigma = \eta_i \circ \mathcal{S}_i = |\Theta_i'|^2$ and $\overline{\eta}_\sigma^n = \max_{\mathfrak{I}_\sigma^n} |\Theta_i'|^2$ due to (1.13a). By summing the elementary inequality

$$(\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))^2 \leqslant \overline{\eta}_\sigma^n \, (p_K^n - p_{K\sigma}^n)^2,$$

over $\sigma \in \mathscr{E}_i$, $i \in \{1, \dots, I\}$ and $n \in \{1, \dots, N\}$ using appropriate weights, we get

$$\sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{\sigma \in \mathscr{E}_i} a_\sigma (\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n))^2 \leqslant \sum_{n=1}^{N} \Delta t^n \sum_{i=1}^{I} \sum_{\sigma \in \mathscr{E}_i} a_\sigma \overline{\eta}_\sigma^n \, (p_K^n - p_{K\sigma}^n)^2,$$

whose right-hand side is obviously less than C_3, thanks to (3.8). This proves (3.12a).

Similarly, the respective definitions of $\overline{\eta}_\sigma^n$ and Υ have been tailored so that $\max_{\mathfrak{I}_\sigma^n} |\Upsilon'|^2 \leqslant \overline{\eta}_\sigma^n$ for all $\sigma \in \mathscr{E}$. As a consequence,

$$(\Upsilon(p_K^n) - \Upsilon(p_{K\sigma}^n))^2 \leqslant \overline{\eta}_\sigma^n (p_K^n - p_{K\sigma}^n)^2.$$

Summing these inequalities over $\sigma \in \mathscr{E}$ and $n \in \{1, \dots, N\}$ with appropriate weights and invoking (3.8), we prove (3.12b).

The argument for (3.13a) is subtler. Starting from the basic inequality

$$\left(\Upsilon\left(p_K^n\right) - \Upsilon\left(p_K^{\mathrm{D}}\right) - \Upsilon\left(p_{K\sigma}^n\right) + \Upsilon\left(p_{K\sigma}^{\mathrm{D}}\right)\right)^2 \leqslant 2\left(\Upsilon(p_K^n) - \Upsilon(p_{K\sigma}^n)\right)^2 + 2\left(\Upsilon\left(p_K^{\mathrm{D}}\right) - \Upsilon\left(p_{K\sigma}^{\mathrm{D}}\right)\right)^2,$$

we apply the discrete Poincaré inequality of Lemma 9.1 from [29] – which is legitimate since Γ^{D} has positive measure – followed by Lemma 9.4 of [29] to obtain

$$\sum_{n=1}^{N} \Delta t^n \sum_{K \in \mathscr{T}} m_K (\Upsilon(p_K^n) - \Upsilon(p_K^{\mathrm{D}}))^2 \leqslant 2C_{\mathrm{P},\mathscr{T}} \left(C_3 + C_\zeta T \|\Upsilon'\|_\infty \|\nabla p^{\mathrm{D}}\|^2\right),$$

where $C_{\mathrm{P},\mathscr{T}}$ denotes the discrete Poincaré constant, and C_ζ is the quantity appearing in Lemma 9.4 of [29] and only depends on $\zeta_{\mathscr{T}}$. This entails (3.13a) with $C_4 = 4C_{\mathrm{P},\mathscr{T}} \left(C_3 + C_\zeta T \|\Upsilon'\|_\infty \|\nabla p^{\mathrm{D}}\|^2\right) + 2m_\Omega T \|\Upsilon(p^{\mathrm{D}})\|_\infty^2$.

The last estimate (3.13b) results from the comparison (1.16) of the nonlinearities Θ_i and Υ. □

The purpose of the next lemma is to work out a weak estimate on the discrete counterpart of $\partial_t s$, which will lead to compactness properties in Section 4.1. For $\varphi \in C_c^\infty(Q_T)$, let

$$\varphi_K^n = \frac{1}{m_K} \int_K \varphi(t^n, x) \, \mathrm{d}x, \qquad \forall K \in \mathscr{T}, \ 1 \leqslant n \leqslant N.$$

Lemma 3.4. *There exists a constant C_6 depending on the same data as C_1 such that*

$$\sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \phi_K \left(s_K^n - s_K^{n-1}\right) \varphi_K^n \leqslant C_6 \|\nabla \varphi\|_{L^\infty(Q_T)^d}, \quad \forall \varphi \in C_c^\infty(Q_T). \tag{3.14}$$

Proof. Multiplying (2.10) by $\Delta t^n \, \varphi_K^n$, summing over $K \in \mathscr{T}$ and $n \in \{1, \cdots, N\}$ and carrying out discrete integration by parts, we end up with

$$\mathtt{A} := \sum_{n=1}^{N} \sum_{K \in \mathscr{T}} m_K \phi_K \left(s_K^n - s_K^{n-1}\right) \varphi_K^n = -\sum_{n=1}^{N} \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\vartheta_K^n - \vartheta_{K\sigma}^n\right) \left(\varphi_K^n - \varphi_{K\sigma}^n\right).$$

Applying the Cauchy–Schwarz inequality and using (3.1b), we get

$$\mathtt{A}^2 \leqslant C_2 \frac{\max_i \lambda_i}{\mu} \sum_n \Delta t^n \sum_{\sigma \in \mathscr{E}} a_\sigma (\varphi_K^n - \varphi_{K\sigma}^n)^2. \tag{3.15}$$

The conclusion (3.14) is then reached by means of the property (see [2], Sect. 4.4)

$$\sum_{n=1}^{N} \Delta t^n \sum_{\sigma\in\mathscr{E}} a_\sigma(\varphi_K^n - \varphi_{K\sigma}^n)^2 \leqslant C\|\nabla\varphi\|^2_{L^\infty(Q_T)^d}$$

for some C depending only on Ω, T and the mesh regularity $\zeta_{\mathscr{T}}$. □

3.2. Existence of a solution to the scheme

The statements of the previous section are all uniform w.r.t. the mesh and are meant to help us passing to the limit in the next section. In contrast, the next lemma provides a bound on the pressure that depends on the mesh size and on the time-step. This property is needed in the process of ensuring the existence of a solution to the numerical scheme.

Lemma 3.5. *There exist two constants C_7, C_8 depending on $\mathscr{T}$, Δt^n as well as on the data of the continuous model λ, μ, p^{D}, ψ, ζ, Ω, T, ϕ, $\|\mathcal{S}_i\|_{L^1(\mathbb{R}_-)}$ and $\|\sqrt{\eta_i\circ\mathcal{S}_i}\|_{L^1(\mathbb{R}_-)}$, $1\leqslant i\leqslant I$, such that*

$$-C_7 \leqslant p_K^n \leqslant C_8, \quad \forall K\in\mathscr{T},\ n\in\{1,\dots,N\}. \tag{3.16}$$

Proof. From (3.13a) and from $\Upsilon(p) = p\sqrt{\min_i \lambda_i/\mu}$ for $p\geqslant 0$, we deduce that

$$p_K^n \leqslant \sqrt{\frac{\mu C_4}{\Delta t^n m_K \min_i \lambda_i}}, \qquad \forall K\in\mathscr{T},\ 1\leqslant n\leqslant N.$$

Hence, the upper-bound C_8 is found by maximizing the right-hand side over $K\in\mathscr{T}$ and $n\in\{1,\dots,N\}$.

To show that p_K^n is bounded from below, we employ a strategy that was developed in [15] and extended to the case of Richards' equation in Lemma 3.10 of [1]. From (2.12), (2.14) and the boundedness of p^{D}, it is easy to see that

$$p_\sigma^n \geqslant \inf_{x\in\partial\Omega} p^{\mathrm{D}}(x), \quad \forall\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}.$$

Estimate (3.8) then shows that for all $K\in\mathscr{T}$ such that $\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}\neq\varnothing$, we have

$$p_K^n \geqslant p_\sigma^n - \sqrt{\frac{C_3}{\Delta t^n a_\sigma \check{\eta}_\sigma(p_\sigma^n)}} =: \pi_K^n, \quad \forall\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}.$$

The quantity π_K^n is well-defined, since $\check{\eta}_\sigma(p_\sigma^n)>0$ for $p_\sigma^n>-\infty$, and does not depend on time, as p^{D} does not either. Furthermore, if p_K^n is bounded from below by some π_K, then the pressure in all its neighboring cells $L\in\mathscr{T}$ such that $\sigma = K|L\in\mathscr{E}_K$ is bounded from below by

$$p_L^n \geqslant \pi_K^n - \sqrt{\frac{C_3}{\Delta t^n a_\sigma \check{\eta}_\sigma(\pi_K^n)}} =: \pi_L^n.$$

Again, π_L^n is well-defined owing to $\check{\eta}_\sigma(\pi_K^n)>0$. Since the mesh is finite and since the domain is connected, only a finite number of edge-crossings is required to create a path from a Dirichlet boundary edge $\sigma\in\mathscr{E}_{\mathrm{ext}}^{\mathrm{D}}$ to any prescribed cell $K\in\mathscr{T}$. Hence, the lower bound C_7 is found by minimizing π_K^n over $K\in\mathscr{T}$ and $n\in\{1,\dots,N\}$. □

Lemma 3.5 is a crucial step in the proof of the existence of a solution $\boldsymbol{p}^n = (p_K^n)_{K\in\mathscr{T}}$ to the scheme (2.9)–(2.14).

Proposition 3.6. *Given $\boldsymbol{s}^{n-1} = (s_K^{n-1})_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$, there exists a solution $\boldsymbol{p}^n\in\mathbb{R}^{\mathscr{T}}$ to the scheme (2.9)–(2.14).*

The proof relies on a standard topological degree argument and is omitted here. However, we make the homotopy explicit for readers' convenience. Let $\gamma \in [0,1]$ be the homotopy parameter. We define the nondecreasing functions $\eta_i^{(\gamma)} : [0,1] \to \mathbb{R}_+$ by setting $\eta_i^{(\gamma)}(s) = (1-\gamma)/\mu + \gamma\eta_i(s)$ for $s \in [0,1]$, and we seek a solution $\boldsymbol{p}^{(\gamma)} = (p_K^{(\gamma)})_{K\in\mathscr{T}}$ to the problem

$$\gamma m_K \phi_K \frac{\mathcal{S}_K(p_K^{(\gamma)}) - s_K^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathscr{E}_K} m_\sigma F_{K\sigma}^{(\gamma)} = 0, \qquad K \in \mathscr{T},\ \gamma \in [0,1], \tag{3.17a}$$

where the fluxes $F_{K\sigma}^{(\gamma)}$ are defined by

$$F_{K\sigma}^{(\gamma)} = \frac{1}{d_\sigma}\lambda_\sigma \eta_\sigma^{(\gamma)} \left(\vartheta_K^{(\gamma)} - \vartheta_{K\sigma}^{(\gamma)}\right), \qquad \sigma \in \mathscr{E}_K,\ K \in \mathscr{T},\ \gamma \in [0,1] \tag{3.17b}$$

with $\vartheta^{(\gamma)} = p^{(\gamma)} + \psi$ and using the upwind mobilities

$$\eta_\sigma^{(\gamma)} = \begin{cases} \eta_K^{(\gamma)}\left(\mathcal{S}_K\left(p_K^{(\gamma)}\right)\right) & \text{if } \vartheta_K^{(\gamma)} > \vartheta_{K\sigma}^{(\gamma)}, \\ \frac{1}{2}\left(\eta_K^{(\gamma)}\left(\mathcal{S}_K\left(p_K^{(\gamma)}\right)\right) + \eta_{K\sigma}^{(\gamma)}\left(\mathcal{S}_{K\sigma}\left(p_K^{(\gamma)}\right)\right)\right) & \text{if } \vartheta_K^{(\gamma)} = \vartheta_{K\sigma}^{(\gamma)}, \\ \eta_{K\sigma}^{(\gamma)}\left(\mathcal{S}_{K\sigma}\left(p_K^{(\gamma)}\right)\right) & \text{if } \vartheta_K^{(\gamma)} < \vartheta_{K\sigma}^{(\gamma)}. \end{cases} \tag{3.17c}$$

At the Dirichlet boundary edges, we still set $p_\sigma^{(\gamma)} = p_\sigma^{\mathrm{D}}$. For $\gamma = 0$, the system is linear and invertible, while for $\gamma = 1$, system (3.17) coincides with the original system (2.9)–(2.14). *A priori* estimates on $\boldsymbol{p}^{(\gamma)}$ that are uniform w.r.t. $\gamma \in [0,1]$ (but not uniform w.r.t. $\mathscr{T}$ nor Δt^n) can be derived on the basis of what was exposed previously, so that one can unfold Leray–Schauder's machinery [22,39] to prove the existence of (at least) one solution to the scheme.

3.3. Uniqueness of the discrete solution

To complete the proof of Theorem 2.2, it remains to show that the solution to the scheme is unique. This is the purpose of the following proposition.

Proposition 3.7. *Given $\boldsymbol{s}^{n-1} = (s_K^{n-1})_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$, the solution $\boldsymbol{p}^n \in \mathbb{R}^{\mathscr{T}}$ to the scheme (2.9)–(2.14) is unique.*

Proof. The proof heavily rests upon the monotonicity properties inherited from the upwind choice (2.11c) for the mobilities. Indeed, due to the upwind choice of the mobility, the flux $F_{K\sigma}^n$ is a function of p_K^n and $p_{K\sigma}^n$ that is nondecreasing w.r.t. p_K^n and nonincreasing w.r.t. $p_{K\sigma}^n$. Moreover, by virtue of the monotonicity of $\mathcal{S}_K$, the discrete volume balance (2.10) can be cast under the abstract form

$$\mathcal{H}_K^n\left(p_K^n, (p_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}\right) = 0, \qquad \forall K \in \mathscr{T}, \tag{3.18}$$

where $\mathcal{H}_K^n$ is nondecreasing w.r.t. its first argument p_K^n and nonincreasing w.r.t. each of the remaining variables $(p_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}$.

