ESAIM: M2AN 52 (2018) 1533–1567
https://doi.org/10.1051/m2an/2017012

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

NUMERICAL ANALYSIS OF A NONLINEARLY STABLE AND POSITIVE CONTROL VOLUME FINITE ELEMENT SCHEME FOR RICHARDS EQUATION WITH ANISOTROPY ⋆

Ahmed Ait Hammou Oulhaj[1], Clément Cancès[1,*]
and Claire Chainais–Hillairet[1]

Abstract. We extend the nonlinear Control Volume Finite Element scheme of [C. Cancès and C. Guichard, *Math. Comput.* **85** (2016) 549–580]. to the discretization of Richards equation. This scheme ensures the preservation of the physical bounds without any restriction on the mesh and on the anisotropy tensor. Moreover, it does not require the introduction of the so-called Kirchhoff transform in its definition. It also provides a control on the capillary energy. Based on this nonlinear stability property, we show that the scheme converges towards the unique solution to Richards equation when the discretization parameters tend to 0. Finally we present some numerical experiments to illustrate the behavior of the method.

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

Received September 27, 2016. Accepted March 15, 2017.

1. Introduction

1.1. Presentation of the continuous problem

We are interested in the numerical approximation of Richards equation. It is a degenerate nonlinear parabolic equation modeling unsaturated flow in porous media. The diffusion terms can be anisotropic and heterogeneous. In order to ease the reading, we restrict our study to the case of a two-dimensional porous medium. However, the extension of our purpose to the three-dimensional framework does not lead to any theoretical difficulty.

Keywords and phrases. Unsaturated porous media flow, Richards equation, nonlinear discretization, nonlinear stability, convergence analysis.

⋆ *This work was supported by the GeoPor project funded by the French National Research Agency (ANR) with the grant ANR-13-JS01-0007-01 (project GEOPOR).*

[1] Univ. Lille, CNRS, UMR 8524, Inria – Laboratoire Paul Painlevé, 59000 Lille, France.

*Corresponding author: clement.cances@inria.fr

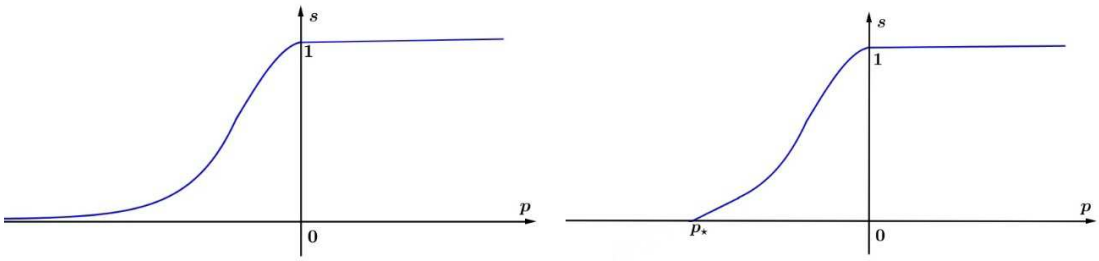

FIGURE 1. Typical water content functions. Two distinct behaviors are allowed in our study: *(left)* either the function s remains strictly positive on $\mathbb{R}$ but tends to 0 as p tends to $-\infty$ and $p_\star = -\infty$, or *(right)* there exists a finite value of $p_\star$ such that $s(p_\star) = 0$.

Let Ω be a polygonal connected open bounded subset of $\mathbb{R}^2$, and $t_f > 0$ a finite time horizon. We define $Q_{t_f} = \Omega \times (0, t_f)$. The Richards equation writes:

$$\begin{cases} \partial_t s(p) - \nabla \cdot (\eta(s(p)) \Lambda (\nabla p - \rho \mathbf{g})) = 0 & \text{in} \quad Q_{t_f}, \\ s(p)_{t=0} = s_0 & \text{in} \quad \Omega, \\ \eta(s(p)) \Lambda (\nabla p - \rho \mathbf{g}) \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0, t_f). \end{cases} \tag{1.1}$$

In (1.1), p denotes the water pressure, s the water content, η the water mobility function, Λ the intrinsic permeability tensor, and $\mathbf{g}$ is the gravity. We do the following assumptions on the data of the continuous problem (1.1):

(**A1**) The function $s : \mathbb{R} \to [0, 1]$ is increasing on $\mathbb{R}_-$ and takes the value 1 on $\mathbb{R}_+$. We assume that there exists $p_\star \in [-\infty, 0)$ such that $s(p_\star) = 0$, and that $s \in L^1(p_\star, 0)$. Figure 1 shows two typical profiles of the function s.

(**A2**) The water mobility function $\eta : [0, 1] \longrightarrow \mathbb{R}^+$ is assumed to be bounded, continuous, nondecreasing, and to fulfill (*cf.* Fig. 2)

$$\eta(0) = 0 \quad \text{and} \quad \eta(s) > 0 \ \text{ if } \ s \neq 0. \tag{1.2}$$

Moreover, we assume all along in this paper that

$$\xi_\star := \int_{p_\star}^{0} \sqrt{\eta(s(a))}\, \mathrm{d}a < +\infty. \tag{1.3}$$

Remark that (1.3) is trivially satisfied if $p_\star > -\infty$.

(**A3**) The permeability tensor Λ belongs $(L^\infty(\Omega))^{2\times 2}$, and it is supposed to be symmetric and uniformly elliptic on Ω, *i.e.*, there exists $(\overline{\Lambda}, \underline{\Lambda}) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ such that

$$\underline{\Lambda} |\mathbf{v}|^2 \leq \Lambda(\mathbf{x}) \mathbf{v} \cdot \mathbf{v} \leq \overline{\Lambda} |\mathbf{v}|^2, \quad \forall \mathbf{v} \in \mathbb{R}^2, \quad \text{for a.e.} \quad \mathbf{x} \in \Omega.$$

(**A4**) The initial data s_0 is supposed to belong to $L^\infty(\Omega; [0, 1])$, and we assume

$$0 < \overline{s}_0 := \frac{1}{\mathrm{meas}(\Omega)} \int_\Omega s_0(x)\, \mathrm{d}x < 1. \tag{1.4}$$

Since s is continuous and increasing on $[p_\star, 0]$, there exists a continuous and increasing function $s^{-1} : [0, 1] \to [p_\star, 0]$ such that, $s \circ s^{-1}(\zeta) = \zeta$ for all $\zeta \in [0, 1]$. Simple calculations (see for instance [13]) show that

$$\|s^{-1}\|_{L^1(0,1)} = \|s\|_{L^1(p_\star,0)} \leq C. \tag{1.5}$$

thanks to (**A1**).

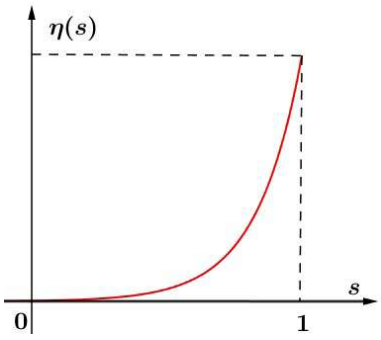

FIGURE 2. Typical water mobility function η.

Remark 1.1. The assumptions (**A1**) and (**A2**) impose some constraints on the nonlinearities $p \mapsto s(p)$ and $s \mapsto \eta(s)$. Let us connect these constraints to two very classical models.

- *van Genuchten–Mualem* model [30, 38] (see also [32]): $p_\star = -\infty$ and the function s is chosen as

$$s(p) = \begin{cases} (1+\alpha|p|^n)^{\frac{1-n}{n}} & \text{if} \quad p < 0, \\ 1 & \text{if} \quad p \geq 0, \end{cases}$$

  where $\alpha > 0$ is a fixed parameter. The condition (**A1**) is fulfilled if $n > 2$. The function $s^{-1} : (0,1] \to (-\infty, 0]$ is then given by

$$s^{-1}(s) = -\left(\frac{s^{-\frac{n}{n-1}} - 1}{\alpha}\right)^{1/n}, \qquad \forall s \in (0,1].$$

  whereas η is given by

$$\eta(s) = \kappa\sqrt{s}\left(\int_0^s \frac{1}{s^{-1}(a)}\mathrm{d}a\right)^2, \qquad \forall s \in [0,1]$$

  for some $\kappa > 0$. Condition (**A2**) —in particular (1.3)— is fulfilled.
- *Brooks-Corey* model [9]: here again, $p_\star = -\infty$. Let $p_b < 0$ and $\lambda > 0$ be given, then the function s is chosen as

$$s(p) = \begin{cases} \left(\dfrac{p+p_b}{p_b}\right)^{-\lambda} & \text{if} \quad p < 0, \\ 1 & \text{if} \quad p \geq 0. \end{cases}$$

  The integrability condition on s in (**A1**) is fulfilled as soon as $\lambda > 1$. The mobility function η is then chosen as

$$\eta(s) = \kappa s^{3+\frac{2}{\lambda}}, \qquad \forall s \in [0,1]$$

  for some $\kappa > 0$. Here again, Condition (1.3) of (**A2**) is fulfilled.

We define the function $\Gamma : \mathbb{R} \to \mathbb{R}_+$ (called *capillary energy function*) by

$$\Gamma(p) = \int_0^p a s'(a)\,\mathrm{d}a. \tag{1.6}$$

The function $\Gamma \circ s^{-1}$ is convex on $[0,1]$, and it follows from the definition (1.6) that

$$\partial_t \Gamma(p) = p \partial_t s(p). \tag{1.7}$$

In order to give a proper mathematical sense to the solution of (1.1), we need to introduce the Lipschitz continuous increasing function $\xi : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\xi(p) = \int_0^p \sqrt{\eta(s(a))}\,\mathrm{d}a, \quad \forall\, p \in \mathbb{R}. \tag{1.8}$$

Since $\sqrt{\eta \circ s \circ \xi^{-1}}$ is uniformly continuous, it admits a modulus of continuity *i.e.*, there exits a continuous function $\mu : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\mu(0) = 0, \quad |\sqrt{\eta \circ s \circ \xi^{-1}}(x) - \sqrt{\eta \circ s \circ \xi^{-1}}(y)| \leq \mu(|x-y|),\ \forall x, y \in [\xi_\star, +\infty). \tag{1.9}$$

We introduce the so-called hydraulic head u defined by

$$u(\boldsymbol{x}, t) = \frac{p(\boldsymbol{x}, t)}{\rho g} + z(\boldsymbol{x}) \quad \text{for all} \quad (\boldsymbol{x}, t) \in Q_{t_f}, \quad \text{for all} \quad t_f > 0,$$

where g denotes the modulus of $\mathbf{g}$ and the function $z(\boldsymbol{x})$ is the projection of the point $\boldsymbol{x}$ on the vertical axis, oriented upward by $-\mathbf{g}/g$. With a simple adimensionalization, we can assume that $\rho g = 1$. The system (1.1) then rewrites:

$$\begin{cases} \partial_t s(p) - \nabla \cdot (\eta(s(p)) \Lambda \nabla u) = 0 & \text{in} \quad Q_{t_f}, \\ s(p)_{t=0} = s_0 & \text{in} \quad \Omega, \\ \eta(s(p)) \Lambda \nabla u \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0, t_f), \\ u = p + z & \text{in} \quad Q_{t_f}. \end{cases} \tag{1.10}$$

Remark 1.2. In Section 5.1, we will present a test case without gravity. *Stricto sensu*, this case is not included in our study. But it corresponds to the simpler case $p = u$ (mainly carried out in [11]) for which our analysis can be straightforwardly adapted.

Multiplying (formally) the equation (1.10) by u and integrating on Ω yields the classical energy/dissipation property:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \left(\Gamma(p) + s(p) z(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} + \int_\Omega \eta(s(p)) \Lambda \nabla u \cdot \nabla u \,\mathrm{d}\boldsymbol{x} = 0, \qquad \forall\, t \in (0, t_f). \tag{1.11}$$

This allows in particular to show that the capillary energy remains bounded and that the function $\xi(p)$ belongs to $L^2((0,T); H^1(\Omega))$, *i.e.*,

$$\int_\Omega \Gamma(p(\boldsymbol{x}, t)) \mathrm{d}\boldsymbol{x} + \int_0^t \int_\Omega |\nabla \xi(p(\boldsymbol{x}, \tau))|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\tau \leq C, \qquad \forall t \in (0, t_f). \tag{1.12}$$

Remark 1.3. It is clear that the quantity $\int_\Omega s(p) z(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ represents the *gravitational (potential) energy* of the fluid. Therefore, it follows from (1.11) that the quantity $\int_\Omega \Gamma(p) \mathrm{d}\boldsymbol{x}$ also corresponds to an energy. Since it originates from capillary (or suction) effects, we call it *capillary energy*. The *free energy* $\int_\Omega \left(\Gamma(p) + s(p) z(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}$ is then obtained as the sum of the capillary and gravitational energies. It is supposed to decay with time thanks to (1.11) (see also Sect. 5.4 for a numerical evidence).

Definition 1.4 (weak solution)**.** A measurable function $p : Q_{t_f} \to \mathbb{R}$ is said to be a weak solution of (1.1) if $p \geq p_\star$ a.e. in Q_{t_f}, if $\xi(p)$ belongs to $L^2((0, t_f); H^1(\Omega))$, and if, for all $\psi \in \mathcal{C}_c^\infty(\overline{\Omega} \times [0, t_f))$, one has

$$\begin{aligned} \iint_{Q_{t_f}} s(p) \partial_t \psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega s_0 \psi(., 0) \,\mathrm{d}\boldsymbol{x} - \iint_{Q_{t_f}} \sqrt{\eta(s(p))} \Lambda \nabla \xi(p) \cdot \nabla \psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\ - \iint_{Q_{t_f}} \eta(s(p)) \rho g \Lambda \nabla z \cdot \nabla \psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = 0. \end{aligned} \tag{1.13}$$

The notion of weak solution is motivated by the following theorem.

Theorem 1.5. *Under assumptions* (**A1**)–(**A4**)*, there exists a unique weak solution to the problem* (1.1) *in the sense of Definition* 1.4.

The existence of a solution is a by-product of the convergence of the scheme proved in Section 4. It can also be obtained by compactness arguments following the program of Alt and Luckhaus [3]. Concerning the uniqueness, since we consider no-flux boundary conditions, we can not directly apply Otto's result [31], where Dirichlet boundary conditions are imposed. However, a slight adaptation of Otto's proof detailed in appendix (*cf.* Prop. A.4) allows us to extend the uniqueness result to our framework.

1.2. Goal and positioning of the paper

Because of its broad interest in the environmental studies, the Richards equation [36] has been the purpose of many research papers, especially in the field of numerical analysis. Richards equation is locally conservative and a particular effort was made to preserve this property in most of the contributions.

A conservative Finite Difference scheme has been studied numerically in [39]. However, there is up to our knowledge no convergence proof for the scheme presented in [39]. Moreover, restrictive conditions have to be prescribed on the grid and on the permeability tensor Λ. The convergence of Two-Point Flux Approximation Finite Volume schemes have been studied in [20] for a scheme that requires the introduction of the Kirchhoff transform, and in [19] for a scheme expressed in physical variables (saturation and pressure), but under the non-physical assumption that the mobility function was not degenerated (*i.e.*, $\eta(s) \geq \eta_\star > 0$ for all s). In both [20] and [19], it was moreover required that the porous medium was isotropic (*i.e.*, $\Lambda = \lambda \mathrm{Id}$) and that the mesh satisfies the so-called orthogonality condition (see, *e.g.*, [17], Def. 9.1 and [16]) so that the two-point flux approximation is consistent. Since they are naturally locally conservative, Mixed Finite Elements have been widely used for the approximation of Richards equation. Let us for instance mention [5,34,35] where the authors managed to provide an error estimate. Nevertheless, the schemes studied in [5, 34, 35] rely on the introduction of the Kirchhoff transform, and on a regularization of the problem in [35] to overcome the difficulties due the degeneracy. Let us also mention the extension of Multi-Point Flux Approximation Finite Volume schemes to the context of Richards equation in [8, 27]. Note that Mixed Finite Elements and Multi-Point Flux Approximation Finite Volumes may produce over- and undershoots on the saturation. We refer to [15] for a review of the numerous Finite Volume methods developed in the last decades that can be applied to the discretization of Richards equation.

The method we study here was designed on the following specifications:

(a) to handle anisotropic and heterogeneous anisotropy tensors;
(b) to avoid the introduction of non-physical quantities like, *e.g.*, the Kirchhoff transform;
(c) to preserve the physical bounds on the saturation;
(d) to conserve locally the mass of fluid;
(e) to converge towards the solution to the continuous problem (mathematical proof and numerical evidence).

The scheme we propose belongs to the family of the so-called Control Volume Finite Element schemes introduced in the context of porous media flows by Forsyth [21, 22]. Roughly speaking, it consists in an interpretation of Finite Elements with mass lumping as a locally conservative method on dual cells. It was already noticed in [21] that the grid had to fulfill some restrictive condition unless the transmissivities may become negative. It results that the reconstructed numerical flux goes the opposite sense to the physical one. More precisely, the triangular grid has to fulfill a so-called Delaunay condition in the two-dimensional isotropic case $\Lambda = \lambda \mathrm{Id}$. But in the case where Λ is a spatially varying full tensor, there is no algorithm up to our knowledge to build a triangulation such that the transmissivity remain nonnegative. As it will be proved in the sequel, the method we propose still converges even in the case where negative transmissivities appear. Our scheme is an extension of the one studied in [10, 11]. It is based on a suitable upwinding of the mobility (*i.e.*, w.r.t. the numerical flux and not

w.r.t. the physical one) that allows to preserve the physical bounds (but not the monotonicity as in [22]). Moreover, we show that our method provides a control on the capillary energy and that this control is sufficient to perform a convergence proof based on compactness arguments. Alternatively, the convergence of the Finite Volume approximation can be obtained by means of error estimates (see [33] in the case where $\mathbf{g} = 0$).

The paper is organized as follows. In Section 2, we introduce the scheme and we state the main results of our paper. Theorem 2.4 states the existence of a solution to scheme which preserves the physical bounds and for which the capillary energy and the energy dissipation are bounded uniformly w.r.t. the grid. Theorem 2.5 states the convergence of a sequence of approximate solutions given by the scheme to the unique weak solution to (1.10) (its uniqueness is proved in Appendix). In Section 3, we derive *a priori* estimates on the discrete solution. They allow us to prove in Section 3.3 the existence of a discrete solution to the nonlinear system corresponding to the scheme. Section 4 is devoted to the convergence proof of the scheme. This proof is based first on the compactness of the sequence of approximate solutions and then on the identification of the limit. We finally present numerical experiments in Section 5, which confirm the theoretical results we proved. We take care to fairly present the advantages and the drawbacks of the method from a computational point of view.

2. The numerical scheme

2.1. Discretization of Q_{t_f}

2.1.1. Discretizations of Ω

The CVFE method requires the introduction of two different space discretizations of Ω: a *primal triangular mesh* and a *dual barycentric mesh*.

The *primal triangular mesh* is denoted by $\mathcal{T}$. It is a conformal triangular discretization of the polygonal domain Ω, consisting in open bounded separated triangles satisfying $\bigcup_{T\in\mathcal{T}} \overline{T} = \overline{\Omega}$. For $T \in \mathcal{T}$, we denote by $\boldsymbol{x}_T$ the center of gravity of T, by h_T the diameter of the triangle T, and by ρ_T the diameter of the largest ball inscribed in the triangle T. Then, we define the mesh diameter h and the mesh regularity $\theta_{\mathcal{T}}$ by

$$h = \max_{T\in\mathcal{T}} h_T, \quad \theta_{\mathcal{T}} = \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}.$$

We denote by $\mathcal{V}$ the set of the vertices of the discretization $\mathcal{T}$, located at positions $(\boldsymbol{x}_K)_{K\in\mathcal{V}}$. The set $\mathcal{E}$ of the edges of $\mathcal{T}$ is made of straight segments σ joining two vertices of $\mathcal{V}$. Given $T, T' \in \mathcal{T}$, we assume that $\overline{T} \cap \overline{T'}$ is either empty, or it is reduced to $\boldsymbol{x}_K$ for some $K \in \mathcal{V}$, or it consists in an edge σ belonging $\mathcal{E}$. For $T \in \mathcal{T}$, we denote by $\mathcal{E}_T$ the set of the edges of T: $\bigcup_{\sigma\in\mathcal{E}_T} \overline{\sigma} = \partial T$. We assume that $\mathcal{E} = \bigcup_{T\in\mathcal{T}} \mathcal{E}_T$. Given two vertices $K, L \in \mathcal{V}$ of a triangle T, then the edge joining $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ is denoted by σ_{KL}. For $K \in \mathcal{V}$, one denotes by $\mathcal{T}_K$ the subset of $\mathcal{T}$ made the triangles admitting K as a vertex, by $\mathcal{E}_K$ the set of edges having the vertex K as an extremity, and by $\mathcal{V}_K$ the subset of $\mathcal{V}$ such that, if $L \in \mathcal{V}_K$, then $[\boldsymbol{x}_K, \boldsymbol{x}_L]$ is an edge of $\mathcal{E}_K$.

Once the *primal triangular mesh* has been built, we can define its *dual barycentric mesh* $\mathcal{M}$ as follows. To each $K \in \mathcal{V}$, we associate a cell ω_K whose vertices are the isobarycenters $\boldsymbol{x}_T$ of the triangles $T \in \mathcal{T}_K$ and the isobarycenters $\boldsymbol{x}_\sigma$ of the edges $\sigma \in \mathcal{E}_K$. Note that $\overline{\Omega} = \bigcup_{K\in\mathcal{V}} \overline{\omega}_K$. We refer to Figure 3 for an illustration of the primary and dual barycentric meshes. The 2-dimensional Lebesgue measure of ω_K is denoted by m_K.