Let $\widetilde{\boldsymbol{p}}^n = (\widetilde{p}_K^n)_{K\in\mathscr{T}}$ be another solution to the system (2.9)–(2.14), *i.e.*,

$$\mathcal{H}_K^n\left(\widetilde{p}_K^n, (\widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}\right) = 0, \qquad \forall K \in \mathscr{T}. \tag{3.19}$$

The nonincreasing behavior of $\mathcal{H}_K^n$ w.r.t. all its variables except the first one implies that

$$\mathcal{H}_K^n\left(p_K^n, (p_{K\sigma}^n \top \widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}\right) \leqslant 0, \qquad \mathcal{H}_K^n\left(\widetilde{p}_K^n, (p_{K\sigma}^n \top \widetilde{p}_{K\sigma}^n)_{\sigma\in\mathscr{E}_K}\right) \leqslant 0,$$

for all $K \in \mathscr{T}$, where $a \top b = \max(a,b)$. Since $p_K^n \top \widetilde{p}_K^n$ is either equal to p_K^n or to $\widetilde{p}_K^n$, we infer from the above inequalities that

$$\mathcal{H}_K^n \left(p_K^n \top \widetilde{p}_K^n, (p_{K\sigma}^n \top \widetilde{p}_{K\sigma}^n)_{\sigma \in \mathscr{E}_K} \right) \leqslant 0, \qquad \forall K \in \mathscr{T}. \tag{3.20}$$

By a similar argument, we can show that

$$\mathcal{H}_K^n \left(p_K^n \bot \widetilde{p}_K^n, (p_{K\sigma}^n \bot \widetilde{p}_{K\sigma}^n)_{\sigma \in \mathscr{E}_K} \right) \geqslant 0, \qquad \forall K \in \mathscr{T}, \tag{3.21}$$

where $a \bot b = \min(a,b)$. Subtracting (3.21) from (3.20) and summing over $K \in \mathscr{T}$, we find

$$\sum_{K \in \mathscr{T}} m_K \phi_K \frac{|s_K^n - \widetilde{s}_K^n|}{\Delta t^n} + \sum_{\sigma \in \mathscr{E}_{\rm ext}^{\rm D}} a_\sigma \lambda_\sigma \mathtt{R}_\sigma^n \leqslant 0, \tag{3.22}$$

where $s_K^n = \mathcal{S}_K(p_K^n)$, $\widetilde{s}_K^n = \mathcal{S}_K(\widetilde{p}_K^n)$ and

$$\begin{aligned} \mathtt{R}_\sigma^n = {} & \eta_K\left(s_K^n \top \widetilde{s}_K^n\right) \left(\vartheta_K^n \top \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ - \eta_K\left(s_\sigma^n\right) \left(\vartheta_\sigma^n - \vartheta_K^n \top \widetilde{\vartheta}_K^n \right)^+ \\ & - \eta_K\left(s_K^n \bot \widetilde{s}_K^n\right) \left(\vartheta_K^n \bot \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ + \eta_K\left(s_\sigma^n\right) \left(\vartheta_\sigma^n - \vartheta_K^n \bot \widetilde{\vartheta}_K^n \right)^+, \end{aligned} \tag{3.23}$$

with $s_\sigma^n = \mathcal{S}_K(p_\sigma^n)$. The top line of (3.23) expresses the upwinded flux of (3.20), while the bottom line of (3.23) is the opposite of the upwinded flux of (3.21). Note that, since $p_\sigma^n = p_\sigma^{\rm D}$ is prescribed at $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$, we have $\vartheta_\sigma^n = \vartheta_\sigma^n \top \widetilde{\vartheta}_\sigma^n = \vartheta_\sigma^n \bot \widetilde{\vartheta}_\sigma^n$. Upon inspection of the rearrangement

$$\begin{aligned} \mathtt{R}_\sigma^n = {} & \left[\eta_K\left(s_K^n \top \widetilde{s}_K^n\right) - \eta_K\left(s_K^n \bot \widetilde{s}_K^n\right) \right] \left(\vartheta_K^n \top \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ \\ & + \eta_K\left(s_K^n \bot \widetilde{s}_K^n\right) \left[\left(\vartheta_K^n \top \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ - \left(\vartheta_K^n \bot \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ \right] \\ & + \eta_K\left(s_\sigma^n\right) \left[\left(\vartheta_\sigma^n - \vartheta_K^n \bot \widetilde{\vartheta}_K^n \right)^+ - \left(\vartheta_\sigma^n - \vartheta_K^n \top \widetilde{\vartheta}_K^n \right)^+ \right], \end{aligned} \tag{3.24}$$

it is trivial that $\mathtt{R}_\sigma^n \geqslant 0$. As a consequence, (3.22) implies that $\mathtt{R}_\sigma^n = 0$ for all $\sigma \in \mathscr{E}_{\rm ext}^{\rm D}$ and that $s_K^n = \widetilde{s}_K^n$ for all $K \in \mathscr{T}$. At this stage, however, we cannot yet claim that $p_K^n = \widetilde{p}_K^n$, as the function $\mathcal{S}_K$ is not invertible.

Taking into account $s_K^n = \widetilde{s}_K^n$, the residue (3.24) becomes

$$\begin{aligned} \mathtt{R}_\sigma^n = {} & \eta_K(s_K^n) \left[\left(\vartheta_K^n \top \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ - \left(\vartheta_K^n \bot \widetilde{\vartheta}_K^n - \vartheta_\sigma^n \right)^+ \right] \\ & + \eta_K\left(s_\sigma^n\right) \left[\left(\vartheta_\sigma^n - \vartheta_K^n \bot \widetilde{\vartheta}_K^n \right)^+ - \left(\vartheta_\sigma^n - \vartheta_K^n \top \widetilde{\vartheta}_K^n \right)^+ \right], \end{aligned} \tag{3.25}$$

which can be lower-bounded by

$$\mathtt{R}_\sigma^n \geqslant \min(\eta_K(s_K^n), \eta_K(s_\sigma^n)) |\vartheta_K^n - \widetilde{\vartheta}_K^n| \tag{3.26}$$

thanks to the algebraic identities $a^+ - (-a)^+ = a$ and $a \top b - a \bot b = |a-b|$. In view of the lower-bound on the discrete pressures of Lemma 3.5, we deduce from (1.6b) that $s_K^n > 0$ and $\widetilde{s}_K^n > 0$. The increasing behavior of η_K implies, in turn, that $\eta_K(s_K^n) > 0$ and $\eta_K(\widetilde{s}_K^n) > 0$. Therefore, the conjunction of $\mathtt{R}_\sigma^n = 0$ and (3.26) yields $\vartheta_K^n = \widetilde{\vartheta}_K^n$ and hence $p_K^n = \widetilde{p}_K^n$ for all cells K having a Dirichlet boundary edge, *i.e.*, $\mathscr{E}_K \cap \mathscr{E}_{\rm ext}^{\rm D} \neq \varnothing$.

It remains to check that $p_K^n = \widetilde{p}_K^n$, or equivalently $\vartheta_K^n = \widetilde{\vartheta}_K^n$ for those cells $K \in \mathscr{T}$ that are far away from the Dirichlet part of the boundary. Subtracting (3.19) from (3.18) and recalling that $s_K^n = \widetilde{s}_K^n$, we arrive at

$$\sum_{\sigma \in \mathscr{E}_K} a_\sigma \lambda_\sigma \left\{ \eta_K\left(s_K^n\right) \left[(\vartheta_K^n - \vartheta_{K\sigma}^n)^+ - \left(\widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n \right)^+ \right] + \eta_{K\sigma}\left(s_{K\sigma}^n\right) \left[\left(\widetilde{\vartheta}_{K\sigma}^n - \widetilde{\vartheta}_K^n \right)^+ - (\vartheta_{K\sigma}^n - \vartheta_K^n)^+ \right] \right\} = 0. \tag{3.27}$$

Consider a cell $K \in \mathscr{T}$ where $\vartheta_K^n - \widetilde{\vartheta}_K^n$ achieves its maximal value, *i.e.*,

$$\vartheta_K^n - \widetilde{\vartheta}_K^n \geqslant \vartheta_L^n - \widetilde{\vartheta}_L^n, \qquad \forall L \in \mathscr{T}. \tag{3.28}$$

This entails that

$$\vartheta_K^n - \vartheta_{K\sigma}^n \geqslant \widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n, \qquad \forall \sigma \in \mathscr{E}_K,$$

so that the two brackets in the right-hand side of (3.27) are nonnegative. In fact, they both vanish by the positivity of $\eta_K(s_K^n)$ and $\eta_{K\sigma}(s_{K\sigma}^n)$. As a result, $\vartheta_K^n - \vartheta_{K\sigma}^n = \widetilde{\vartheta}_K^n - \widetilde{\vartheta}_{K\sigma}^n$ for all $\sigma \in \mathscr{E}_K$. This implies that $\vartheta_K^n - \widetilde{\vartheta}_K^n = \vartheta_L^n - \widetilde{\vartheta}_L^n$ for all the cells $L \in \mathscr{T}$ sharing an edge $\sigma = K|L$ with K, and thus that the cell L also achieves the maximality condition (3.28). The process can then be repeated over and over again. Since Ω is connected, we deduce that $\vartheta_K^n - \widetilde{\vartheta}_K^n$ is constant over $K \in \mathscr{T}$. The constant is finally equal to zero since $\vartheta_K^n = \widetilde{\vartheta}_K^n$ on the cells having a Dirichlet edge. □

4. Convergence analysis

Once existence and uniqueness of the discrete solution have been settled, the next question to be addressed is the convergence of the discrete solution towards a weak solution of the continuous problem, as the mesh-size and the time-step are progressively refined. In accordance with the general philosophy expounded in [29], the proof is built on compactness arguments. We start by highlighting compactness properties in Section 4.1, before identifying the limit values as weak solutions in Section 4.2.

4.1. Compactness properties

Let us define $G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} : Q_T \to \mathbb{R}^d$ and $J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} : Q_T \to \mathbb{R}^d$ by

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} d\dfrac{\Theta_i(p_{K\sigma}^n) - \Theta_i(p_K^n)}{d_\sigma} n_{K\sigma}, & \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

for $\sigma \in \mathscr{E}_{i,m}$, $1 \leqslant n \leqslant N_m$ and, respectively,

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = d\frac{\Upsilon(p_{K\sigma}^n) - \Upsilon(p_K^n)}{d_\sigma} n_{K\sigma}, \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \tag{4.2}$$

for $\sigma \in \mathscr{E}_m$, $1 \leqslant n \leqslant N_m$. We remind that $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \mathcal{S}(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, x)$ is the sequence of approximate saturation fields computed from that of approximate pressure fields $p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$ by (2.15b).

Proposition 4.1. *There exists a measurable function $p : Q_T \to \mathbb{R}$ such that $\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T);V)$ and $\Theta_i(p) \in L^2((0,T);H^1(\Omega_i))$, $1 \leqslant i \leqslant I$, such that, up to a subsequence,*

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \mathcal{S}(p,x) \qquad \text{a.e. in } Q_T, \tag{4.3a}$$

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \nabla\Theta_i(p) \qquad \text{weakly in } L^2(Q_{i,T})^d, \tag{4.3b}$$

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \nabla\Upsilon(p) \qquad \text{weakly in } L^2(Q_T)^d. \tag{4.3c}$$

Proof. We know from Corollary 3.3 that $\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ and $\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ are bounded w.r.t. m in $L^2(Q_{i,T})$ and $L^2(Q_T)$ respectively, while $G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ and $J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ are respectively bounded in $L^2(Q_{i,T})^d$ and $L^2(Q_T)^d$. In particular, there exist $\widehat{\Theta}_i \in L^2(Q_{i,T})$, $\widehat{\Upsilon} \in L^2(Q_T)$, $J \in L^2(Q_{i,T})^d$, and $J \in L^2(Q_T)^d$ such that

$$\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to+\infty}{\longrightarrow} \widehat{\Theta}_i \qquad \text{weakly in } L^2(Q_{i,T}), \tag{4.4a}$$

$$\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}) \underset{m\to+\infty}{\longrightarrow} \widehat{\Upsilon} \qquad \text{weakly in } L^2(Q_T), \tag{4.4b}$$

$$G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} G \qquad \text{weakly in } L^2(Q_{i,T})^d, \tag{4.4c}$$

$$J_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} J \qquad \text{weakly in } L^2(Q_T)^d. \tag{4.4d}$$

Establishing that $\widehat{\Theta}_i \in L^2((0,T);H^1(\Omega_i))$ and $\widehat{\Upsilon} \in L^2((0,T);H^1(\Omega))$ with $G = \nabla\widehat{\Theta}_i$ and $J = \nabla\widehat{\Upsilon}$ is now classical, see for instance Lemma 2 of [27] or Lemma 4.4 of [18].

The key points of this proof are the identification $\widehat{\Theta}_i = \Theta_i(p)$ and $\widehat{\Upsilon} = \Upsilon(p)$ for some measurable p, as well as the proofs of the almost everywhere convergence property (4.3a). The identification of the limit and the almost everywhere convergence can be handled simultaneously by using twice ([2], Thm. 3.9), once for $\Theta_i(p)$ and once for $\Upsilon(p)$. More precisely, Lemma 3.4 provides a control on the time variations of the approximate saturation $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$, whereas Corollary 3.3 provides some compactness w.r.t. space on $\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$ and $\Upsilon(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})$. Using further that $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \mathcal{S}_i \circ \Theta_i^{-1}\left(\Theta_i(p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})\right)$ with $\mathcal{S}_i \circ \Theta_i^{-1}$ nondecreasing and continuous, then one infers from Theorem 3.9 of [2] that

$$s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m\to+\infty}{\longrightarrow} \mathcal{S}_i \circ \Theta_i^{-1}(\widehat{\Theta}_i) \quad \textit{a.e.} \text{ in } Q_{i,T}.$$

Let $p = \Theta_i^{-1}(\widehat{\Theta}_i)$. Then, (4.3a) and (4.3b) hold. Proving (4.3a) and (4.3c) is similar, and the properties (4.3) can be assumed to hold for the same function p up to the extraction of yet another subsequence.