Let us now introduce some useful functional spaces. The space $V_{\mathcal{T}} \subset \mathcal{C}(\overline{\Omega})$ is made of piecewise affine functions on the primal mesh, *i.e.*,

$$V_{\mathcal{T}} = \{f \in H^1(\Omega) \mid f_{|_T} \text{ is affine}, \ \forall T \in \mathcal{T}\}.$$

For all $K \in \mathcal{V}$, we denote by e_K the unique element of $V_{\mathcal{T}}$ such that $e_K(\boldsymbol{x}_K) = 1$ and $e_K(\boldsymbol{x}_L) = 0$ if $L \in \mathcal{V}\backslash\{K\}$. The geometrical construction of ω_K ensures that

$$\int_\Omega e_K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\omega_K} \mathrm{d}\boldsymbol{x} =: m_K, \qquad \forall K \in \mathcal{V}.$$

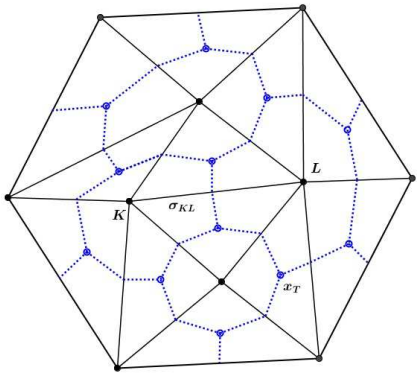

FIGURE 3. The triangular mesh $\mathcal{T}$ (*solid line*) and its corresponding dual barycentric dual mesh $\mathcal{M}$ (*dashed line*).

We can also define the set of the piecewise constant functions on $\mathcal{M}$, $X_{\mathcal{M}}$, by

$$X_{\mathcal{M}} = \{f : \Omega \longrightarrow \mathbb{R} \text{ measurable} \mid f_{|\omega_K} \in \mathbb{R} \text{ is constant}, \quad \forall K \in \mathcal{V}\}.$$

Given a vector $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, there exists a unique $u_{\mathcal{T}} \in V_{\mathcal{T}}$ and a unique $u_{\mathcal{M}} \in X_{\mathcal{M}}$ such that $u_{\mathcal{T}}(\boldsymbol{x}_K) = u_{\mathcal{M}}(\boldsymbol{x}_K) = u_K$ for all $K, L \in \mathcal{V}$. Let us note that $u_{\mathcal{T}} = \sum_{K\in\mathcal{V}} u_K e_K$. Moreover, for all $q \in [1, \infty)$, there exist C_1 and C_2 depending only on q and on $\theta_{\mathcal{T}}$ such that

$$C_1\|u_{\mathcal{T}}\|_{L^q(\Omega)} \leq \|u_{\mathcal{M}}\|_{L^q(\Omega)} \leq C_2\|u_{\mathcal{T}}\|_{L^q(\Omega)}, \qquad \forall\ (u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}. \tag{2.1}$$

A proof of the above inequalities can be found for instance in ([12], Lem. A.6).

2.1.2. Space-time discretizations

In order to avoid heavier notations, we restrict our study to the case of a uniform time discretization of $(0, t_f)$. However, all the results presented in this paper can be extended to general time discretizations without any technical difficulty. In what follows, we assume that the spatial mesh is fixed and does not change with the time step.

Let N be a nonnegative integer, then we define $\Delta t = \dfrac{t_f}{N+1}$, and $t_n = n\Delta t$ for all $n \in \{0, \ldots, N+1\}$, so that $t_0 = 0$, and $t_{N+1} = t_f$.

We define the space and time discrete spaces $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ as the set of piecewise constant functions in time with values in $V_{\mathcal{T}}$ and $X_{\mathcal{M}}$ respectively:

$$\begin{aligned} V_{\mathcal{T},\Delta t} =& \{f : Q_{t_f} \to \overline{\mathbb{R}} \mid f(x,t) = f(x,t^{n+1}) \in V_{\mathcal{T}}, \quad \forall t \in (t_n, t_{n+1}]\}, \\ X_{\mathcal{M},\Delta t} =& \left\{f : Q_{t_f} \to \overline{\mathbb{R}} \ \mid f(x,t) = f(x,t^{n+1}) \in X_{\mathcal{M}}, \quad \forall t \in (t_n, t_{n+1}]\right\}. \end{aligned}$$

For a given $(u_K^{n+1})_{n\in\{0,\ldots,N\},K\in\mathcal{V}} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$, we denote by $u_{\mathcal{T},\Delta t}$ and $u_{\mathcal{M},\Delta t}$ the unique elements of $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ respectively such that

$$u_{\mathcal{T},\Delta t}(x_K, t) = u_{\mathcal{M},\Delta t}(x_K, t) = u_K^{n+1}, \quad \forall K \in \mathcal{V},\ \forall t \in (t_n, t_{n+1}]. \tag{2.2}$$

2.2. Finite elements

The method we propose, and more generally the CVFE method, is based on P_1-finite elements. We introduce in this section the technical material that is needed in order to define the scheme and to perform its analysis.

We define the transmissibility coefficients

$$a_{KL}^T = -\int_T \Lambda \nabla e_K \cdot \nabla e_L \,\mathrm{d}\boldsymbol{x} = a_{LK}^T, \quad \forall T \in \mathcal{T},\ \forall (K,L) \in \mathcal{V}^2, \tag{2.3}$$

and

$$a_{KL} = a_{LK} = -\int_\Omega \Lambda \nabla e_K \cdot \nabla e_L \,\mathrm{d}\boldsymbol{x} = \sum_{T \in \mathcal{T}} a_{KL}^T, \quad \forall (K,L) \in \mathcal{V}^2. \tag{2.4}$$

Note that $a_{KL} = 0$ unless $\sigma_{KL} \in \mathcal{E}$. Moreover, since $\sum_{K \in \mathcal{V}} \nabla e_K = 0$, we have that :

$$-a_{KK} = \sum_{L \neq K} a_{KL} > 0. \tag{2.5}$$

As a consequence of (2.4)−(2.5), given $u_\mathcal{T}$ and $v_\mathcal{T}$ two elements of $V_\mathcal{T}$, one has

$$\int_\Omega \Lambda \nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T} \,\mathrm{d}\mathbf{x} = \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL}(u_K - u_L)(v_K - v_L) = \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL} \in \mathcal{E}_T} a_{KL}^T (u_K - u_L)(v_K - v_L). \tag{2.6}$$

The following lemma plays a crucial role in the numerical analysis carried out in this paper. We refer to ([11], Lem. 3.2) for its proof.

Lemma 2.1. *There exists C_3 depending only on $\theta_\mathcal{T}$, $\Lambda_\star$ and $\Lambda^\star$ such that, for all $u_\mathcal{T} \in V_\mathcal{T}$, one has*

$$\sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}|(u_K - u_L)^2 \leq \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL} \in \mathcal{E}_T} |a_{KL}^T|(u_K - u_L)^2 \leq C_3 \int_\Omega \Lambda \nabla u_\mathcal{T} \cdot \nabla u_\mathcal{T} \,\mathrm{d}\boldsymbol{x}.$$

2.3. The nonlinear CVFE scheme

In this section, we explicit the discretization of the problem (1.1) we will study in this paper. The time discretization relies on backward Euler scheme, while the space discretization relies on finite elements with mass lumping and a suitable upwinding of the mobility.

The discretization $s_\mathcal{M}^0 \in X_\mathcal{M}$ of the initial data is defined by

$$s_K^0 = \frac{1}{m_K} \int_{\omega_K} s_0(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{V}. \tag{2.7}$$

In the sequel, we will make use of the shortened notation

$$z_K = z(\boldsymbol{x}_K), \qquad \forall K \in \mathcal{V}.$$

Let us now introduce the scheme. For all $n \in \{0, \dots, N\}$, a solution $\left(p_K^{n+1}\right)_{K \in \mathcal{V}}$ to the scheme at the time step $n+1$ has to satisfy the following equations: for all $K \in \mathcal{V}$,

$$\frac{s(p_K^{n+1}) - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} \eta_{KL}^{n+1} a_{KL}(u_K^{n+1} - u_L^{n+1}) = 0, \tag{2.8a}$$

$$u_K^{n+1} = p_K^{n+1} + \rho g z_K, \tag{2.8b}$$

$$s_K^{n+1} = s(p_K^{n+1}), \tag{2.8c}$$

$$\eta_{KL}^{n+1} = \begin{cases} \eta(s_K^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) \geq 0, \\ \eta(s_L^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) < 0. \end{cases} \tag{2.8d}$$

Remark 2.2. It follows from the monotonicity of the mobility and water content functions η and s that (2.8d) is equivalent to

$$\eta_{KL}^{n+1} = \begin{cases} \max\limits_{p\in I_{KL}^{n+1}} \eta(s(p)) & \text{if} \quad a_{KL}(p_K^{n+1}-p_L^{n+1})(u_K^{n+1}-u_L^{n+1}) \geq 0, \\ \min\limits_{p\in I_{KL}^{n+1}} \eta(s(p)) & \text{if} \quad a_{KL}(p_K^{n+1}-p_L^{n+1})(u_K^{n+1}-u_L^{n+1}) \leq 0, \end{cases} \tag{2.9}$$

where

$$I_{KL}^{n+1} = \left[\min\left(p_K^{n+1}, p_L^{n+1}\right), \max\left(p_K^{n+1}, p_L^{n+1}\right)\right].$$

It is then worth noticing that the monotonicity assumption on η can be bypassed if one enforces (2.9) directly instead of (2.8d) for the definition of the upwind mobility.

This scheme, whose construction is based on finite elements *via* (2.4), can be interpreted as a finite volume scheme. Indeed denoting by

$$F_{KL}^{n+1} = a_{KL}\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1}),$$

the scheme (2.8) can be rewritten under the locally conservative form on the dual cells ω_K:

$$\begin{cases} F_{KL}^{n+1} + F_{LK}^{n+1} = 0, & \text{for all} \quad \sigma_{KL} \in \mathcal{E}_K \\ \dfrac{s_K^{n+1}-s_K^n}{\Delta t} m_K + \displaystyle\sum_{\sigma_{KL}\in\mathcal{E}_K} F_{KL}^{n+1} = 0, & \text{for all} \quad K \in \mathcal{V}. \end{cases} \tag{2.10}$$

As a straightforward consequence, we can claim that the scheme (2.8) is globally conservative, *i.e.*,

$$\sum_{K\in\mathcal{V}} m_K s_K^{n+1} = \sum_{K\in\mathcal{V}} m_K s_K^n = \int_\Omega s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall n \geq 0. \tag{2.11}$$

Remark 2.3. It will appear in the analysis that the discrete pressures p_K^{n+1} are always bounded (see Lems. 3.10 and 3.11). Therefore, all the terms appearing in the scheme are finite, hence the products and sums are well defined.

2.4. Main results

The scheme (2.8) amounts to a nonlinear system to be solved at each time step. The existence of a solution to this system is therefore non trivial. The first result we highlight is thus the existence of a solution to the scheme (2.8) and the stability in terms of the discrete capillary energy.

Theorem 2.4. *There exists (at least) one solution* $\left(p_K^{n+1}\right)_{K\in\mathcal{V},n\in\{0,\dots,N\}}$ *to the scheme* (2.8a)*. Moreover,* $0 \leq s_K^n \leq 1$ *for all* $K \in \mathcal{V}$ *and for all* $n \in \{0,\dots,M\}$*, and there exists* C *depending only on* $\theta_{\mathcal{T}}$*,* Λ*,* Ω*,* t_f*,* $\|s\|_{L^1(p_\star,0)}$*, and* $\|\eta\|_\infty$ *such that*

$$\sup_{n\in\{0,\dots,N\}} \sum_{K\in\mathcal{V}} m_K \Gamma(p_K^{n+1}) + \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 \leq C.$$

Once we have the discrete solution $\left(p_K^{n+1}\right)_{K\in\mathcal{V},n\in\{0,\dots,N\}}$ at hand for all meshes and all time discretizations, then we can study the convergence of the scheme when the discretization parameters tend to 0. More precisely, consider a sequence $(\mathcal{T}_m)_{m\geq 1}$ of triangulations of Ω such that

$$h_m = \max_{T\in\mathcal{T}_m} \mathrm{diam}(T) \underset{m\to\infty}{\longrightarrow} 0, \tag{2.12}$$

and such that there exists $\theta^\star > 0$ such that

$$\theta_{\mathcal{T}_m} \leq \theta^\star, \quad \forall m \geq 1. \tag{2.13}$$

A sequence of dual meshes $(\mathcal{M}_m)_{m\geq 1}$ corresponding to the triangular meshes $(\mathcal{T}_m)_{m\geq 1}$ is built as in Section 2.1.1. Let $(N_m)_{m\geq 1}$ be an increasing sequence of integers, then we define the corresponding sequence of time steps $\Delta t_m = \frac{t_f}{N_m+1}$ tending to 0 as m tends to ∞. To this sequence of discretizations of Q_{t_f} corresponds a sequence of solutions $\left(p_K^{n+1}\right)_{K\in\mathcal{V}_m, n\in\{0,\dots,N_m\}}$ to the scheme. Thanks to these solutions, we can construct the functions $s_{\mathcal{M}_m,\Delta t_m} \in X_{\mathcal{M}_m,\Delta t_m}$ and $\xi_{\mathcal{T}_m,\Delta t_m} \in V_{\mathcal{T}_m,\Delta t_m}$ defined by

$$s_{\mathcal{M}_m,\Delta t_m}(\boldsymbol{x}_K, t_{n+1}) = s(p_K^{n+1}) = s_K^{n+1}, \qquad \forall K \in \mathcal{V}_m,\ \forall n \in \{0,\dots,N_m\}, \tag{2.14}$$

and

$$\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_K, t_{n+1}) = \xi(p_K^{n+1}) = \xi_K^{n+1}, \qquad \forall K \in \mathcal{V}_m,\ \forall n \in \{0,\dots,N_m\}. \tag{2.15}$$

Once these sequences of discrete functions at hand, we can state the second main result of this paper, namely the convergence of the scheme (2.8).

Theorem 2.5. *Let $(\mathcal{T}_m)_{m\geq 1}$ be a sequence conformal triangular discretization of Ω such that (2.12) and (2.13) hold. Let $(s_{\mathcal{M}_m,\Delta t_m})_m$ and $(\xi_{\mathcal{T}_m,\Delta t_m})_m$ be the functions reconstructed from the solutions $\left(\left(p_K^{n+1}\right)_{K,n}\right)_m$ to the scheme (2.8) thanks to formulas (2.14)–(2.15). Then*

$$\begin{aligned} &s_{\mathcal{M}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} s(p) \quad \text{a.e in } Q_{t_f}, \\ &\xi_{\mathcal{T}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} \xi(p) \quad \text{weakly in } L^2((0,t_f);H^1(\Omega)) \text{ and strongly in } L^2(Q_{t_f}), \end{aligned}$$

where p is the unique solution to the continuous problem (1.1).

The proof of Theorem 2.4 is addressed in Section 3. The convergence of the scheme towards a weak solution is the purpose of Section 4, while the uniqueness of the weak solution is proved in appendix, *cf.* Proposition A.4. Numerical illustrations are provided in Section 5.

3. Discrete properties, a *priori* estimates and existence

In this section, we establish *a priori* estimates, among which the positivity of the saturation and the stability of the capillary energy. These estimates allow to prove the existence of a solution to the nonlinear system (2.8). They are also keystones in order to perform the convergence analysis later on.

3.1. A uniform L^∞-estimate on $s_{\mathcal{M},\Delta t}$

In what follows, $(p_K^{n+1})_{K\in\mathcal{V},n\geq 0}$ denotes a solution to the scheme (2.8) (whose existence will be established later). This allows to define the quantities $s_K^{n+1} = s(p_K^{n+1})$ and $\xi_K^{n+1} = \xi(p_K^{n+1})$ for all $K \in \mathcal{V}$ and all $n \in \{0,\dots,N\}$.

Proposition 3.1. *For all $K \in \mathcal{V}$, and all $n \in \{0,\dots,N\}$, one has*

$$0 \leq s_K^n \leq 1. \tag{3.1}$$

Equivalently, one has

$$p_\star \leq p_K^{n+1}, \qquad \forall K \in \mathcal{V},\ \forall n \in \{0,\dots,N\}. \tag{3.2}$$

Proof. First of all, note that there is nothing to prove if $p_\star = -\infty$. Therefore, we restrict our attention to the case of a finite $p_\star$. The property (3.1) holds for $n = 0$ thanks to the discretization (2.7) of the initial data. Assume now (3.1) holds at time step n. It is equivalent to prove $p_K^{n+1} \geq p_\star$. Assume that

$$p_{K_m}^{n+1} = \min_{L\in\mathcal{V}} p_L^{n+1} < p_\star \quad \Leftrightarrow \quad s_{K_m}^{n+1} < 0. \tag{3.3}$$

In view of the definition (2.9) of $\eta_{K_mL}^{n+1}$, and of the fact that $\eta(s) = 0$ if $s < 0$, it follows from (2.8d) that

$$\eta_{K_mL}^{n+1} = 0 \quad \text{if} \quad a_{K_mL}(u_{K_m}^{n+1} - u_L^{n+1}) \geq 0.$$

Therefore, the scheme (2.8) at vertex K_m rewrites

$$s_{K_m}^{n+1} = s_{K_m}^{n} - \frac{\Delta t}{m_{K_m}} \sum_{\sigma_{K_mL}\in\mathcal{E}} \eta_{K_mL}^{n+1} a_{K_mL}(u_{K_m}^{n+1} - u_L^{n+1}) \geq 0.$$

This yields a contradiction with (3.3). Hence, the L^∞ estimate (3.1) holds at the time step $n+1$, thus for all n. □

3.2. Capillary energy estimate and the control of the dissipation

The goal of this section is to get an *a priori* control for the capillary energy of the discrete solution and to derive some estimates coming from the dissipation of the energy. We were not able to derive the discrete counterpart of the energy/dissipation estimate (1.11). However, we can prove a discrete counterpart of (1.12) (*cf.* Prop. 3.2) that appears to be sufficient to establish Theorems 2.4 and 2.5. In what follows, we assume that $(s_K^n)_{K\in\mathcal{V}}$ is known and $\left(p_K^{n+1}\right)_{K\in\mathcal{V}}$ denotes an arbitrary solution to the scheme (2.8).

Proposition 3.2. *There exists C_4 depending only on $\theta_\mathcal{T}$, Λ, Ω, t_f, $\|s\|_{L^1(p_\star,0)}$, and $\|\eta\|_\infty$ such that*

$$\sup_{n\in\{0,\dots,N\}} \sum_{K\in\mathcal{V}} m_K \Gamma(p_K^{n+1}) + \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL} \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 \leq C_4.$$

The proof of Proposition 3.2 is based on several Lemmas stated below. This section also contains technical lemmas that will be useful in the convergence proof of Section 4.

Lemma 3.3. *There exists C_5 depending only on Ω, s such that, for all $\nu \in \{0,\dots,N\}$, one has*

$$\sum_{K\in\mathcal{V}} m_K \Gamma(p_K^{\nu+1}) + \sum_{n=0}^{\nu} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) \leq C_5. \tag{3.4}$$

Proof. We multiply the scheme (2.8a) by $\Delta t p_K^{n+1}$ and sum on $K \in \mathcal{V}$. This yields:

$$A + B = 0,$$

where

$$A = \sum_{K\in\mathcal{V}} m_K(s_K^{n+1} - s_K^n)p_K^{n+1}, \quad B = \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})p_K^{n+1}.$$

Since $a_{KL} = a_{LK}$ and $\eta_{KL}^{n+1} = \eta_{LK}^{n+1}$, we can rewrite

$$B = \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}).$$

By convexity of $\Gamma \circ s^{-1}$ one deduces this estimation

$$A \geq \sum_{K \in V} m_K (\Gamma(p_K^{n+1}) - \Gamma \circ s^{-1}(s_K^n)).$$

Summing over $n \in \{0, \dots, \nu\}$ provides

$$\sum_{K \in \mathcal{V}} m_K \Gamma(p_K^{\nu+1}) + \sum_{n=0}^{\nu} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \eta_{KL}^{n+1} (u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) \leq \sum_{K \in \mathcal{V}} m_K \Gamma \circ s^{-1}(s_K^0). \tag{3.5}$$

It remains to check that for $b \in [0,1]$,

$$0 \leq \Gamma \circ s^{-1}(b) = \int_0^{s^{-1}(b)} a s'(a) \, \mathrm{d}a = \int_1^b s^{-1}(a) \, \mathrm{d}a \leq \|s^{-1}\|_{L^1(0,1)} < +\infty,$$

ensuring that

$$\sum_{K \in \mathcal{V}} m_K \Gamma \circ s^{-1}(s_K^0) \leq \int_\Omega \Gamma \circ s^{-1}(s_0) \, \mathrm{d}\boldsymbol{x} \leq |\Omega| \|s^{-1}\|_{L^1(0,1)}$$

thanks to Jensen's inequality and to (1.5). □

From the previous lemma, we can get an estimate on the spatial variations of the function $\xi_{\mathcal{T},\Delta t}$. In order to ease the reading, we use the shortened notation

$$\xi_K^{n+1} = \xi(p_K^{n+1}), \qquad \forall K \in \mathcal{V}, \ \forall n \in \{0, \dots, N\},$$

and we define

$$\tilde{\eta}_{KL}^{n+1} = \begin{cases} \left(\dfrac{\xi_K^{n+1} - \xi_L^{n+1}}{p_K^{n+1} - p_L^{n+1}} \right)^2 & \text{if} \quad p_K^{n+1} \neq p_L^{n+1}, \\ \eta(s_K^{n+1}) & \text{if} \quad p_K^{n+1} = p_L^{n+1}. \end{cases} \tag{3.6}$$

Lemma 3.4. *There exists C_6 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\iint_{Q_{t_f}} \Lambda \nabla \xi_{\mathcal{T},\Delta t} \cdot \nabla \xi_{\mathcal{T},\Delta t} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} (\xi_K^{n+1} - \xi_L^{n+1})^2 \leq C_6. \tag{3.7}$$

Proof. The definition (2.9) of the mobilities η_{KL}^{n+1} has been chosen so that

$$C_5 \geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \eta_{KL}^{n+1} (u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) \geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \check{\eta}_{KL}^{n+1} (u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}),$$

where $\check{\eta}_{KL}^{n+1} = \eta(s(p_{KL}))$ whatever $p_{KL} \in I_{KL}^{n+1}$. Therefore, using the definition (2.8b) of u_K^{n+1} and Young's inequality leads to

$$\begin{aligned} C_5 &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \check{\eta}_{KL}^{n+1} \left((p_K^{n+1} - p_L^{n+1})^2 + (p_K^{n+1} - p_L^{n+1})(z_K - z_L) \right) \\ &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \check{\eta}_{KL}^{n+1} (p_K^{n+1} - p_L^{n+1})^2 \\ &\quad - \frac{\alpha}{2} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}| \check{\eta}_{KL}^{n+1} (p_K^{n+1} - p_L^{n+1})^2 - \frac{\|\eta\|_\infty}{2\alpha} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}| (z_K - z_L)^2 \end{aligned}$$

where α is a positive parameter to be fixed. We choose $\check{\eta}_{KL}^{n+1} = \tilde{\eta}_{KL}^{n+1}$ defined in (3.6), leading to

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 - \frac{\alpha}{2}\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \leq C_5 + \frac{\|\eta\|_\infty}{2\alpha}\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(z_K - z_L\right)^2.$$

Using Lemma 2.1, we get that

$$\left(1 - \frac{\alpha C_3}{2}\right)\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \leq C_5 + \frac{\|\eta\|_\infty}{2\alpha} t_f C_3 |\Omega|.$$

We conclude the proof by setting $\alpha = \frac{1}{C_3}$. □

The function Γ takes non-negative values, hence so does the first term in (3.4). But since a_{KL} may become negative, we are not able to claim that the second term is non-negative (this would end the proof of Prop. 3.2). Nevertheless, we can prove that this term is uniformly bounded. This information, combined with Lemma 3.4, is sufficient to conclude the proof of Proposition 3.2.