Finally, by applying the arguments developed in Section 4.2 of [8], we show that $\Upsilon(p)$ and $\Upsilon(p^{\mathrm{D}})$ share the same trace on $(0,T)\times\Gamma^{\mathrm{D}}$, hence $\Upsilon(p) - \Upsilon(p^{\mathrm{D}}) \in L^2((0,T);V)$. □

Let us now define

$$\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \eta_\sigma^n \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma \tag{4.5}$$

for $\sigma \in \mathscr{E}_m$, $1 \leqslant n \leqslant N_m$.

Lemma 4.2. *Up to a subsequence, the function p whose existence is guaranteed by Proposition 4.1 satisfies*

$$\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \eta(\mathcal{S}(p,x)) \qquad \text{in } L^q(Q_T),\ 1 \leqslant q < +\infty. \tag{4.6}$$

Proof. Because of (4.3a), $\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} = \eta(s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, x)$ converges almost everywhere to $\eta(\mathcal{S}(p,x),x)$. Since η is bounded, Lebesgue's dominated convergence theorem ensures that the convergence holds in $L^q(Q_T)$ for all $q \in [1,+\infty)$. The reconstruction $\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ of the mobility is also uniformly bounded, so we have just to show that $\|\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} - \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}\|_{L^1(Q_T)} \to 0$ as $m \to +\infty$. Letting $\Delta_{K\sigma} = K \cap \Delta_\sigma$ denote the half-diamond cell, we have

$$\begin{aligned}\|\eta_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} - \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}\|_{L^1(Q_T)} &\leqslant \sum_{n=1}^{N_m} \Delta t_m^n \sum_{K\in\mathscr{T}_m} \sum_{\sigma\in\mathscr{E}_K} m_{\Delta_{K\sigma}} |\eta_K(s_K^n) - \eta_\sigma^n| \\ &\leqslant \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)| \ \leqslant \ \sum_{i=1}^{I} \mathtt{R}_{i,m} + \mathtt{R}_{\Gamma,m},\end{aligned}$$

where

$$\mathtt{R}_{i,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)|,$$

$$\mathtt{R}_{\Gamma,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} m_{\Delta_\sigma} |\eta_K(s_K^n) - \eta_{K\sigma}(s_{K\sigma}^n)|.$$

Let us define

$$r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = |\eta_K^n - \eta_{K\sigma}^n| = r_\sigma^n \quad \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma,$$

then $r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ is uniformly bounded by $\|\eta\|_\infty = 1/\mu$. Therefore,

$$\mathtt{R}_{\Gamma,m} \leqslant \frac{T}{\mu} \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} m_{\Delta_\sigma} \leqslant \frac{2T\,\nu^{d-1}(\Gamma)}{\mu d}\, h_{\mathscr{T}_m}$$

where $h_{\mathscr{T}_m}$ is the size of $\mathscr{T}_m$ as defined in (2.8). Besides, for $i \in \{1,\dots,I\}$, $\eta_i \circ \mathcal{S}_i \circ \Theta_i^{-1}$ is continuous, monotone and bounded, hence uniformly continuous. This provides the existence of a modulus of continuity $\varpi_i : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varpi_i(0) = 0$ such that

$$r_\sigma^n := |\eta \circ \mathcal{S} \circ \Theta_i^{-1}(\Theta_K^n) - \eta \circ \mathcal{S} \circ \Theta_i^{-1}(\Theta_{K\sigma}^n)\,| \leqslant \varpi_i\left(|\Theta_K^n - \Theta_{K\sigma}^n|\right) \tag{4.7}$$

for $\sigma \in \mathscr{E}_{i,m}$. Therefore, if the function

$$q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} |\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n)| & \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{4.8a}$$

for $\sigma \in \mathscr{E}_{i,m}$, $1 \leqslant n \leqslant N_m$, could be proven to converge to 0 almost everywhere in $Q_{i,T}$, then it would also be the case for $r_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ and $\mathtt{R}_{i,m}$ as $m \to +\infty$, thanks to Lebesgue's dominated convergence theorem. Now, it follows from (3.12a) and from the elementary geometric relation

$$m_{\Delta_\sigma} = \frac{a_\sigma}{d} d_\sigma^2 \leqslant 4\frac{a_\sigma}{d} h_{\mathscr{T}_m}^2,$$

that

$$\|q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m}\|_{L^2(Q_{i,T})}^2 = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} |\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n)|^2 \leqslant \frac{4C_3}{d} h_{\mathscr{T}_m}^2.$$

Therefore, $q_{\mathscr{E}_{i,m},\boldsymbol{\Delta t}_m} \to 0$ in $L^2(Q_{i,T})$, thus also almost everywhere up to extraction of a subsequence. This provides the desired result. □

4.2. Identification of the limit

So far, we have exhibited some "limit" value p for the approximate solution $p_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}$ in Proposition 4.1. Next, we show that the scheme is consistent with the continuous problem by showing that any limit value is a weak solution.

Proposition 4.3. *The function p whose existence is guaranteed by Proposition 4.1 is a weak solution of the problem (1.9a)–(1.9c) in the sense of Definition 1.1.*

Proof. Let $\varphi \in C_c^\infty(\{\Omega \cup \Gamma^{\mathrm{N}}\} \times [0,T))$ and denote by $\varphi_K^n = \varphi(t_m^n, x_K)$, for all $K \in \mathscr{T}_m$ and all $n \in \{0,\dots,N_m\}$. We multiply (2.10) by $\Delta t_m^n \varphi_K^{n-1}$ and sum over $n \in \{1,\dots,N_m\}$ and $K \in \mathscr{T}_m$ to obtain

$$\mathtt{A}_m + \mathtt{B}_m = 0, \qquad m \geqslant 1, \tag{4.9}$$

where we have set

$$\mathtt{A}_m = \sum_{n=1}^{N_m} \sum_{K\in\mathscr{T}_m} m_K \phi_K \left(s_K^n - s_K^{n-1}\right) \varphi_K^{n-1}, \tag{4.10a}$$

$$\mathtt{B}_m = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_m} a_\sigma \lambda_\sigma \eta_\sigma^n \left(\vartheta_K^n - \vartheta_{K\sigma}^n\right)\left(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}\right). \tag{4.10b}$$

The quantity $\mathtt{A}_m$ in (4.10a) can be rewritten as

$$\begin{aligned}\mathtt{A}_m &= -\sum_{n=1}^{N_m}\Delta t_m^n\sum_{K\in\mathscr{T}_m} m_K\phi_K s_K^n\frac{\varphi_K^n-\varphi_K^{n-1}}{\Delta t_m^n} - \sum_{K\in\mathscr{T}_m} m_K\phi_K s_K^0\varphi_K^0\\ &= -\iint_{Q_T}\phi\, s_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}\delta\varphi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\, s^0_{\mathscr{T}_m}\varphi^0_{\mathscr{T}_m}\,\mathrm{d}x\end{aligned}$$

where

$$\begin{aligned}\delta\varphi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}(t,x) &= \frac{\varphi_K^n-\varphi_K^{n-1}}{\Delta t_m^n}, \qquad \text{if } (t,x)\in(t_m^{n-1},t_m^n)\times K,\\ \varphi^0_{\mathscr{T}_m} &= \varphi(0,x_K) \qquad\qquad \text{if } x\in K.\end{aligned}$$

Thanks to the regularity of φ, the function $\delta\varphi_{\mathscr{T}_m,\Delta t_m}$ converges uniformly to $\partial_t\varphi$ on $\Omega\times[0,T]$. Moreover, by virtue of (4.3a) and the boundedness of $s_{\mathscr{T}_m,\boldsymbol{\Delta t}_n}$ we can state that

$$\iint_{Q_T}\phi\, s_{\mathscr{T}_m,\Delta t_m}\delta\varphi_{\mathscr{T}_m,\Delta t_m}\,\mathrm{d}x\,\mathrm{d}t \underset{m\to+\infty}{\longrightarrow} \iint_{Q_T}\phi\,\mathcal{S}(p,x)\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t,$$

and, in view of the definition (2.13) of $s^0_{\mathscr{T}_m}$ and of the uniform convergence of $\varphi^0_{\mathscr{T}_m}$ towards $\varphi(0,\cdot)$,

$$\int_\Omega \phi_{\mathscr{T}_m} s^0_{\mathscr{T}_m}\varphi^0_{\mathscr{T}_m}\,\mathrm{d}x \underset{m\to+\infty}{\longrightarrow} \int_\Omega \phi\, s^0\varphi(0,\cdot)\,\mathrm{d}x.$$

From the above, we draw that

$$\lim_{m\to+\infty}\mathtt{A}_m = -\iint_{Q_T}\phi\,\mathcal{S}(p,x)\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega\phi\, s^0\varphi(0,\cdot)\,\mathrm{d}x. \tag{4.11}$$

Let us now turn our attention to the quantity $\mathtt{B}_m$ of (4.10b), which can be split into $\mathtt{B}_m=\mathtt{B}_m^1+\mathtt{B}_m^2$ using

$$\begin{aligned}\mathtt{B}_m^1 &= \sum_{n=1}^{N_m}\Delta t_m^n\sum_{\sigma\in\mathscr{E}_m} a_\sigma\lambda_\sigma\eta_\sigma^n\left(p_K^n-p_{K\sigma}^n\right)\left(\varphi_K^{n-1}-\varphi_{K\sigma}^{n-1}\right),\\ \mathtt{B}_m^2 &= \sum_{n=1}^{N_m}\Delta t_m^n\sum_{\sigma\in\mathscr{E}_m} a_\sigma\lambda_\sigma\eta_\sigma^n\left(\psi_K-\psi_{K\sigma}\right)\left(\varphi_K^{n-1}-\varphi_{K\sigma}^{n-1}\right).\end{aligned}$$

Consider first the convective term $\mathtt{B}_m^2$. It follows from the definition of the discrete gravitational potential

$$\psi_K = -\varrho g\cdot x_K, \quad \psi_\sigma = -\varrho g\cdot x_\sigma, \qquad K\in\mathscr{T}_m,\ \sigma\in\mathscr{E}_{\mathrm{ext},m}^{\mathrm{D}}$$

and from the orthogonality of the mesh that

$$\psi_K-\psi_{K\sigma} = d_\sigma\varrho g\cdot n_{K\sigma}, \qquad \forall\sigma\in\mathscr{E}_K\backslash\mathscr{E}_{\mathrm{ext}}^{\mathrm{N}},\ K\in\mathscr{T}_m.$$

Therefore, $\mathtt{B}_m^2$ can be transformed into

$$\begin{aligned}\mathtt{B}_m^2 &= \sum_{n=1}^{N_m}\Delta t_m^n\sum_{\sigma\in\mathscr{E}_m} m_{\Delta_\sigma}\lambda_\sigma\eta_\sigma^n d\frac{\varphi_K^{n-1}-\varphi_{K\sigma}^{n-1}}{d_\sigma}n_{K\sigma}\cdot\varrho g\\ &= -\iint_{Q_T}\lambda_{\mathscr{E}_m}\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}H_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}\cdot\varrho g\,\mathrm{d}x\,\mathrm{d}t,\end{aligned} \tag{4.12}$$

where

$$\lambda_{\mathscr{E}_m}(x) = \lambda_\sigma \qquad \text{if } x \in \Delta_\sigma,\ \sigma \in \mathscr{E}_m,$$
$$H_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = (d/d_\sigma)(\varphi_{K\sigma}^{n-1} - \varphi_K^{n-1})n_{K\sigma} \qquad \text{if } (t,x) \in [t_m^{n-1}, t_m^n) \times \Delta_\sigma.$$

After ([18], Lem. 4.4), $H_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ converges weakly in $L^2(Q_T)^d$ towards $\nabla\varphi$, while $\lambda_{\mathscr{E}_m}$ and $\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ converge strongly in $L^4(\Omega)$ and $L^4(Q_T)$ towards λ and $\eta(\mathcal{S}(p,x))$ respectively (*cf.*, Lem. 4.2). Thus, we can pass to the limit in (4.12) and

$$\lim_{m\to+\infty} \mathtt{B}_m^2 = -\iint_{Q_T} \lambda\eta(\mathcal{S}(p,x))\varrho g \cdot \nabla\varphi \,\mathrm{d}x\,\mathrm{d}t. \tag{4.13}$$

The capillary diffusion term $\mathtt{B}_m^1$ appears to be the most difficult one to deal with. Taking inspiration from [15], we introduce the auxiliary quantity

$$\begin{aligned}\widetilde{\mathtt{B}}_m^1 &= \sum_{i=1}^I \widetilde{\mathtt{B}}_{i,m}^1 \\ &= \sum_{i=1}^I \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma \sqrt{\lambda_i \eta_\sigma^n}\left(\Theta_i(p_K^n) - \Theta_i\left(p_{K\sigma}^n\right)\right)\left(\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}\right).\end{aligned}$$

Analogously to [24], we can define a piecewise-constant vector field $\overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ such that

$$\overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x)\cdot n_{K\sigma} = \varphi_{K\sigma}^{n-1} - \varphi_K^{n-1}, \qquad \text{if } (t,x) \in [t_m^{n-1}, t_m^n) \times \Delta_\sigma,\ \sigma \in \mathscr{E}_m,$$

and such that $\overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}$ converges uniformly towards $\nabla\varphi$ on $\overline{Q}_T$. Under these circumstances, $\widetilde{\mathtt{B}}_{i,m}^1$ reads

$$\widetilde{\mathtt{B}}_{i,m}^1 = \int_0^T \int_{\Omega_{i,m}} \sqrt{\lambda_i \eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}}\, G_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \cdot \overline{H}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \,\mathrm{d}x\,\mathrm{d}t$$

where $\Omega_{i,m} = \bigcup_{\sigma\in\mathscr{E}_{i,m}} \Delta_\sigma \subset \Omega_i$. The strong convergence of $\sqrt{\eta_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}}$ in $L^2(Q_{i,T})$ towards $\sqrt{\eta_i(\mathcal{S}_i(p))}$ directly follows from the boundedness of η_i combined with (4.3a). Combining this with (4.3b) results in

$$\widetilde{\mathtt{B}}_{i,m}^1 \underset{m\to+\infty}{\longrightarrow} \iint_{Q_{i,T}} \sqrt{\lambda_i \eta_i(\mathcal{S}_i(p))}\nabla\Theta_i(p)\cdot\nabla\varphi \,\mathrm{d}x\,\mathrm{d}t = \iint_{Q_{i,T}} \nabla\Phi_i(p)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t. \tag{4.14}$$