Lemma 3.5. *There exists C_7 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}|u_K^{n+1} - u_L^{n+1}||p_K^{n+1} - p_L^{n+1}| \leq C_7.$$

Proof. Since $|x| = x + 2x^-$, $x^- = \max(-x, 0)$, one has

$$\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}|u_K^{n+1} - u_L^{n+1}||p_K^{n+1} - p_L^{n+1}| = \sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) + 2\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \eta_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^-. \tag{3.8}$$

It follows from the definition (2.9) of η_{KL}^{n+1} that

$$\eta_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^- \leq \tilde{\eta}_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^-,$$

with $\tilde{\eta}_{KL}^{n+1}$ defined by (3.6). Moreover, using the definition (2.8b) of u_K^{n+1} together with Young's inequality, we obtain that

$$\begin{aligned}
&\tilde{\eta}_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^- \\
&\leq \tilde{\eta}_{KL}^{n+1}|a_{KL}|(p_K^{n+1} - p_L^{n+1})^2 + \tilde{\eta}_{KL}^{n+1}|a_{KL}||z_K - z_L||p_K^{n+1} - p_L^{n+1}| \\
&\leq |a_{KL}|(\xi_K^{n+1} - \xi_L^{n+1})^2 + \frac{1}{2}|a_{KL}|\tilde{\eta}_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 + \frac{1}{2}|a_{KL}|\tilde{\eta}_{KL}^{n+1}(z_K - z_L)^2 \\
&\leq \frac{3}{2}|a_{KL}|(\xi_K^{n+1} - \xi_L^{n+1})^2 + \frac{\|\eta\|_\infty}{2}|a_{KL}|(z_K - z_L)^2.
\end{aligned}$$

We deduce from Lemma 2.1 that

$$2\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\eta_{KL}^{n+1}[a_{KL}(u_K^{n+1}-u_L^{n+1})(p_K^{n+1}-p_L^{n+1})]^- \le 3C_3\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}a_{KL}(\xi_K^{n+1}-\xi_L^{n+1})^2+C_3\|\eta\|_\infty t_f|\Omega|. \tag{3.9}$$

Then we combine (3.8), (3.9), Lemma 3.4, and Lemma 3.3 to conclude. □

The *a priori* estimate of Proposition 3.2 follows easily from Lemmas 3.3, 3.4, and 3.5. It is sufficient to prove the existence of a solution to the scheme (2.8) (see Sect. 3.3). Nevertheless, before going to this existence proof, we still provide some additional *a priori* estimates to be used later on in Section 4.

Lemma 3.6. *There exists C_8 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2\le C_8, \tag{3.10}$$

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|a_{KL}|\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1})^2\le C_8. \tag{3.11}$$

Proof. The definition (2.8b) of u_K^{n+1} yields

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma\in\mathcal{E}_{KL}}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2=A+B,$$

where

$$A=\sum_{n=0}^{N}\Delta t\sum_{\sigma\in\mathcal{E}_{KL}}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})(u_K^{n+1}-u_L^{n+1}),$$

$$B=-\sum_{n=0}^{N}\Delta t\sum_{\sigma\in\mathcal{E}_{KL}}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})(z_K^{n+1}-z_L^{n+1}).$$

Thanks to Lemma 3.5, one has $A\le C_7$. Moreover, combining once again Young inequality with Lemma 2.1, we get that

$$B\le\frac{1}{2}\sum_{n=0}^{N}\Delta t\sum_{\sigma\in\mathcal{E}_{KL}}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2+C_3\|\eta\|_\infty t_f|\Omega|,$$

hence (3.10) holds with $C_8=2C_7+2C_3\|\eta\|_\infty t_f|\Omega|$. The proof of (3.11) is similar. □

The last lemma of this section is devoted to the control of the L^2 norm of $\xi_{\mathcal{T},\Delta t}$. Lemma 3.4 only provides a control on the gradient of $\xi_{\mathcal{T},\Delta t}$, but not on $\xi_{\mathcal{T},\Delta t}$ directly. The control on $\xi_{\mathcal{T},\Delta t}$ is provided by an argument *à la* Poincaré, *cf.* Appendix A.1.

Lemma 3.7. *There exists C_9 depending only on $\Omega, t_f, s, \Lambda, \theta_{\mathcal{T}}, \eta, \overline{s}_0$, and $\xi_\star$ such that*

$$\|\xi_{\mathcal{T},\Delta t}\|_{L^2(Q_{t_f})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\le C_9, \tag{3.12}$$

$$\|\xi_{\mathcal{M},\Delta t}\|_{L^2(Q_{t_f})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\le C_9. \tag{3.13}$$

Proof. Let us first etablish (3.13). Thanks to assumption (1.4), we know that $\int_\Omega s_0 \,\mathrm{d}\boldsymbol{x} < \mathrm{meas}(\Omega)$. The global conservativity property (2.11) allows to claim that

$$\sum_{K\in\mathcal{V}} s_K^{n+1} m_K = \overline{s}_0 = \int_\Omega s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < \mathrm{meas}(\Omega)$$

for any $n \in \{0,\dots,N\}$. Using that $\xi_K^{n+1} < 0$ if and only if $s_K^{n+1} < 1$ (recall that $\xi(p) < 0$ iff $p < 0$ iff $s < 1$), one gets

$$\mathrm{meas}\left\{\xi_{\mathcal{M},\Delta t}(\cdot,t_{n+1}) < 0\right\} \geq \mathrm{meas}(\Omega) - \overline{s}_0 > 0. \tag{3.14}$$

Denote by $\xi_K^{+,n+1} = \max(0,\xi_K^{n+1})$, and by $\xi^+_{\mathcal{M},\Delta t}$ and $\xi^+_{\mathcal{T},\Delta t}$ the unique elements of $X_{\mathcal{M},\Delta t}$ and $V_{\mathcal{T},\Delta t}$ respectively such that

$$\xi^+_{\mathcal{M},\Delta t}(\boldsymbol{x}_K,t_{n+1}) = \xi^+_{\mathcal{T},\Delta t}(\boldsymbol{x}_K,t_{n+1}) = \xi_K^{+,n+1}, \qquad \forall K\in\mathcal{V},\ \forall n\in\{0,\dots,N\}.$$

Note that $\xi^+_{\mathcal{T},\Delta t} \neq (\xi_{\mathcal{T},\Delta t})^+ = \max(0,\xi_{\mathcal{T},\Delta t})$ in general, but that $\xi^+_{\mathcal{M},\Delta t} = (\xi_{\mathcal{M},\Delta t})^+$ and that $\xi^-_{\mathcal{M},\Delta t} = (\xi_{\mathcal{M},\Delta t})^- = \max(0,-\xi_{\mathcal{M},\Delta t})$. Using assumption (**A3**), the 1-Lipschitz continuity of $x\mapsto x^+$, and Lemmas 2.1 and 3.4, we obtain

$$\begin{aligned}
\iint_{Q_{t_f}} |\nabla \xi^+_{\mathcal{T},\Delta t}|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t &\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi_K^{+,n+1} - \xi_L^{+,n+1}\right)^2 \\
&\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(\xi_K^{+,n+1} - \xi_L^{+,n+1}\right)^2 \\
&\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \leq \frac{C_3C_6}{\underline{\Lambda}}.
\end{aligned}$$

Therefore, we can apply Lemma A.3 stated in appendix. This provides

$$\iint_{Q_{t_f}} \left(\xi^+_{\mathcal{M},\Delta t}\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C. \tag{3.15}$$

On the other hand, because of (1.3), we know that $\xi^-_{\mathcal{M},\Delta t} \leq \xi_\star$ a.e. in Q_{t_f}, hence

$$\iint_{Q_{t_f}} \left(\xi^-_{\mathcal{M},\Delta t}\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq (\xi_\star)^2\,\mathrm{meas}(\Omega)t_f. \tag{3.16}$$

Combining (3.15) with (3.16) provides (3.13). In order to recover (3.12), in only remains to use (2.1) and (3.13). □

3.3. Existence of a discrete solution

In order to prove the existence of a solution $(p_K^{n+1})_K$ to the scheme (2.8), we need an additional mesh-depending estimate on the solution. Following [11], we introduce now the notion of *transmissive path*.

Definition 3.8. A transmissive path w joining $K_i \in \mathcal{V}$ to $K_f \in \mathcal{V}$ consists in a list of vertices $(K_q)_{0\leq q\leq M}$ such that $K_i = K_0, K_f = K_M$, with $K_q \neq K_\ell$ if $q \neq \ell$, and such that $\sigma_{K_qK_{q+1}} \in \mathcal{E}$ with $a_{K_qK_{q+1}} > 0$ for all $q \in \{0,\dots,M-1\}$. We denote by $\mathcal{W}(K_i,K_f)$ the set of the transmissive path joining $K_i \in \mathcal{V}$ to $K_f \in \mathcal{V}$.

We now state a result which is proved in ([11], Lem. 3.5).

Lemma 3.9. *For all $(K_i,K_f) \in \mathcal{V}^2$ there exists a transmissive path $w \in \mathcal{W}(K_i,K_f)$.*

Lemma 3.10. *There exists $C_\star > -\infty$ depending only on $\mathcal{T}, \Delta t, \Omega, s, \overline{s}_0, t_f, \Lambda, \theta_{\mathcal{T}}$, η and z such that*

$$p_K^{n+1} \geq C_\star, \qquad \forall K \in \mathcal{V}, \quad \forall n \in \{0, \dots, N\}.$$

Proof. Let us prove that $p_K^{n+1} \geq C_\star$. Assume first that $p_\star > -\infty$, then we can choose $C_\star = p_\star$ thanks to (3.2), so that we can focus on the case $p_\star = -\infty$.