Therefore, to finish the proof of Proposition 4.3, it only remains to check that B_m^1 and $\widetilde{B}_m^1$ share the same limit. To this end, we observe that, by the triangle inequality, we have

$$|\mathtt{B}_m^1 - \widetilde{\mathtt{B}}_m^1| \leqslant \mathtt{R}_{\Gamma,m} + \sum_{i=1}^I \mathtt{R}_{i,m}, \tag{4.15}$$

where

$$\mathtt{R}_{\Gamma,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} a_\sigma \lambda_\sigma \eta_\sigma^n |p_K^n - p_{K\sigma}^n||\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}|,$$
$$\mathtt{R}_{i,m} = \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma \sqrt{\lambda_i\eta_\sigma^n}|\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n) - \sqrt{\lambda_i\eta_\sigma^n}(p_K^n - p_{K\sigma}^n)||\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}|.$$

Applying the Cauchy–Schwarz inequality and using Proposition 3.1, we find

$$|\mathtt{R}_{\Gamma,m}|^2 \leqslant C_1 \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{\Gamma,m}} a_\sigma \lambda_\sigma \eta_\sigma^n |\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}|^2 \leqslant 2C_1 T \|\nabla\varphi\|_\infty^2 \frac{\max_i \lambda_i}{\mu} \nu^{d-1}(\Gamma) h_{\mathscr{T}_m},$$

so $\mathtt{R}_{\Gamma,m} \to 0$ as $m \to +\infty$. Besides, we also apply the Cauchy–Schwarz inequality to $\mathtt{R}_{i,m}$ in order to obtain

$$|\mathtt{R}_{i,m}|^2 \leqslant C_1 \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} a_\sigma\, \lambda_i \big|\sqrt{\eta_\sigma^n} - \sqrt{\widetilde{\eta}_\sigma^n}\,\big|^2 \left|\varphi_K^{n-1} - \varphi_{K\sigma}^{n-1}\right|^2$$
$$\leqslant d\lambda_i C_1 \|\nabla\varphi\|_\infty^2 \sum_{n=1}^{N_m} \Delta t_m^n \sum_{\sigma\in\mathscr{E}_{i,m}} m_{\Delta_\sigma} \left|\eta_\sigma^n - \widetilde{\eta}_\sigma^n\right|,$$

where we have set

$$\widetilde{\eta}_\sigma^n = \begin{cases} \eta_i(s_K^n) & \text{if } p_K^n = p_{K\sigma}^n, \\ \left[\dfrac{\Theta_i(p_K^n) - \Theta_i(p_{K\sigma}^n)}{\sqrt{\lambda_i}(p_K^n - p_{K\sigma}^n)}\right]^2 & \text{otherwise.} \end{cases}$$

Define

$$\widetilde{\eta}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m}(t,x) = \begin{cases} \widetilde{\eta}_\sigma^n & \text{if } (t,x) \in (t_m^{n-1}, t_m^n] \times \Delta_\sigma,\ \sigma \in \bigcup_{i=1}^I \mathscr{E}_{i,m}, \\ 0 & \text{otherwise.} \end{cases}$$

Reproducing the proof of Lemma 4.2, we can show that

$$\widetilde{\eta}_{\mathscr{E}_m,\boldsymbol{\Delta t}_m} \underset{m\to\infty}{\longrightarrow} \eta(\mathcal{S}(p,x)) \qquad \text{in } L^q(Q_T),\ 1 \leqslant q < +\infty.$$

Therefore, $\mathtt{R}_{i,m} \to 0$ as $m \to +\infty$. Putting things together in (4.15), we conclude that $\mathtt{B}_m^1$ and $\widetilde{\mathtt{B}}_m^1$ share the same limit, which completes the proof of Proposition 4.3. □

5. Practical aspects of numerical resolution

We provide some details on the resolution strategy for the discrete problem (2.9)–(2.11c). Different techniques have been proposed in the literature to numerically solve the Richards equation (*e.g.*, see [40,43]). Our strategy is based on a parametrization technique to automatically choose the most convenient variable during the Newton iterations (Sect. 5.1) to enhance Newton's convergence and on the addition of cells on the interfaces between different rock types (Sect. 5.2) to improve the pressure continuity.

5.1. Switch of variable and parametrization technique

A natural choice to solve the nonlinear system (2.9)–(2.11c) is to select the pressure $(p_K)_{K\in\mathscr{T}}$ as primary unknown and to solve it *via* an iterative method such as Newton's one. Nevertheless, the pressure variable is known to be an inefficient choice for $s \ll 1$ because of the degeneracy of Richards' equation. For dry soils, this strategy is outperformed by schemes in which saturation is the primary variable. On the other hand, the knowledge of the saturation is not sufficient to describe the pressure curve in saturated regions where the pressure-saturation relation cannot be inverted. This motivated the design of schemes involving a switch of variable [23,34]. In this work, we adopt the technique proposed in [7], in which a third generic variable τ is introduced to become the primary unknown of the system. Then the idea is to choose a parametrization of the graph $\{p, \mathcal{S}(p)\}$, *i.e.*, to construct two functions $\mathfrak{s} : I \to [s_{\rm rw}, 1 - s_{\rm rn}]$ and $\mathfrak{p} : I \to \mathbb{R}$ such that $\mathfrak{s}(\tau) = \mathcal{S}(\mathfrak{p}(\tau))$ and $\mathfrak{s}'(\tau) + \mathfrak{p}'(\tau) > 0$ for all $\tau \in I \subset \mathbb{R}$. Such a parametrization is not unique, for instance one can take $I = \mathbb{R}$, $\mathfrak{p} = Id$ which amounts to solving the system always in pressure, but this is not recommended as explained before. Here, we set $I = (s_{\rm rw}, +\infty)$ and

$$\mathfrak{s}(\tau) = \begin{cases} \tau & \text{if } \tau \leqslant s_{\rm s}, \\ \mathcal{S}\left(p_{\rm s} + \dfrac{\tau - s_{\rm s}}{\mathcal{S}'(p_{\rm s}^-)}\right) & \text{if } \tau \geqslant s_{\rm s}, \end{cases} \qquad \mathfrak{p}(\tau) = \begin{cases} \mathcal{S}^{-1}(\tau) & \text{if } \tau \leqslant s_{\rm s}, \\ p_{\rm s} + \dfrac{\tau - s_{\rm s}}{\mathcal{S}'(p_{\rm s}^-)} & \text{if } \tau \geqslant s_{\rm s}, \end{cases}$$

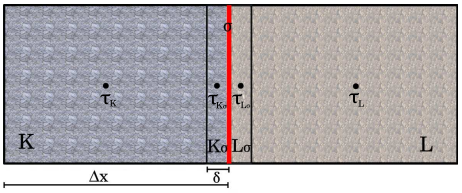

FIGURE 1. Mesh refinement on both sides of an interface face for a 2D case.

where $\mathcal{S}'(p_{\mathrm{s}}^-)$ denotes the limit as p tends to $p_{\mathrm{s}} = \mathcal{S}(s_{\mathrm{s}})$ from below of $\mathcal{S}'(p)$. Since the switch point s_{s} is taken as the inflexion point of $\mathcal{S}$, both $\mathfrak{s}$ and $\mathfrak{p}$ are C^1 and concave, and even C^2 if $\mathcal{S}$ is given by the van Genuchten–Mualem model. Moreover, for all $p \in \mathbb{R}$, there exists a unique $\tau \in (s_{\mathrm{rw}}, +\infty)$ such that $(p, \mathcal{S}(p)) = (\mathfrak{p}(\tau), \mathfrak{s}(\tau))$. The resulting system $\mathcal{F}_n(\boldsymbol{\tau}^n) = \mathbf{0}$ made up of $N_{\mathscr{T}} = \mathrm{Card}(\mathscr{T})$ nonlinear equations admits a unique solution $\boldsymbol{\tau}^n$, since it is fully equivalent to (2.9)–(2.11c). The stopping criterion for Newton iterations is based on the L^∞-norm of the residual with a convergence threshold fixed to $\epsilon = 10^{-12}$. A direct linear solver based on the LU factorization has been used. More details about the practical resolution of this nonlinear system *via* the Newton method can be found in [5].

5.2. Pressure continuity at rock type interfaces

Physically, the pressure should remain continuous on both sides of an interface between two different rock types. But this continuity is here not imposed at the discrete level. The two-point flux approximation based on the cell unknowns is strongly dependent on the mesh resolution and can induce a large error close to the rock type interface. We here propose a very simple method to improve this continuity condition in pressure. It consists in adding two thin cells of resolution δ on each side of the rock-type interface with $\delta \ll \Delta x$ as shown in Figure 1.

The idea is here to add two cells unknowns in the neighborhood of the interface to have a more precise approximation of the pressure gradient on each side of the faces where changes of rock types occur. In this way, we avoid the introduction of face unknowns in our solver which remains unchanged. For these interface cells, tangential fluxes are neglected.

Readers who are interested in a more advanced discussion on numerical strategies (among which the one briefly described in the current section and referred as Method B in what follows) to solve the transmission problem (1.10) shall refer to [6]. The study presented therein in particular covers the robustness of the nonlinear solvers.

6. NUMERICAL RESULTS

In this section, we present the results obtained for different test cases. For all these cases, we consider a two-dimensional layered domain $\Omega = [0\,\mathrm{m}, 5\,\mathrm{m}] \times [-3\,\mathrm{m}, 0\,\mathrm{m}]$ made up of two rock types denoted by RT0 and RT1 respectively, RT0 being less permeable than RT1. Using these two lithologies, the domain Ω is partitioned into three connected subdomains: $\Omega_1 = [1\,\mathrm{m}, 4\,\mathrm{m}] \times [-1\,\mathrm{m}, 0\,\mathrm{m}]$, $\Omega_2 = [0\,\mathrm{m}, 5\,\mathrm{m}] \times [-3\,\mathrm{m}, -2\,\mathrm{m}]$ and $\Omega_3 = \Omega \setminus (\Omega_1 \cup \Omega_2)$, as depicted in Figure 2.

The Brooks–Corey [12] and van Genuchten–Mualem [47] petro-physical models are used to model the flow characteristics of both rock types. In these models, the water saturation and the water pressure are linked pointwise by the relation $s = \mathcal{S}(p)$ where $\mathcal{S} : \mathbb{R} \to [0, 1]$ is nondecreasing and satisfies $\mathcal{S}(p) = 1 - s_{\mathrm{rn}}$ if $p \geqslant p_b$ and $\mathcal{S}(p) \to s_{\mathrm{rw}}$ as $p \to -\infty$, s_{rw} being the residual wetting saturation, s_{rn} the residual non-wetting saturation and p_b the entry pressure. More precisely, we have,

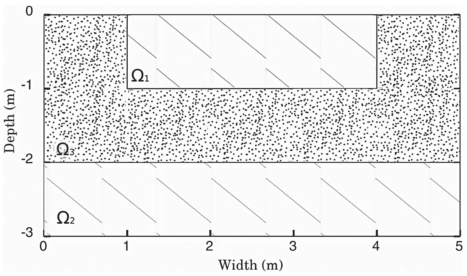

FIGURE 2. Simulation domain $\Omega = [0\,\mathrm{m}, 5\,\mathrm{m}] \times [-3\,\mathrm{m}, 0\,\mathrm{m}]$.

TABLE 1. Parameters used for the Brooks–Corey model.

	$1 - s_{\mathrm{rn}}$	s_{rw}	p_b[Pa]	n	λ[m^2]	ϕ
RT0	1.0	0.1	-1.4708×10^3	3.0	10^{-11}	0.35
RT1	1.0	0.2	-3.4301×10^3	1.5	10^{-13}	0.35

TABLE 2. Parameters used for the van Genuchten–Mualem model.

	$1 - s_{\mathrm{rn}}$	s_{rw}	n	λ [m^2]	α [m^{-1}]	ϕ
RT0 (Sand)	1.0	0.0782	2.239	6.3812×10^{-12}	2.8	0.3658
RT1 (Clay)	1.0	0.2262	1.3954	1.5461×10^{-13}	1.04	0.4686

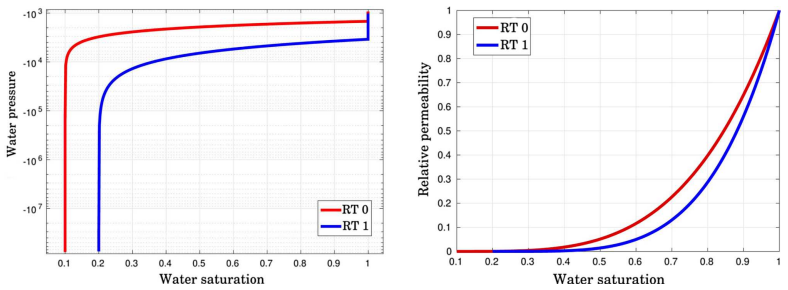

FIGURE 3. Water pressure and relative permeability curves for the Brooks–Corey model.