In view of the global conservation property (2.11), one has that

$$\sum_{K \in \mathcal{V}} (s_K^{n+1} - \overline{s}_0) m_K = 0.$$

This ensures the existence of at least one vertex K_i such that $s_{K_i}^{n+1} \geq \overline{s}_0 > 0$. In particular,

$$-\infty < s^{-1}(\overline{s}_0) \leq p_{K_i}^{n+1}. \tag{3.17}$$

Let $K_f \in \mathcal{V} \backslash \{K_i\}$, then thanks to Lemma 3.9, there exists a transmissive path $w \in \mathcal{W}(K_i, K_f) = (K_q)_{0 \leq q \leq M}$ of finite length in the sense of Definition 3.8. Let us show that for all $p_{K_q}^{n+1} > -\infty$ for all $q \in \{0, \dots, M\}$.

First, we have checked in (3.17) that $p_{K_0}^{n+1} > -\infty$. Assume now that $p_{K_q}^{n+1} > -\infty$ for some $q \in \{0, \dots, M-1\}$, then it follows from Lemma 3.5 that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}| \eta_{KL}^{n+1} |u_K^{n+1} - u_L^{n+1}| |p_K^{n+1} - p_L^{n+1}| \leq C_6.$$

This ensures in particular that

$$\Delta t a_{K_q K_{q+1}} \eta_{K_q K_{q+1}}^{n+1} (u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \leq C_6.$$

Thanks to the definition (2.9) of $\eta_{K_q K_{q+1}}^{n+1}$, one has

$$a_{K_q K_{q+1}} \eta_{K_q K_{q+1}}^{n+1} (u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \geq a_{K_q K_{q+1}} \eta(s_{K_q}^{n+1})(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}).$$

Since $a_{K_q K_{q+1}} > 0$, we obtain that

$$(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) = (p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})^2 + (p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})(z_{K_q} - z_{K_{q+1}}) \ \leq \frac{C_6}{\Delta t a_{K_q K_{q+1}} \eta(s(p_{K_q}^{n+1}))}.$$

Using Young inequality one has

$$(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \geq \frac{1}{2}(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})^2 - \frac{1}{2}(z_{K_q} - z_{K_{q+1}})^2,$$

thus

$$p_{K_{q+1}}^{n+1} \geq p_{K_q}^{n+1} - \sqrt{(z_{K_q} - z_{K_{q+1}})^2 + \frac{2C_6}{\Delta t a_{K_q K_{q+1}} \eta(s(p_{K_q}^{n+1}))}}.$$

This ensures that $p_{K_{q+1}}^{n+1} > -\infty$.

We have proved the existence of a finite quantity $(C_{K_i, K_f, w})_{K_f \in \mathcal{V}}$ (depending on the data of the continuous problem $\Omega, s, \overline{s}_0, t_f, \Lambda, \theta_{\mathcal{T}}, \eta$ but also on the discretization $\mathcal{T}$ and on Δt) such that

$$s(p_{K_i}^{n+1}) \geq \overline{s}_0 \quad \Longrightarrow \quad p_{K_f}^{n+1} \geq -C_{K_i, K_f, w}.$$

As a consequence, since there exists a finite number of transmissive paths between two vertices, we get the estimate

$$p_K^{n+1} \geq -\max_{K_i \in \mathcal{V}} \max_{K_f \in \mathcal{V}} \min_{w \in \mathcal{W}(K_i, K_f)} C_{K_i, K_f, w} > -\infty, \quad \forall K \in \mathcal{V}, \ \forall n \in \{0, \dots, N\}. \qquad \square$$

In the previous lemma, we managed to bound the $\{p_K^{n+1}\}$ from below. The next lemma provides a bound from above.

Lemma 3.11. *There exists $p^\star < \infty$ depending only on $\mathcal{T}, \Delta t, \Omega, t_f, s, \Lambda, \eta, \overline{s}_0$ and $\xi_\star$ such that*

$$p_K^{n+1} \le p^\star \quad \forall K \in \mathcal{V}, \ \ \forall n \in \{0, \dots, N\}.$$

Proof. Since $s(p) = 1$ if $p \ge 0$, one has $\xi(p) = p\sqrt{\eta(1)}$ if $p \ge 0$. By (3.13), one has

$$\Delta t m_K \xi(p_K^{n+1})^2 \le \|\xi_{\mathcal{M},\Delta t}\|^2_{L^2(Q_{t_f})} \le (C_9)^2.$$

Therefore, we get $p_K^{n+1} \le \dfrac{C_9}{\sqrt{\Delta t m_K}} \dfrac{1}{\eta(1)}$. □

Now, one can apply the same strategy as in ([11], Lem. 3.11) for proving the existence of a solution to the scheme (2.8).

Proposition 3.12. *Let $(s_K^n)_{K\in\mathcal{V}} \in [0,1]^{\#\mathcal{V}}$ be such that $\sum_{K\in\mathcal{V}} m_K s_K^n = \mathrm{meas}(\Omega)\overline{s}_0$, there exists (at least) one solution $(p_K^{n+1})_{K\in\mathcal{V}} \in [p_\star, p^\star]^{\#\mathcal{V}}$ of the scheme (2.8). Moreover, it satisfies $\sum_{K\in\mathcal{V}} m_K s_K^{n+1} = \mathrm{meas}(\Omega)\overline{s}_0$.*

The proof of Proposition 3.12 is not detailed here since it mimics the one of ([11], Lem. 3.11). Let us just mention that it is based on a topological degree argument [14, 28].

4. Convergence towards a weak solution

The proof of the convergence properties stated in Theorem 2.5 is based on compactness arguments. As a first step, we show in Section 4.1 the appropriate compactness properties on the reconstructed discrete solutions. Then we identify in Section 4.2 the limit value (whose existence is ensured thanks to the compactness properties) as the unique weak solution to the problem (1.1).

4.1. Compactness properties of discrete solutions

As it is classical for unsteady problems, we need to prove some time-compactness for the approximate solutions. Because of the degeneracy of the problem we consider, we cannot use a strategy *à la* Aubin-Simon [37] (see [23] for an extension of this strategy to the discrete setting). A classical way to circumvent this problem is to estimate the time-translates (see [3] in the continuous setting and [17] in the discrete setting). This strategy could have been used here, but we rather make use of the time-compactness result for degenerate parabolic equations proposed in [4]. To this end, we need the following lemma.

Lemma 4.1. *There exists C_{10} depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}, z$ and η such that*

$$\sum_{n=0}^{N} \sum_{K\in\mathcal{V}} (s_K^{n+1} - s_K^n)\psi(\boldsymbol{x}_K, t_{n+1}) m_K \le C_{10} \|\nabla\psi\|_\infty, \quad \forall\, \psi \in \mathcal{C}_c^\infty(Q_{t_f}). \tag{4.1}$$

Proof. For the sake of readability, we denote by $\psi_K^{n+1} = \psi(\boldsymbol{x}_K, t_{n+1})$ for all $K \in \mathcal{V}$ and all $n \in \{0, \dots, M\}$. We multiply (2.8a) by $\Delta t \psi_K^{n+1}$ and sum for $K \in \mathcal{V}$, for $n \in \{0, \dots, N\}$. This yields

$$A = B,$$

where

$$A = \sum_{n=0}^{N} \sum_{K\in\mathcal{V}} m_K(s_K^{n+1} - s_K^n)\psi_K^{n+1},$$
$$B = -\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(\psi_K^{n+1} - \psi_L^{n+1}).$$

Using the Cauchy−Schwarz inequality, we get

$$|B|^2 \le \left(\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})^2\right) \times \left(\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}(\psi_K^{n+1} - \psi_L^{n+1})^2\right).$$

Using Lemma 3.6, the boundedness of η and Lemma 2.1, we obtain that

$$|B|^2 \le \|\eta\|_\infty C_8 C_3 \iint_{Q_{t_f}} \Lambda \nabla\psi_{\mathcal{T},\Delta t} \cdot \nabla\psi_{\mathcal{T},\Delta t} \le \|\eta\|_\infty C_8 C_3 \mathrm{meas}(\Omega) t_f \overline{\Lambda} \|\nabla\psi\|_\infty^2.$$

Therefore (4.1) holds with $C_{10} = \sqrt{\|\eta\|_\infty C_8 C_3 \mathrm{meas}(\Omega) t_f \overline{\Lambda}}$. □

We can now state the expected compactness properties.

Proposition 4.2. *There exists a measurable function $p : Q_{t_f} \longrightarrow [p_\star, p^\star]$ such that, up to an unlabeled subsequence, one has*

$$s_{\mathcal{M}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} s(p) \quad \text{a.e in } Q_{t_f},$$
$$\xi_{\mathcal{T}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} \xi(p) \quad \text{weakly in } L^2((0,t_f);H^1(\Omega)).$$

Proof. Thanks to (3.7), the sequence $(\nabla\xi_{\mathcal{T}_m,\Delta t_m})_{m\ge1}$ is bounded in $(L^2(Q_{t_f}))^2$. Moreover, it follows from (3.12) that $(\xi_{\mathcal{T}_m,\Delta t_m})_{m\ge1}$ is uniformly bounded in $L^2(Q_{t_f})$, providing the boundedness of $(\xi_{\mathcal{T}_m,\Delta t_m})_{m\ge1}$ in $L^2((0,t_f);H^1(\Omega))$. Therefore, there exists $\Xi \in L^2((0,t_f);H^1(\Omega))$ such that

$$\xi_{\mathcal{T}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} \Xi \quad \text{weakly in } L^2((0,t_f);H^1(\Omega)).$$

By (3.1) we obtain directly that $0 \le s_{\mathcal{M}_m,\Delta t_m} \le 1$, ensuring the L^∞-weak-$\star$ convergence of an unlabeled subsequence towards $s \in L^\infty(Q_{t_f};[0,1])$. Thanks to Lemma 4.1, we can apply ([4], Thm. 3.9). It gives the existence of $p : Q_{t_f} \longrightarrow [p_\star, p^\star]$ such that, up to an unlabeled subsequence,

$$s_{\mathcal{M}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} s(p) \quad \text{a.e in } Q_{t_f},$$

and $\Xi = \xi(p)$. □

4.2. Identification as a weak solution

Proposition 4.3. *Let p be as in Proposition 4.2, then p is the unique weak solution to (1.1) in the sense of Definition 1.4.*

Proof. Let $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_f))$, and denote by $\psi_K^n = \psi(\boldsymbol{x}_K, t_n)$, for all $K \in \mathcal{V}_m$ and all $n \in \{0, \dots, N_m\}$. We multiply (2.8a) by $\Delta t_m \psi_K^n$ and sum over $n \in \{0, \dots, N_m\}$ and $K \in \mathcal{V}_m$ to obtain

$$A_m + B_m + C_m + D_m = 0, \tag{4.2}$$

where, denoting by $\xi_K^{n+1} = \xi(p_K^{n+1})$, we have set

$$\begin{aligned}
A_m &= \sum_{n=0}^{N_m} \sum_{K \in \mathcal{V}_m} (s_K^{n+1} - s_K^n) \psi_K^n m_K, \\
B_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \left(\eta_{KL}^{n+1} (p_K^{n+1} - p_L^{n+1}) - \sqrt{\eta_{KL}^{n+1}} (\xi_K^{n+1} - \xi_L^{n+1}) \right) (\psi_K^n - \psi_L^n), \\
C_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \sqrt{\eta_{KL}^{n+1}} \, (\xi_K^{n+1} - \xi_L^{n+1}) \, (\psi_K^n - \psi_L^n), \\
D_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \eta_{KL}^{n+1} \, (z_K - z_L) \, (\psi_K^n - \psi_L^n).
\end{aligned}$$