– for the Brooks–Corey model,

$$s = \mathcal{S}(p) = \begin{cases} s_{\mathrm{rw}} + (1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}) \left(\frac{p}{p_b}\right)^{-n} & \text{if } p \leqslant p_b, \\ 1 - s_{\mathrm{rn}} & \text{if } p > p_b, \end{cases}$$

$$k_r(s) = s_{\mathrm{eff}}^{3+\frac{2}{n}}, \qquad s_{\mathrm{eff}} = \frac{s - s_{\mathrm{rw}}}{1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}};$$

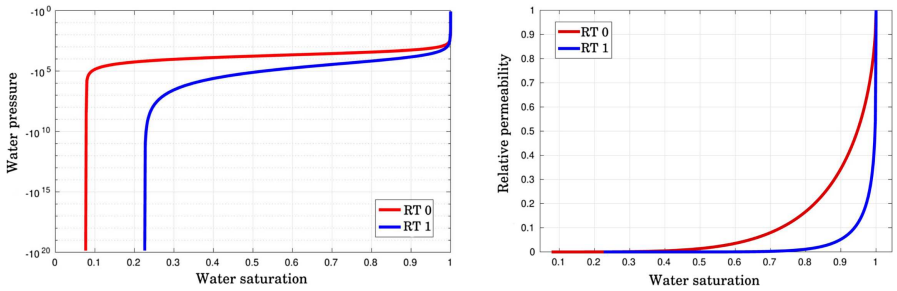

FIGURE 4. Water pressure and relative permeability curves for the van Genuchten–Mualem model.

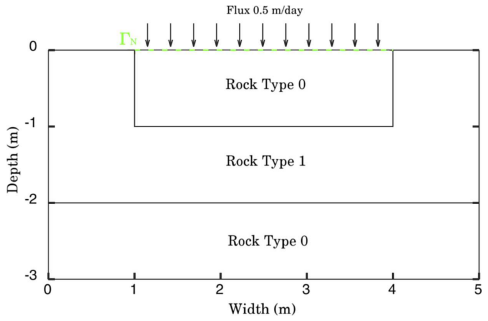

FIGURE 5. Boundary condition for the filling case.

– for the van Genuchten–Mualem model,

$$s = \mathcal{S}(p) = \begin{cases} s_{\mathrm{rw}} + (1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}) \left[1 + \left|\frac{\alpha}{\rho g} p\right|^n\right]^{-m} & \text{if } p \leqslant 0, \\ 1 - s_{\mathrm{rn}} & \text{if } p > 0, \end{cases}$$

$$k_r(s) = s_{\mathrm{eff}}^{\frac{1}{2}} \left\{1 - \left[1 - s_{\mathrm{eff}}^{\frac{1}{m}}\right]^m\right\}^2, \qquad s_{\mathrm{eff}} = \frac{s - s_{\mathrm{rw}}}{1 - s_{\mathrm{rn}} - s_{\mathrm{rw}}}, \qquad m = 1 - \frac{1}{n};$$

where $\eta(\cdot) = k_r(\cdot)/\mu$, $\mu = 10^{-3}$ Pa·s being water viscosity, is the relative permeability. The parameters used for both rock types are given in Table 1 for the Brooks–Corey model and in Table 2 for the van Genuchten–Mualem model. With these choices of parameters, water is more likely to be in RT1 than in RT0, in the sense tha, at a given pressure, the water saturation is higher in RT1 than in RT0, as it can be seen on the plots of the capillary-pressure functions depicted in Figures 3 and 4 for these two petro-physical models. Figures 3 and 4 also show the relative permeability functions. Note the non-Lipschitz character of the relative permeability in the van Genuchten–Mualem framework.For the numerical tests, in order to avoid infinite values for the derivative of $k_r(s)$ when $s \to 1 - s_{\mathrm{rn}}$, we approximate it for $s \in [s_{\mathrm{lim}}, 1 - s_{\mathrm{rn}}]$ using a second degree polynomial $\widetilde{k_r}(s)$. Such a polynomial satisfies the following constraints: $k_r(s_{\mathrm{lim}}) = \widetilde{k_r}(s_{\mathrm{lim}})$ and $\widetilde{k_r}(1 - s_{\mathrm{rn}}) = 1$. The value s_{lim} corresponds to $s_{\mathrm{eff}} = 0.998$.

6.1. Configurations of the test cases

For both petro-physical models, we consider two configurations further referred as filling and drainage cases, which are described in the following.

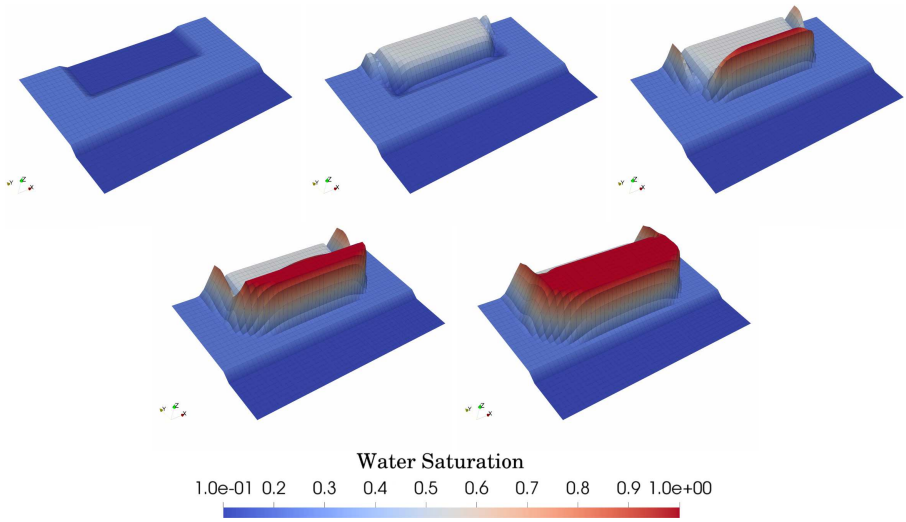

FIGURE 6. Evolution of the saturation profile for $t \in \{0\,\mathrm{s}, 20 \times 10^3\,\mathrm{s}, 40 \times 10^3\,\mathrm{s}, 60 \times 10^3\,\mathrm{s}, 86 \times 10^3\,\mathrm{s}\}$ for filling case, using the Brooks–Corey model, method B and the 50×30 cells mesh.

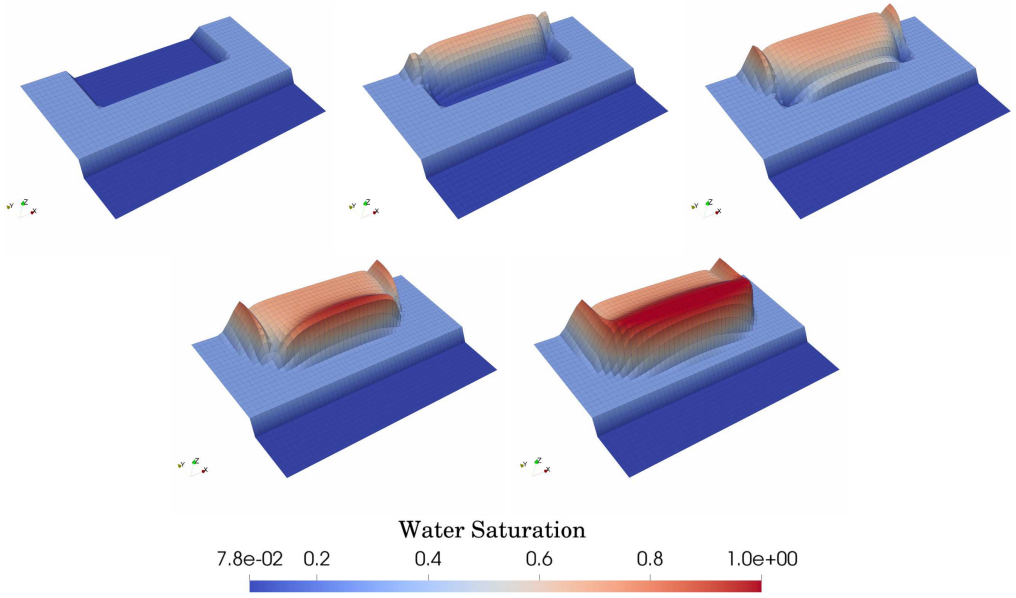

FIGURE 7. Evolution of the saturation profile for $t \in \{0\,\mathrm{s}, 20 \times 10^3\,\mathrm{s}, 40 \times 10^3\,\mathrm{s}, 60 \times 10^3\,\mathrm{s}, 86 \times 10^3\,\mathrm{s}\}$ for filling case using the van Genuchten–Mualem model, method B and the 50×30 cells mesh.

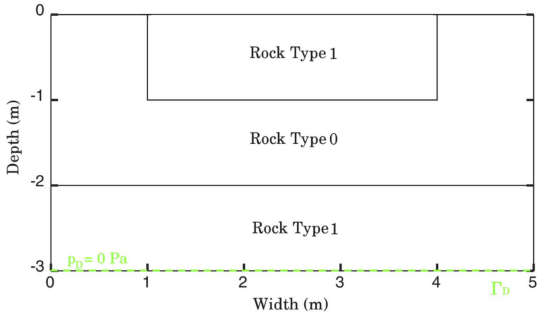

FIGURE 8. Boundary condition for the drainage test.

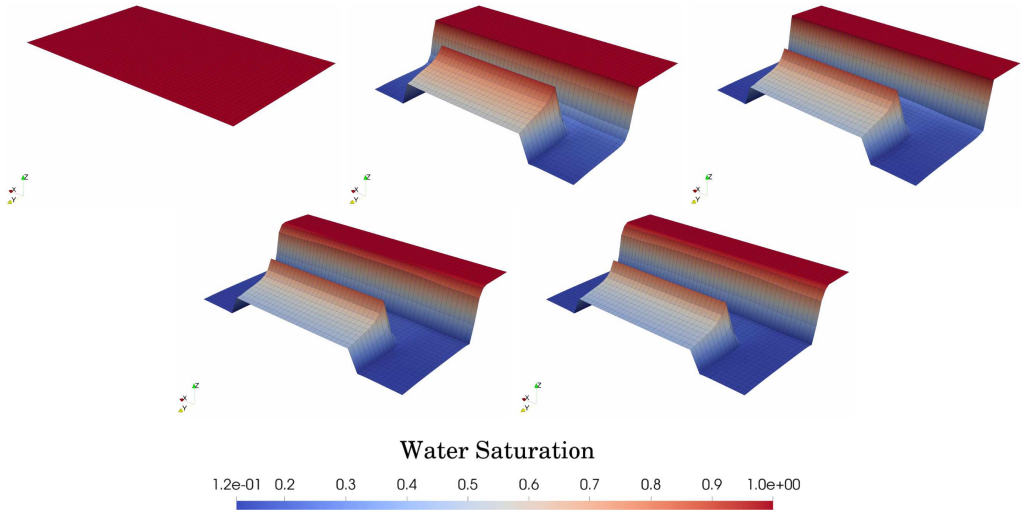

FIGURE 9. Evolution of the saturation profile for $t \in \{0\,\text{s}, 26.2 \times 10^4\,\text{s}, 52.4 \times 10^4\,\text{s}, 78.6 \times 10^4\,\text{s}, 105 \times 10^4\,\text{s}\}$ for drying case, using the Brooks–Corey model, method B and the 50×30 cells mesh.

6.1.1. Filling case

The filling test case has already been considered in [17, 34, 38, 42]. Starting from an initially dry domain Ω, whose layers' composition is reported in Figure 5, water flows from a part of the top boundary during the entire simulation time that is equal to one day. A no-flow boundary condition is applied elsewhere. More precisely, the initial capillary pressure is set to -47.088×10^5 Pa and the water flux rate to 0.5 m/day through $\Gamma_N = \{(x, y) \,|\, x \in [1\,\text{m}, 4\,\text{m}], y = 0\,\text{m}\}$. For this simulation a uniform time-step $\Delta t = 1000\,\text{s}$ is prescribed for the test using the Brooks–Corey model and $\Delta t = 500\,\text{s}$ for the one using the van Genuchten–Mualem model.

The test case follows the following dynamics. Water starts invading the void porous space in Ω_1. When it reaches the interface with Ω_3, capillarity involves a suction force on water from Ω_1 to Ω_3. Since clay (RT1) has low permeability, water encounters difficulties to progress within Ω_3. This yields a front moving downward in Ω_1 which is stiffer for the Brooks–Corey model than for the van Genuchten–Mualem one. In both cases, the simulation is stopped before water reaches the bottom part corresponding to Ω_2. In Figure 6 we can observe the evolution of the saturation profile during the simulation performed on a 50×30 cells mesh with the Brooks–Corey model, whereas the evolution corresponding to van Genuchten–Mualem nonlinearities is depicted in Figure 7.

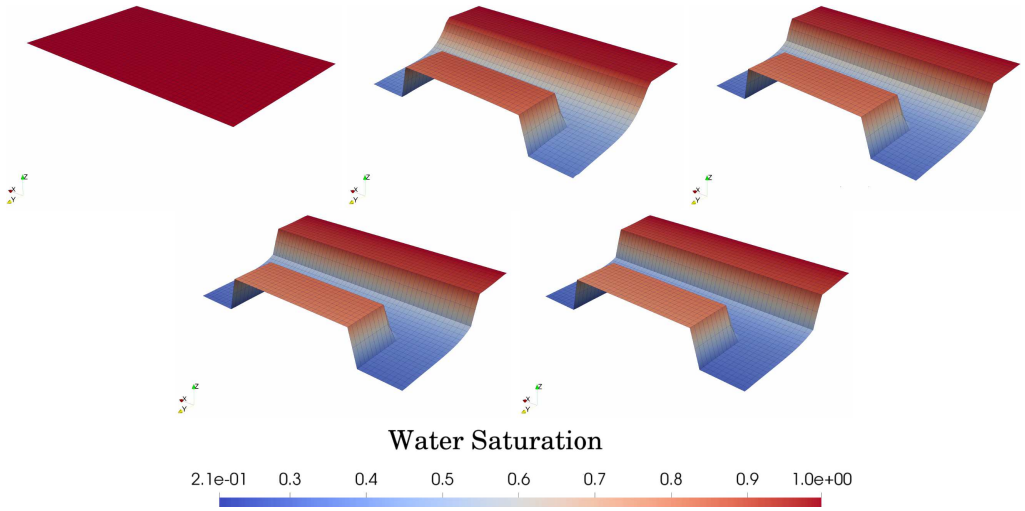

FIGURE 10. Evolution of the saturation profile for $t \in \{0\,\mathrm{s}, 26.16 \times 10^4\,\mathrm{s}, 52.4 \times 10^4\,\mathrm{s}, 78.56 \times 10^4\,\mathrm{s}, 105 \times 10^4\,\mathrm{s}\}$ for drying case using the van Genuchten–Mualem model, method B and the 50×30 cells mesh.

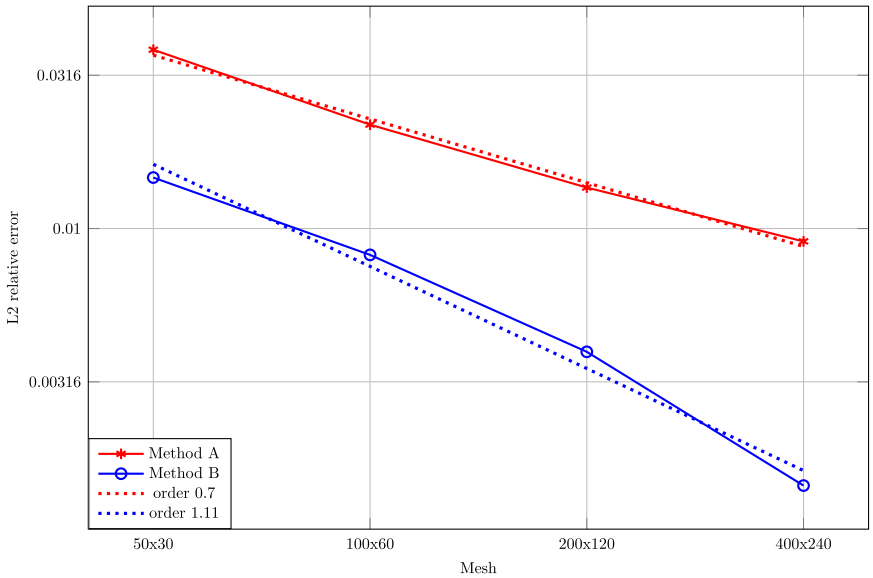

FIGURE 11. $L^2(Q_T)$ relative error in saturation for the drainage case using the Brooks–Corey model.

6.1.2. Drainage case

This test case is designed as a two-dimensional extension of a one-dimensional test case proposed by [41] and addressed in [17, 42]. We simulate a vertical drainage starting from initially and boundary saturated conditions during 105×10^4 s. At the initial time, the pressure varies with depth with $p^0(z) = -\rho g z$. A Dirichlet boundary condition $p_D = 0$ Pa is imposed on the bottom of the domain, more precisely on $\Gamma_D = \{(x, y) \,|\, x \in [0\,\mathrm{m}, 5\,\mathrm{m}], y =$

TABLE 3. Newton's iterations for the mesh 200 × 120 for the drainage case using the Brooks–Corey model.

	♯ total	♯ avg	♯ max
Method A	1927	3	29
Method B	2038	3	29

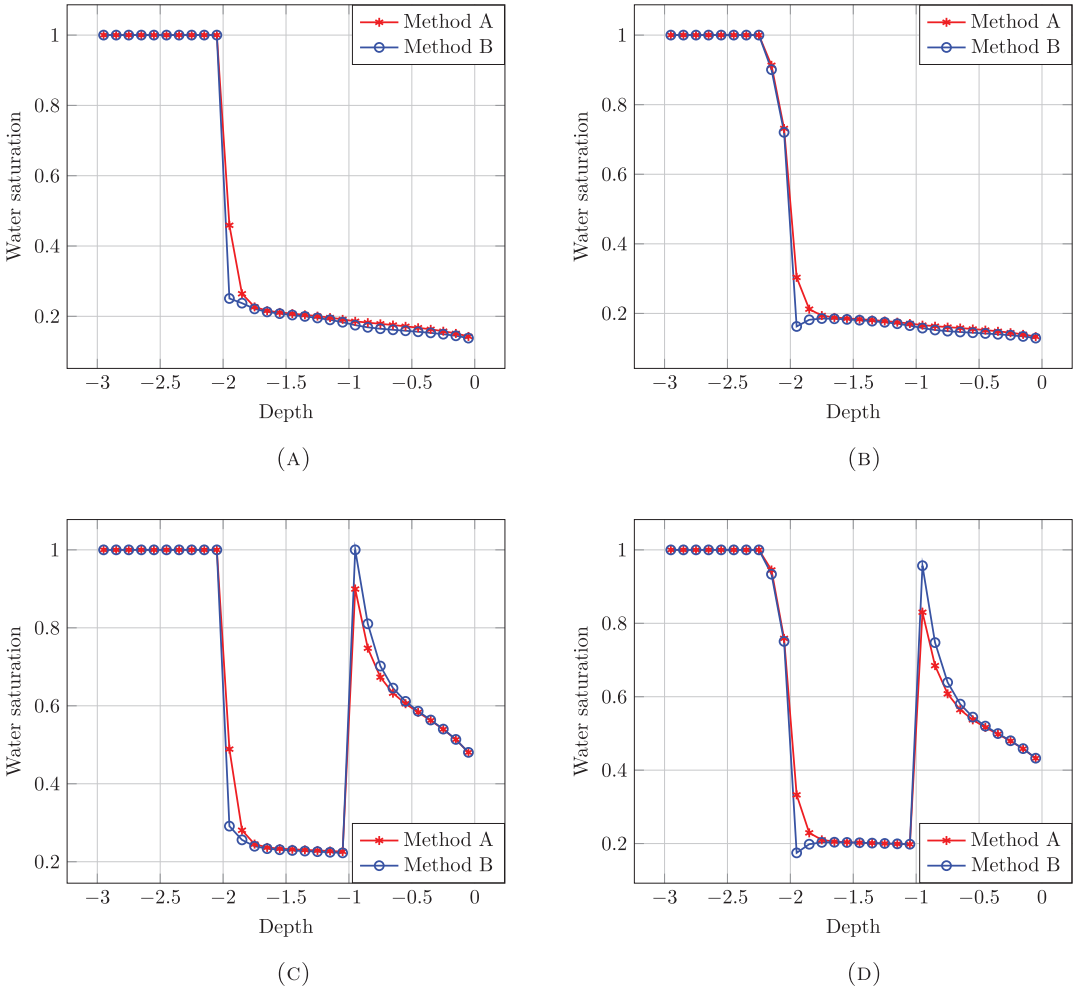

FIGURE 12. Water saturation profile obtained in the drainage test case with the Brooks–Corey model using methods A and B along vertical cross-sections at different times. (A) Cross-section at $x = 0.85\,\text{m}$, $t = 53.2\times10^4\,\text{s}$. (B) Cross-section at $x = 0.85\,\text{m}$, $t = 105\times10^4\,\text{s}$. (C) Cross-section at $x = 2.55\,\text{m}$, $t = 53.2\times10^4\,\text{s}$. (D) Cross-section at $x = 2.55\,\text{m}$, $t = 105\times10^4\,\text{s}$.

$-3\,\text{m}\}$. The layers' composition of Ω is reported in Figure 8. For this simulation a uniform time-step $\Delta t = 2000\,\text{s}$ is used for the test with the Brooks–Corey model and $\Delta t = 800\,\text{s}$ for the one with the van Genuchten–Mualem model.

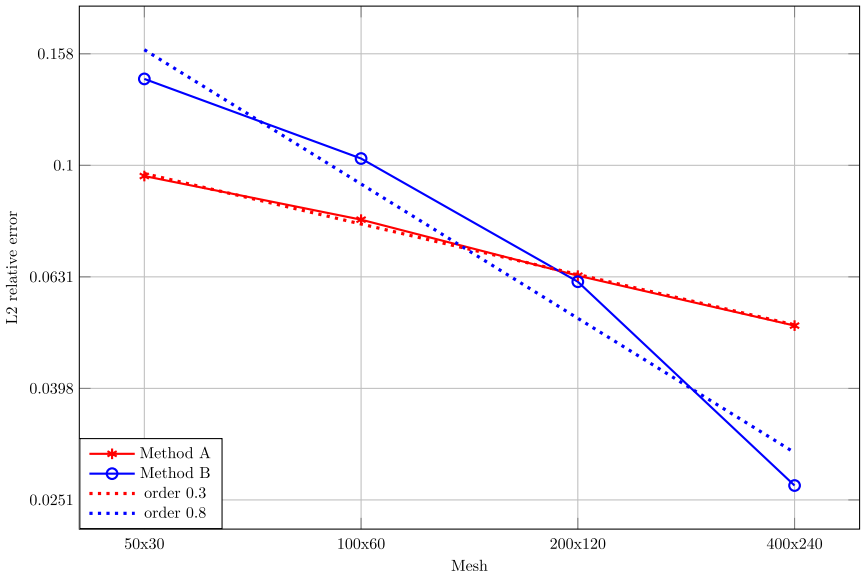

FIGURE 13. $L^2(Q_T)$ relative error in saturation for the filling case using the Brooks–Corey model.

TABLE 4. Newton's iterations for the mesh 200×120 for the filling case using the Brooks–Corey model.

	♯ total	♯ avg	♯ max
Method A	659	7	31
Method B	788	9	32

At the top interface between Ω_1 and Ω_3, capillarity acts in opposition to gravity and to the evolution of the system into a dryer configuration. The interface between Ω_2 and Ω_3 acts in the reverse way: suction accelerates the gravity driven drainage of RT0.

In Figure 9 we can observe the evolution of the saturation profile during the simulation performed on a 50×30 cells mesh with the Brooks–Corey model, whereas the evolution corresponding to van Genuchten–Mualem nonlinearities is depicted in Figure 10.

6.2. Comparisons of the numerical treatments of the interfaces

For each petro-physical model and configuration, a numerical convergence analysis is carried out for the schemes with (method B) or without (method A) thin cells, whose thickness is fixed to $\delta = 10^{-6}$ m, at rock type interfaces. Five structured meshes with the following resolutions are considered for this analysis: 50×30, 100×60, 200×120, 400×240, 800×480. The evolution of the error is measured using the $L^2(Q_T)$-norm of the relative difference between the saturations obtained on a given mesh and the ones computed with Method A and a mesh of resolution 800×480. The number of Newton iterations obtained with both methods is also compared.

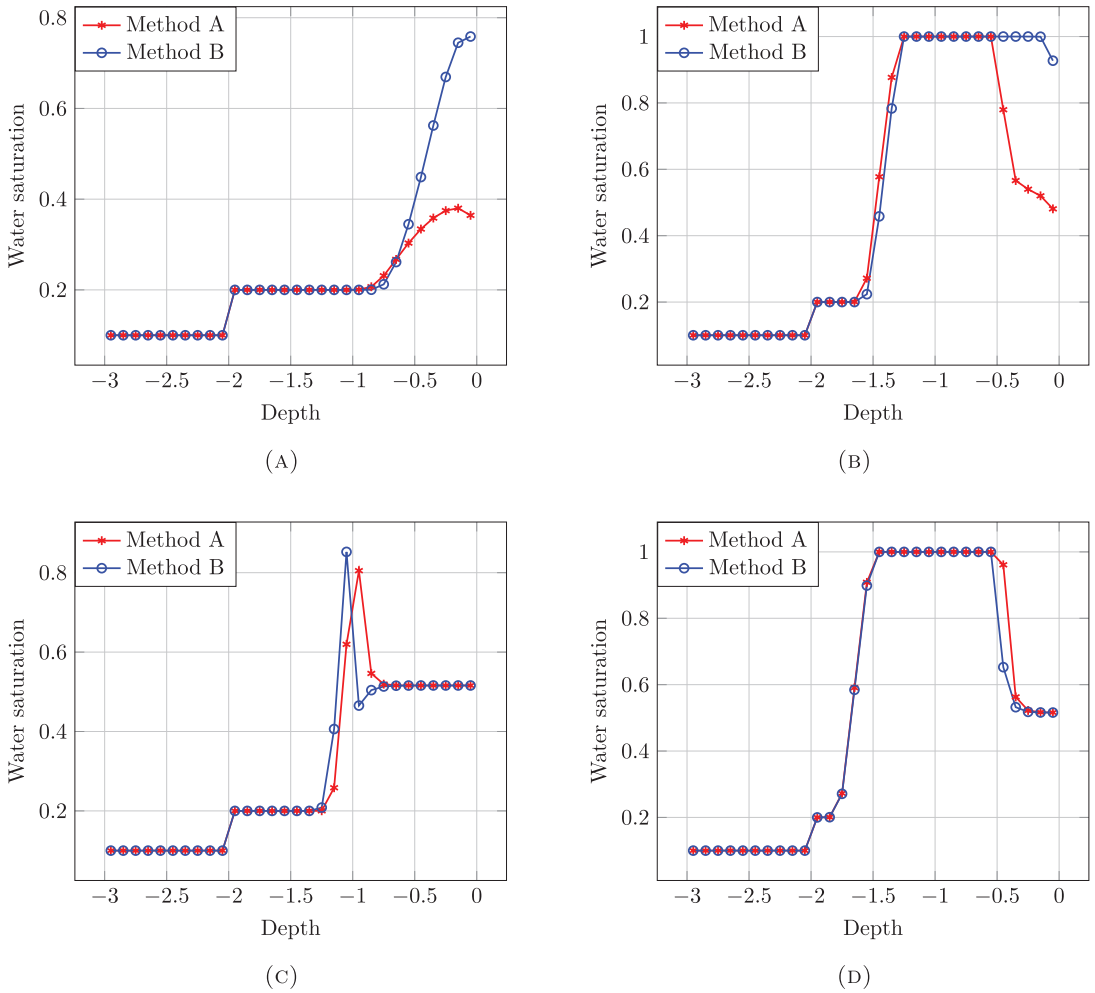

FIGURE 14. Water saturation profile obtained in the filling test case with the Brooks–Corey model using Method A and B along vertical cross-sections at different times. (A) Cross-section at $x = 0.95\,\mathrm{m}$, $t = 30 \times 10^3\,\mathrm{s}$. (B) Cross-section at $x = 0.95\,\mathrm{m}$, $t = 86.4 \times 10^3\,\mathrm{s}$. (C) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 30 \times 10^3\,\mathrm{s}$. (D) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 86.4 \times 10^3\,\mathrm{s}$.