Note that $\psi_K^{N_m+1} = 0$ for all $K \in \mathcal{V}_m$, then a discrete integration parts yields

$$\begin{aligned}
A_m &= - \sum_{n=0}^{N_m} \Delta t_m \sum_{K \in \mathcal{V}_m} s_K^{n+1} \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m} m_K - \sum_{K \in \mathcal{V}_m} s_K^0 \psi_K^0 m_K \\
&= - \iint_{Q_{t_f}} s_{\mathcal{M}_m, \Delta t_m} \delta\psi_{\mathcal{M}_m, \Delta t_m} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \int_\Omega s_{\mathcal{M}_m}^0 \psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x},
\end{aligned}$$

where the function $\delta\psi_{\mathcal{M}_m, \Delta t_m}$ of $X_{\mathcal{M}_m, \Delta t_m}$ is defined by

$$\delta\psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, t) = \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m} \quad \text{if } (\boldsymbol{x}, t) \in \omega_K \times (t_n, t_{n+1}).$$

Thanks to the regularity of ψ, the function $\delta\psi_{\mathcal{M}_m, \Delta t_m}$ converges uniformly towards $\partial_t \psi$ on Q_{t_f}. Moreover, we have

$$s_{\mathcal{M}_m, \Delta t_m} \longrightarrow s(p) \quad \text{in } L^r(Q_{t_f}) \text{ as } m \to \infty,$$

for all $r \in [1, \infty)$ thanks to Proposition 4.2. Therefore,

$$\iint_{Q_{t_f}} s_{\mathcal{M}_m, \Delta t_m} \delta\psi_{\mathcal{M}_m, \Delta t_m} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \longrightarrow \iint_{Q_{t_f}} s(p) \partial_t \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \quad \text{as } m \to \infty. \tag{4.3}$$

Moreover, $s_{\mathcal{M}_m}^0$ converges strongly in $L^1(\Omega)$ towards the initial data s_0 and $\psi_{\mathcal{M}_m, \Delta t_m}(\cdot, 0)$ converges uniformly towards $\psi(\cdot, 0)$. Therefore, we get that

$$\int_\Omega s_{\mathcal{M}_m}^0(\boldsymbol{x}) \psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x} \longrightarrow \int_\Omega s_0(\boldsymbol{x}) \psi(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty. \tag{4.4}$$

We deduce from (4.3) and (4.4) that

$$A_m \longrightarrow - \iint_{Q_{t_f}} s(p) \partial_t \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \int_\Omega s_0 \psi(., 0) \, \mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty. \tag{4.5}$$

The term B_m rewrites

$$B_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} a_{KL}\sqrt{\eta_{KL}^{n+1}}\left(\sqrt{\eta_{KL}^{n+1}} - \sqrt{\tilde{\eta}_{KL}^{n+1}}\right)(p_K^{n+1} - p_L^{n+1})(\psi_K^n - \psi_L^n),$$

where $\tilde{\eta}_{KL}^{n+1}$ is defined by (3.6). Using the Cauchy−Schwarz inequality, we get

$$|B_m|^2 \le \left(\sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2\right) \times \underbrace{\left(\sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} |a_{KL}|\left(\sqrt{\eta_{KL}^{n+1}} - \sqrt{\tilde{\eta}_{KL}^{n+1}}\right)^2 (\psi_K^n - \psi_L^n)^2\right)}_{:=R_m}. \tag{4.6}$$

The first term in the right-hand side of (4.6) is bounded by C_8 thanks to Lemma 3.6. Therefore, in order to prove that $\lim_{m\to\infty} B_m = 0$, it suffices to prove that $\lim_{m\to\infty} R_m = 0$. For $T \in \mathcal{T}_m$, we denote by

$$\overline{\xi}_T^{n+1} = \max_{\boldsymbol{x}\in T}\left(\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t_{n+1})\right), \qquad \underline{\xi}_T^{n+1} = \min_{\boldsymbol{x}\in T}\left(\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t_{n+1})\right),$$

and we define the piecewise constant functions $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}$ and $\underline{\xi}_{\mathcal{T}_m,\Delta t_m}$ by

$$\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t) = \overline{\xi}_T^{n+1} \quad \text{and} \quad \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t) = \underline{\xi}_T^{n+1} \quad \text{if } (\boldsymbol{x}, t) \in T \times (t_n, t_{n+1}),$$

We can estimate

$$\left|\sqrt{\eta_{KL}^{n+1}} - \sqrt{\tilde{\eta}_{KL}^{n+1}}\right| \le \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right), \quad \forall \sigma_{KL} \in \mathcal{E}_T, \tag{4.7}$$

where μ is the continuity modulus defined in (1.9). Using (4.7) in the definition (4.6) of R_m, we get

$$0 \le R_m \le \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(\psi_K^n - \psi_L^n)^2. \tag{4.8}$$

Following the proof of Lemma 2.1 (*cf.* [11], Lem. 3.2), we can prove that

$$\sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(\psi_K^n - \psi_L^n)^2 \le C_3\overline{\Lambda}\|\nabla\psi\|_\infty^2 \mathrm{meas}(T), \qquad \forall T \in \mathcal{T}. \tag{4.9}$$

Therefore, we deduce from (4.8) that

$$0 \le R_m \le C \iint_{Q_{t_f}} \mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m} - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}t, \tag{4.10}$$

where C depends only on Λ, θ and ψ. Since μ is bounded (as η is bounded), continuous, with $\mu(0) = 0$, it suffices to show that $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t) - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t)$ tends to 0 almost everywhere in Q_{t_f} as $m \to \infty$ (up to an unlabeled subsequence). Thanks to ([11], Lem. A.1) and to Lebesgue's dominated convergence theorem, one has

$$\iint_{Q_{t_f}} \left|\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t) - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}, t)\right| \mathrm{d}\boldsymbol{x}\,\mathrm{d}t \le Ch_m \iint_{Q_{t_f}} |\nabla\xi_{\mathcal{T}_m,\Delta t_m}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \underset{m\to+\infty}{\longrightarrow} 0. \tag{4.11}$$

As a consequence of (4.6), (4.10) and (4.11), and still up to the extraction of an unlabeled subsequence, one has

$$\lim_{m\to\infty} B_m = \lim_{m\to\infty} R_m = 0. \tag{4.12}$$

Let us now focus on the term C_m. We define the piecewise constant functions $\Xi_{\mathcal{T}_m,\Delta t_m}$ and $H_{\mathcal{T}_m,\Delta t_m}$ by

$$\Xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) = \xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_T,t), \quad \forall \boldsymbol{x}\in T, \forall t\in(t_n,t_{n+1}),$$

$\boldsymbol{x}_T$ being the center of mass of the triangle T, and by $H_{\mathcal{T}_m,\Delta t_m} = \eta\circ s\circ \xi^{-1}(\Xi_{\mathcal{T}_m,\Delta t_m})$. Clearly, one has

$$\underline{\xi}_{\mathcal{T}_m,\Delta t_m} \le \Xi_{\mathcal{T}_m,\Delta t_m} \le \overline{\xi}_{\mathcal{T}_m,\Delta t_m}.$$

It follows from (4.11) that both $\underline{\xi}_{\mathcal{T}_m,\Delta t_m}$ and $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}$ converge almost everywhere to $\xi(p)$, hence so does $\Xi_{\mathcal{T}_m,\Delta t_m}$. This provides that

$$H_{\mathcal{T}_m,\Delta t_m} \longrightarrow \eta(s(p)) \quad \text{ in } L^1(Q_{t_f}) \text{ as } m\to\infty. \tag{4.13}$$

We introduce the term

$$C'_m = \iint_{Q_{t_f}} \sqrt{H_{\mathcal{T}_m,\Delta t_m}}\Lambda\nabla\xi_{\mathcal{T}_m,\Delta t_m}\cdot\nabla\psi_{\mathcal{T}_m,\Delta t_m}(.,t-\Delta t_m)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

Since $\nabla\xi_{\mathcal{T}_m,\Delta t_m}$ converges weakly in $L^2(Q_{t_f})$ towards $\nabla\xi(p)$, since $\nabla\psi_{\mathcal{T}_m,\Delta t_m}$ converges uniformly towards $\nabla\psi$, and thanks to (4.13), we obtain that

$$\lim_{m\to\infty} C'_m = \iint_{Q_{t_f}} \sqrt{\eta(s(p))}\Lambda\nabla\xi(p)\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{4.14}$$

Let us now check that $|C_m - C'_m|$ tends to 0 as m tends to ∞. We denote by

$$\eta_T^{n+1} = H_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_T,t_{n+1}), \quad \forall T\in\mathcal{T}_m, \forall n\in\{0,\dots,N_m\}.$$

The term C'_m can be rewritten

$$C'_m = \sum_{n=0}^{N_m}\Delta t_m \sum_{T\in\mathcal{T}_m}\sqrt{\eta_T^{n+1}}\sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T\left(\xi_K^{n+1}-\xi_L^{n+1}\right)(\psi_K^n-\psi_L^n),$$

so that

$$C_m - C'_m = \sum_{n=0}^{N_m}\Delta t_m \sum_{T\in\mathcal{T}_m}\sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\eta_T^{n+1}}\right)\left(\xi_K^{n+1}-\xi_L^{n+1}\right)(\psi_K^n-\psi_L^n).$$

For all $n\in\{0,\dots,N_m\}$, for all $T\in\mathcal{T}_m$, and for all $\sigma_{KL}\in\mathcal{E}_T$, one has

$$\left|\sqrt{\eta_{KL}^{n+1}}-\sqrt{\eta_T^{n+1}}\right| \le \mu(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}) \tag{4.15}$$

where μ is the continuity modulus defined in (1.9). Then one obtains that

$$|C_m - C'_m| \le \sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\left[\mu(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1})\times\sum_{\sigma_{KL}\in\mathcal{E}_T}|a_{KL}^T|\left|\xi_K^{n+1}-\xi_L^{N+1}\right|\left|\psi_K^n-\psi_L^n\right|\right].$$

The Cauchy-Schwarz inequality provides

$$|C_m - C'_m|^2 \leq \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(\psi_K^n - \psi_L^n)^2$$
$$\times \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2.$$

Using Lemmas 2.1 and 3.4, together with (4.9), one deduces that

$$|C_m - C'_m|^2 \leq C \iint_{Q_{t_f}} \mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m} - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}t,$$

thus $|C_m - C'_m|$ tends to 0 thanks to the arguments already developed to prove that R_m tends to 0. Finally, we obtain that

$$\lim_{m\to\infty} C_m = \iint_{Q_{t_f}} \sqrt{\eta(s(p))}\Lambda\nabla\xi(p)\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{4.16}$$

Let us focus on the last term D_m. We introduce the term

$$D'_m = \iint_{Q_{t_f}} H_{\mathcal{T}_m,\Delta t_m}\Lambda\nabla z\cdot\nabla\psi_{\mathcal{T}_m,\Delta t_m}(.,t-\Delta t_m)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