6.2.1. Brooks–Corey model: drainage case

For the drainage case with the Brooks–Corey model, the convergence error is given in Figure 11. First we notice that, for all meshes, the error is smaller with method B than with method A and that we have a linear rate of convergence with the first one whereas this rate is smaller with the latter one. The total, average and maximal number of Newton iterations are also given in Table 3. Method B appears to be slightly more expensive.

Let us now look into the results obtained with Method A and Method B. In Figure 12 we plot the saturation profile at $x \in \{0.85\,\mathrm{m}, 2.55\,\mathrm{m}\}$ (see Fig. 2) for two different times, namely when the cells line in Ω_2 below its interface with Ω_3 starts drying and at the final time.

6.2.2. Brooks–Corey model: filling case

For the filling case with the Brooks–Corey model, the convergence error is given in Figure 13. As for the previous case, Method B enables to recover a linear convergence rate. Except for the first two meshes where

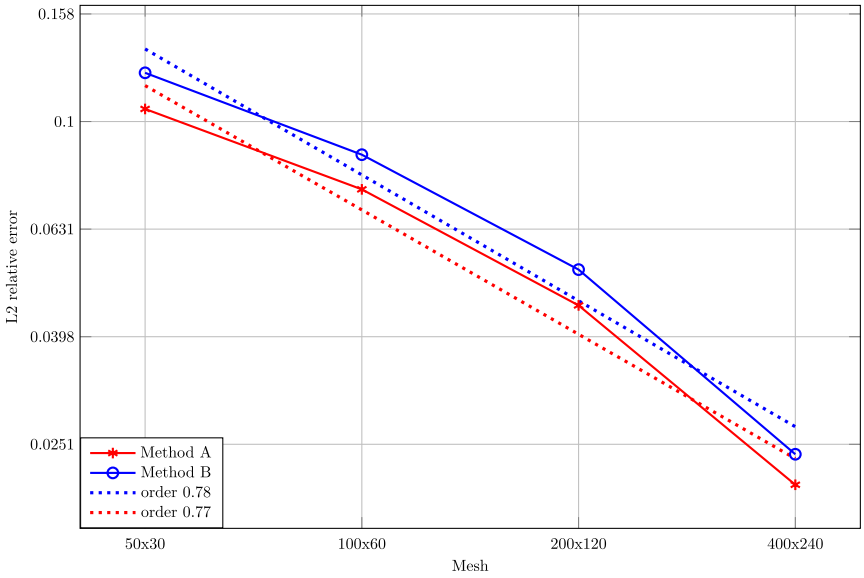

FIGURE 15. $L^2(Q_T)$ relative error in saturation for the filling case using the van Genuchten–Mualem model.

TABLE 5. Newton's iterations for the mesh 200×120 for the filling case using the van Genuchten–Mualem model.

	♯ total	♯ avg	♯ max
Method A	782	4	15
Method B	959	5	15

the error obtained with Method B is slightly larger, for all other meshes, this error is smaller than the one obtained with method A. The total, average and maximal number of Newton iterations are given in Table 4. The algorithm behaves here in the same way as before.

Let us now look into the results obtained with Method A and Method B. In Figure 14 we plot the saturation profile at $x \in \{0.95\,\mathrm{m}, 2.55\,\mathrm{m}\}$ (see Fig. 2) for two different times: when cells around the interface between Ω_1 and Ω_3 are almost saturated and at the final time.

6.2.3. Van Genuchten–Mualem model: filling case

For the filling case with the van Genuchten–Mualem model, the convergence error is given in Figure 15. Both methods exhibit a linear rate of convergence. On the other hand, the error is slightly larger with method B than with method A. The total, average and maximal number of Newton iterations are given in Table 5.

Figure 16 shows the localization of the differences between the numerical solutions provided by methods A and B. In the picture we report the vertical section of the saturation solution at $x \in \{0.65\,\mathrm{m}, 0.85\,\mathrm{m}, 2.55\,\mathrm{m}\}$ (see Fig. 2) for two different times: when the line of cells below and above the interface between Ω_1 and Ω_3 are almost saturated and at the final time. Unsurprisingly, the difference is located in the neighborhood of the interfaces. Moreover, as suggested by Figures 13 and 15, the influence of the introduction of additional interface unknowns (method B) has a lower impact for van Genuchten–Mualem nonlinearities than for Brook–Corey nonlinearities.

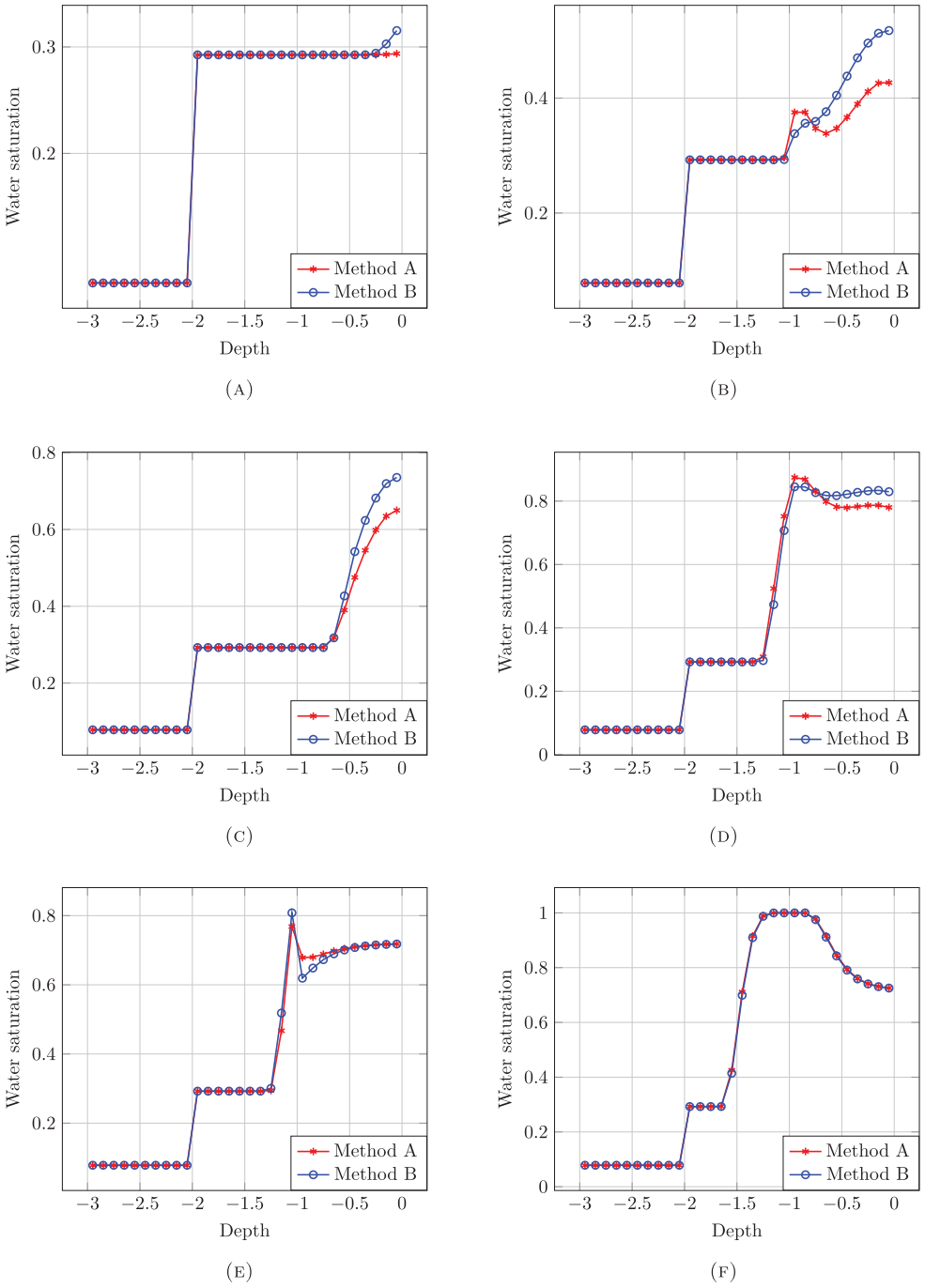

FIGURE 16. Water saturation profile obtained in the filling test case with the van Genuchten Mualem model using Method A and B along vertical cross-sections at different times. (A) Cross-section at $x = 0.65\,\mathrm{m}$, $t = 45 \times 10^3\,\mathrm{s}$. (B) Cross-section at $x = 0.65\,\mathrm{m}$, $t = 86.4 \times 10^3\,\mathrm{s}$. (C) Cross-section at $x = 0.85\,\mathrm{m}$, $t = 45 \times 10^3\,\mathrm{s}$. (D) Cross-section at $x = 0.85\,\mathrm{m}$, $t = 86.4 \times 10^3\,\mathrm{s}$. (E) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 45 \times 10^3\,\mathrm{s}$. (F) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 86.4 \times 10^3\,\mathrm{s}$.

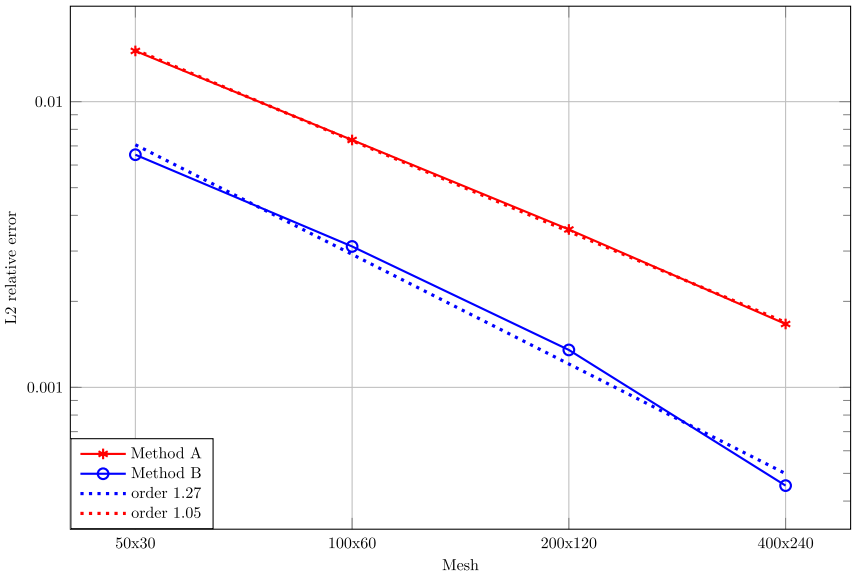

FIGURE 17. $L^2(Q_T)$ relative error in saturation for the drainage case using the van Genuchten–Mualem model.

TABLE 6. Newton's iterations for the mesh 200×120 for the filling case using the van Genuchten–Mualem model.

	♯ total	♯ avg	♯ max
Method A	2845	2	29
Method B	3523	2	20

6.2.4. Van Genuchten–Mualem model: drainage case

For the drainage case with the van Genuchten–Mualem model, the convergence error is given in Figure 17. Both methods exhibit a linear rate of convergence. Moreover, the error is slightly larger with method A than with method B. The total, average and maximal number of Newton iterations are given in Table 6.

Let us now look into the results obtained with method A and method B. In Figure 18 we plot the saturation profile at $x \in \{0.95\,\mathrm{m}, 2.55\,\mathrm{m}\}$ (see Fig. 2) for two different times: when the cell line in Ω_2 below its interface with Ω_3 significantly starts drying and at the final time.

6.2.5. Influence of the parameter δ

Let us now analyze how the thickness of the thin cells employed in method B affects the accuracy of the solution obtained with this method. We consider the filling and drainage cases along with the Brooks–Corey model and evaluate the relative $L^2(Q_T)$ error between the solution obtained on the 200×120 cells mesh using $\delta \in \{10^{-2}\,\mathrm{m}, 10^{-4}\,\mathrm{m}, 10^{-6}\,\mathrm{m}\}$ with respect to the reference solution obtained on the 800×480 cells mesh using $\delta_{\mathrm{ref}} = 10^{-6}\,\mathrm{m}$. As shown in Figure 19, the value of δ does not have a significant influence on the overall error as soon as δ is small enough. We also observe a moderate influence on the robustness of the non-linear solver for the values considered here.

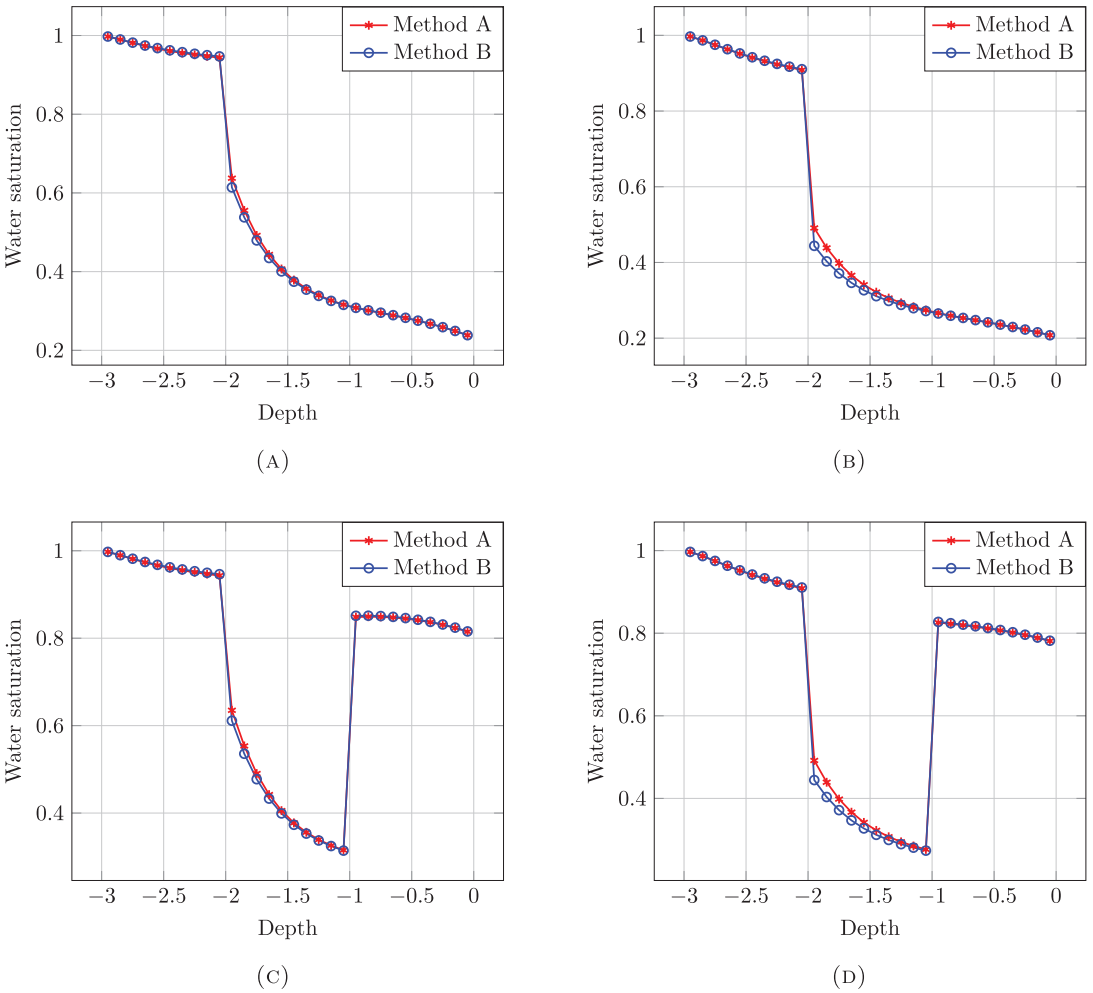

FIGURE 18. Water saturation profile obtained in the drainage test case with the van Genuchten Mualem model using method A and B along vertical cross-sections at different times. (A) Cross-section at $x = 0.95\,\mathrm{m}$, $t = 52 \times 10^4\,\mathrm{s}$. (B) Cross-section at $x = 0.95\,\mathrm{m}$, $t = 105 \times 10^4\,\mathrm{s}$. (C) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 52 \times 10^4\,\mathrm{s}$. (D) Cross-section at $x = 2.55\,\mathrm{m}$, $t = 105 \times 10^4\,\mathrm{s}$.

7. Conclusions and perspectives

This article aimed at proving that standard upstream mobility finite volume schemes for variable saturated porous media flows still converge in highly heterogeneous contexts without any specific treatment of the rock type discontinuities. The scheme is indeed shown to satisfy some energy stability which provides enough *a priori* estimates to carry out its numerical analysis. First, the existence of a unique solution to the nonlinear system stemming from the scheme is established thanks to a topological degree argument and from the monotonicity of the scheme. Besides, a rigorous mathematical convergence proof is conducted, based on compactness arguments. No error estimate can then be deduced from our analysis.

Because of the choice of a backward Euler in time discretization and from the upwind choice of the mobilities, a first order in time and space accuracy is expected in the case of homogeneous computational domains. We show in numerical experiments that without any particular treatment of the interfaces at rock discontinuities, this first

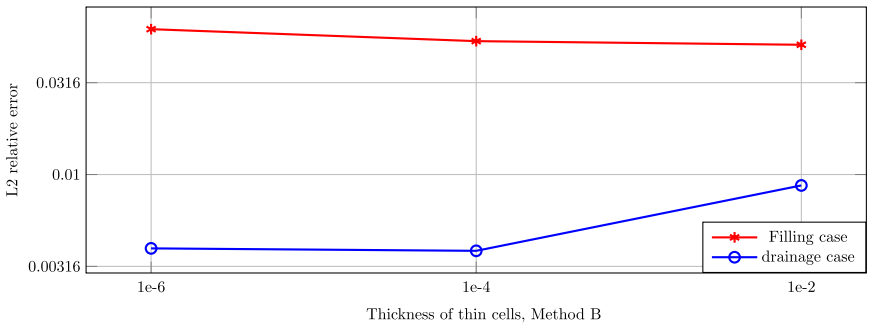

FIGURE 19. $L^2(Q_T)$ relative error in saturation as a function of the thickness δ of the thin cells with Method B using the 200×120 cells mesh.

order accuracy can be lost, especially in the case of Brooks–Corey nonlinearities. This motivates the introduction of a specific treatment of the interfaces. The approach we propose here is based on the introduction of additional unknowns located in fictitious small additional cells on both sides of each interface. Even though the rigorous convergence proof of this approach is not provided here in the multidimensional setting – such a proof can for instance be done by writing the scheme with the specific treatment of the interface (method B) as a perturbation of the scheme without any particular treatment of the interface (method A) –, the numerical experiments show that it allows to recover the first order accuracy without having major impacts on the implementation and on the behavior of the numerical solver.

For future researches, we suggest to test the so-called method B on a two-phase flow test and to compare it to the approaches presented in [10]. Moreover, in [6], we propose two other methods to really impose the pressure continuity condition at interfaces. A comparison between all methods will be shown.

Acknowledgements. C. Cancès acknowledges support from Labex CEMPI (ANR-11-LABX-0007-01)

REFERENCES

[1] A. Ait Hammou Oulhaj, C. Cancès and C. Chainais-Hillairet, Numerical analysis of a nonlinearly stable and positive control volume finite element scheme for Richards equation with anisotropy. *ESAIM: M2AN* **52** (2018) 1532–1567.

[2] B. Andreianov, C. Cancès and A. Moussa, A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. *J. Funct. Anal.* **273** (2017) 3633–3670.

[3] T. Arbogast, M. Obeyesekere and M.F. Wheeler, Numerical methods for the simulation of flow in root-soil systems. *SIAM J. Numer. Anal.* **30** (1993) 1677–1702.

[4] T. Arbogast, M.F. Wheeler and N.-Y. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.* **33** (1996) 1669–1687.

[5] S. Bassetto, C. Cancès, G. Enchéry and Q. H. Tran, Robust Newton solver based on variable switch for a finite volume discretization of Richards equation, In: Finite Volumes for Complex Applications IX – Methods, Theoretical Aspects, Examples edited by R. Klöfkorn, E. Keilegavlen, F.A. Radu and J. Fuhrmann. Vol. 323 of *Springer Proceedings in Mathematics & Statistics* (2020) 385–394.

[6] S. Bassetto, C. Cancès, G. Enchéry and Q.H. Tran, On several numerical strategies to solve Richards' equation in heterogeneous media with Finite Volumes. Working paper or preprint (2021) https://hal.archives-ouvertes.fr/hal-03259026.

[7] K. Brenner and C. Cancès, Improving Newton's method performance by parametrization: the case of the Richards equation. *SIAM J. Numer. Anal.* **55** (2017) 1760–1785.

[8] K. Brenner, C. Cancès and D. Hilhorst, Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure. *Comput. Geosci.* **17** (2013) 573–597.

[9] K. Brenner, M. Groza, L. Jeannin, R. Masson and J. Pellerin, Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media. *Comput. Geosci.* **21** (2017) 1075–1094.

[10] K. Brenner, R. Masson, E. H. Quenjel and J. Droniou, Total velocity-based finite volume discretization of two-phase Darcy flow in highly heterogeneous media with discontinuous capillary pressure. *IMA Journal of Numerical Analysis* (2021) https://doi.org/10.1093/imanum/drab018.

[11] K. Brenner, R. Masson and E.H. Quenjel, Vertex approximate gradient discretization preserving positivity for two-phase Darcy flows in heterogeneous porous media. *J. Comput. Phys.* **409** (2020) 109357.

[12] R.H. Brooks and A.T. Corey, Hydraulic properties of porous media. *Hydrol. Paper* **7** (1964) 26–28.

[13] C. Cancès, Nonlinear parabolic equations with spatial discontinuities. *Nonlinear Diff. Equ. Appl.* **15** (2008) 427–456.

[14] C. Cancès, Finite volume scheme for two-phase flow in heterogeneous porous media involving capillary pressure discontinuities. *ESAIM: M2AN* **43** (2009) 973–1001.

[15] C. Cancès and C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comput.* **85** (2016) 549–580.

[16] C. Cancès, F. Nabet and M. Vohralík, Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations. *Math. Comput.* **90** (2021) 517–563.

[17] V. Casulli and P. Zanolli, A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form. *SIAM J. Sci. Comput.* **32** (2010) 2255–2273.

[18] C. Chainais-Hillairet, J.-G. Liu and Y.-J. Peng, Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *ESAIM: M2AN* **37** (2003) 319–338.

[19] G. Chavent and J. Jaffré, Mathematical models and finite elements for reservoir simulation: single phase, multiphase and multicomponent flows through porous media. In: Vol. 17 of Studies in Mathematics and its Applications. North-Holland, Amsterdam (1986).

[20] Z. Chen and R.E. Ewing, Fully discrete finite element analysis of multiphase flow in groundwater hydrology. *SIAM J. Numer. Anal.* **34** (1997) 2228–2253.

[21] Z. Chen and R.E. Ewing, Degenerate two-phase incompressible flow. III. Sharp error estimates. *Numer. Math.* **90** (2001) 215–240.

[22] K. Deimling, Nonlinear Functional Analysis. Springer-Verlag, Berlin (1985).

[23] H.-J.G. Diersch and P. Perrochet, On the primary variable switching technique for simulating unsaturated–saturated flows. *Adv. Water Resour.* **23** (1999) 271–301.

[24] J. Droniou and R. Eymard, The asymmetric gradient discretisation method, edited by C. Cancès and P. Omnes. In: Finite Volumes for Complex Applications VIII – Methods and Theoretical Aspects. Vol. 199 of *Springer Proc. Math. Stat.* Springer, Cham (2017) 311–319.

[25] G. Enchéry, R. Eymard and A. Michel, Numerical approximation of a two-phase flow in a porous medium with discontinuous capillary forces. *SIAM J. Numer. Anal.* **43** (2006) 2402–2422.

[26] B.G. Ersland, M.S. Espedal and R. Nybø, Numerical methods for flow in a porous medium with internal boundaries. *Comput. Geosci.* **2** (1998) 217–240.

[27] R. Eymard and T. Gallouët, H-convergence and numerical schemes for elliptic problems. *SIAM J. Numer. Anal.* **41** (2003) 539–562.

[28] R. Eymard, M. Gutnic and D. Hilhorst, The finite volume method for Richards equation. *Comput. Geosci.* **3** (1999) 259–294.

[29] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods, In: Techniques of Scientific Computing (Part 3) edited by P.G. Ciarlet and J.-L. Lions. Vol. VII of *Handbook of Numerical Analysis*. North-Holland, Elsevier, Amsterdam (2000) 713–1018.

[30] R. Eymard, T. Gallouët, R. Herbin, M. Gutnic and D. Hilhorst, Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies. *M3AS: Math. Models Meth. Appl. Sci.* **11** (2001) 1505–1528.

[31] R. Eymard, R. Herbin and A. Michel, Mathematical study of a petroleum-engineering scheme. *ESAIM: M2AN* **37** (2003) 937–972.

[32] R. Eymard, T. Gallouët, C. Guichard, R. Herbin and R. Masson, TP or not TP, that is the question. *Comput. Geosci.* **18** (2014) 285–296.

[33] R. Eymard, C. Guichard, R. Herbin and R. Masson, Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. *Z. Angew. Math. Mech.* **94** (2014) 560–585.

[34] P.A. Forsyth, Y.S. Wu and K. Pruess, Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. *Adv. Water Resour.* **18** (1995) 25–38.

[35] K. Gärtner and L. Kamenski, Why do we need Voronoi cells and Delaunay meshes? In: Numerical Geometry, Grid Generation and Scientific Computing edited by V.A. Garanzha, L. Kamenski and H. Si. *Lecture Notes in Computational Science and Engineering*. Springer International Publishing, Cham (2019) 45–60.

[36] V. Girault, B. Riviere and L. Cappanera, A finite element method for degenerate two-phase flow in porous media. Part II: Convergence, *Journal of Numerical Mathematics* (2021) https://doi.org/10.1515/jnma-2020-0005.

[37] H. Hoteit and A. Firoozabadi, Numerical modeling of two-phase flow in heterogeneous permeable media with different capillarity pressures. *Adv. Water Resour.* **31** (2008) 56–73.

[38] M.R. Kirkland, R.G. Hills and P.J. Wierenga, Algorithms for solving Richards equation for variably saturated soils. *Water Resour. Res.* **28** (1992) 2049–2058.

[39] J. Leray and J. Schauder, Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup.* **51** (1934) 45–78.

[40] F. List and F.A. Radu, A study on iterative methods for solving Richards' equation. *Comput. Geosci.* **20** (2016) 341–353.

[41] F. Marinelli and D.S. Dunford, Semianalytical solution to Richards equation for layered porous media. *J. Irrig. Drain. Eng.* **124** (1998) 290–299.

[42] D. McBride, M. Cross, N. Croft, C. Bennett and J. Gebhardt, Computational modelling of variably saturated flow in porous media with complex three-dimensional geometries. *Int. J. Numer. Meth. Fluids* **50** (2006) 1085–1117.

[43] I.S. Pop, F.A. Radu and P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure. *J. Comput. Appl. Math.* **168** (2004) 365–373.
[44] F.A. Radu and W. Wang, Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media. *Nonlin. Anal.: Real World Appl.* **15** (2014) 266–275.
[45] 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) 1452–1478.
[46] L.A. Richards, Capillary conduction of liquids through porous mediums. *Physics* **1** (1931) 318–333.
[47] M.T. van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Amer. J.* **44** (1980) 892–898.