It follows from (4.13) and from the uniform convergence of $\nabla\psi_{\mathcal{T}_m,\Delta t_m}$ towards $\nabla\psi$ as m tends to $+\infty$ that

$$\lim_{m\to\infty} D'_m = \iint_{Q_{t_f}} \eta(s(p))\Lambda\nabla z\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

We will now check that $|D_m - D'_m| \longrightarrow 0$ as $m\to\infty$. The term D'_m rewrites

$$D'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \eta_T^{n+1} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T\,(z_K - z_L)\,(\psi_K^n - \psi_L^n),$$

so that

$$D_m - D'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T\left(\eta_{KL}^{n+1} - \eta_T^{n+1}\right)(z_K - z_L)\,(\psi_K^n - \psi_L^n).$$

For all $\sigma_{KL}\in\mathcal{E}_T$, one has

$$\left|\eta_{KL}^{n+1} - \eta_T^{n+1}\right| \leq \left|\sqrt{\eta_{KL}^{n+1}} - \sqrt{\eta_T^{n+1}}\right|\left(\sqrt{\eta_{KL}^{n+1}} + \sqrt{\eta_T^{n+1}}\right) \leq 2\|\eta\|_\infty \mu(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}).$$

Therefore, using the Cauchy-Schwarz inequality, one has

$$|D_m - D'_m|^2 \leq 4\|\eta\|_\infty^2 \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(\psi_K^n - \psi_L^n)^2$$
$$\times \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(z_K - z_L)^2.$$

We use Lemma 2.1 to get

$$\sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(z_K - z_L)^2 \leq C_3 \iint_{Q_{t_f}} \Lambda\nabla z\cdot\nabla z \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \leq C.$$

We deduce from (4.9) that

$$|D_m - D'_m|^2 \leq C \iint_{Q_{t_f}} \mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m} - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}t \underset{m\to\infty}{\longrightarrow} 0,$$

and then that

$$\lim_{m\to\infty} D_m = \iint_{Q_{t_f}} \eta(s(p))\Lambda\nabla z\cdot\nabla\psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{4.17}$$

Putting (4.5), (4.12), (4.16) and (4.17) together in (4.2) provides that p satisfies the weak formulation (1.13), then it is the unique weak solution to the problem (*cf.* Thm. 1.5). □

Finally, let us remark that since the weak solution p is unique, all the convergence in functional space that were proved to occur up to the extraction of a subsequence are valid for the whole sequences. Concerning the almost everywhere convergence, we cannot do better than saying that it holds up to a subsequence.

5. Numerical results

Let us provide some illustrations of the behavior of the numerical scheme (2.8). The scheme leads to a nonlinear system that we solve thanks to the Newton-Raphson method with Matlab. As proved in Proposition 3.1, the approximate pressure remains greater than $p_\star$. Therefore, we project the discrete solution at each Newton iteration on the set $\{p \geq p_\star\}$. We refer to [7,29] for a study on iterative methods for solving Richard's equation.

In all our test cases, the domain is the unit square, *i.e.*, $\Omega = (0,1)^2$. We use meshes coming from the 2D benchmark on anisotropic diffusion problems [24]. An illustration of the meshes is given in Figure 4. These triangle meshes show no symmetry which could artificially increase the convergence rate. All angles are acute, so that, in the case of an isotropic tensor Λ, the coefficients a_{KL} are all non-negative. This is no longer the case when Λ is chosen to be anisotropic. To be more precise concerning the diffusion tensor, we have considered constant diagonal tensors

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & 0 \\ 0 & \Lambda_{yy} \end{pmatrix}$$

where Λ_{xx} and Λ_{yy} are chosen constant in Ω, and the gravity acceleration $\mathbf{g}$ is defined by $\mathbf{g} = (g,0)^T$ for all $\boldsymbol{x} \in \Omega$ with $g \in \mathbb{R}_+$.

The numerical analysis of the scheme was carried out for a uniform time discretization of $(0,t_f)$ only in order to avoid heavy notations. In order to increase the robustness of the algorithm and to ensure the convergence of the Newton-Raphson iterative procedure, we used an adaptive time step procedure in the practical implementation. More precisely, to each mesh, we associate a maximal time step Δt_k, k being the index of the mesh (1 for the coarsest, 8 for the finest). If the Newton-Raphson method fails to converge after 30 iterations —we choose that the ℓ^∞ norm of the residual has to be smaller than 10^{-7} as stopping criterion—, the time step is divided by two. If the Newton-Raphson method converges, the time step is multiplied by two and projected on $[0,\Delta t_k]$.

In Sections 5.1, 5.2 and 5.3, we give evidence of the convergence of scheme (2.8) on test cases where exact analytical solutions are known. We are interested in the convergence speed of our method when the discretization parameters h and Δt tend to 0. We focus on the error caused by the spatial discretization (the time discretization is a classical first order accurate backward Euler method). As we will see, our scheme is at most first order accurate. In order to be sure that the error caused by the time discretization will not be of leading order, we

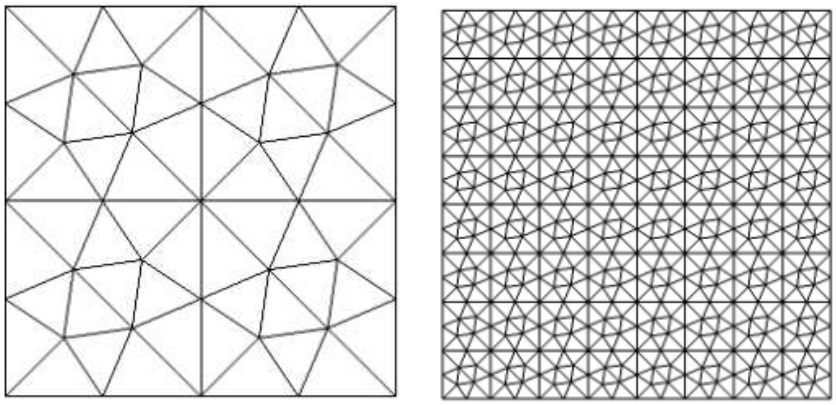

FIGURE 4. Second and fourth meshes used in the numerical tests.

choose $\Delta t_{k+1} = \Delta t_k/4$ while $h_{k+1} = h_k/2$, h_k being the size the mesh $k \in \{1, \dots, 8\}$. The first time step Δt_1 to 0.01024 in all the test cases presented below.

The test cases we chose to present here do not perfectly match with the assumptions presented at the beginning of the paper. They rather isolate the difficulties of the problem and give a better view of the behavior of the scheme. More precisely, the so called Hornung-Messing problem presented in Section 5.1 aims to illustrate the behavior of the scheme when an elliptic degeneracy occurs. The linear Fokker-Planck problem of Section 5.2 illustrates the behavior of the scheme for a stiff problem when $p_\star = -\infty$. The porous medium equation with drift presented in Section 5.3 allows to illustrate the behavior of the scheme near a hyperbolic degeneracy at $s(p) = 0$. The test case presented in Section 5.4 is there to illustrate numerically the decay of the free energy. Finally, we illustrate the behavior of Newton's method in Section 5.5. Let us stress that the numerical analysis we developed in the paper can be adapted without any major modification to all the cases we present here.

In the case where $p_\star = -\infty$, it was proved in Lemma 3.10 that there exists $C_\star > -\infty$ depending only on $\mathcal{T}, \Delta t, \Omega, s, \overline{s}_0, t_f, \Lambda, \theta_{\mathcal{T}}, \eta$ and z such that

$$p_K^{n+1} \geq C_\star, \qquad \forall K \in \mathcal{V}, \quad \forall n \in \{0, \dots, N\}.$$

Therefore we initialize the Newton-Raphson algorithm by

$$p_K^{n+1,0} = \max(s^{-1}(\epsilon), p_K^n), \qquad \text{where} \quad \epsilon = 10^{-14}.$$

Let us mention that in the tests 2, 3, and 4, we considered problems without elliptic degeneracy. The corresponding functions s are increasing on $(p_\star, +\infty)$. Therefore, we can choose $S = s(p)$ rather that p as a primary unknown in these cases. Denoting by $p = s^{-1}$, the problem solved numerically in Section 5.2, Sections 5.3 and 5.4 can then be written

$$\partial_t S - \nabla \cdot (\Lambda \eta(S)(\nabla p(S) - \mathbf{g})) = 0 \qquad \text{in} \quad Q_{t_f}. \tag{5.1}$$

Finally, we have set the gravity $\mathbf{g} = \mathbf{e}_x$ horizontal from the left to the right in the tests 2, 3, and 4. As a consequence, the scheme we considered in Section 5.2, Sections 5.3, and 5.4 is

$$\frac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} \eta_{KL}^{n+1} a_{KL}(u_K^{n+1} - u_L^{n+1}) = 0, \tag{5.2a}$$

$$u_K^{n+1} = p_K^{n+1} - x_K, \tag{5.2b}$$

$$p_K^{n+1} = p(s_K^{n+1}), \tag{5.2c}$$

$$\eta_{KL}^{n+1} = \begin{cases} \eta(s_K^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) \geq 0, \\ \eta(s_L^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) < 0. \end{cases} \tag{5.2d}$$

5.1. Test 1: A test case with saturated zones

The first test-case we propose here is the so-called Hornung-Messing problem [25]. In this problem, gravity is neglected (*i.e.* $g = 0$ and $u_K^{n+1} = p_K^{n+1}$ for all $K \in \mathcal{V}$ and $n \geq 0$). We consider the following nonlinearities

$$\eta(p) = \begin{cases} \dfrac{2}{1+p^2} & \text{if} \quad p < 0, \\ 2 & \text{if} \quad p \geq 0, \end{cases} \qquad s(p) = \begin{cases} \left(\dfrac{\pi^2}{4} - \arctan^2(p)\right)(\Lambda_{xx} + \Lambda_{yy}) & \text{if} \quad p < 0, \\ \dfrac{\pi^2}{4}(\Lambda_{xx} + \Lambda_{yy}) & \text{if} \quad p \geq 0. \end{cases}$$

and the exact solution to the Richards equation

$$p_{\text{ex}} = \begin{cases} -\dfrac{x-y-t}{2} & \text{if } x-y-t < 0, \\ -\tan\left(\dfrac{\mathrm{e}^{x-y-t}-1}{\mathrm{e}^{x-y-t}+1}\right) & \text{if } x-y-t \geq 0, \end{cases} \qquad \forall (x,y) \in \Omega, \ \forall t \in (0, t_f), \tag{5.3}$$

where t_f was set to 0.05. This exact solution does not satisfies the no-flux boundary conditions. Therefore, we prescribe the exact solution p_{ex} as Dirichlet boundary conditions on $\partial\Omega \times (0, t_f)$. In Tables 1 and 2, we report the errors

$$err_{L^p} = \|p_{\mathcal{M},\Delta t} - p_{\text{ex}}\|_{L^p(Q_{t_f})} \quad \text{for } p = 1, 2, \infty$$

for 7 successively refined meshes in the Isotropic case $\Lambda = \text{Id}$ and in the anisotropic case $\Lambda = \text{diag}(1, 10^{-3})$.

We observe that numerical order of convergence is close to 1 for the three norms whatever the anisotropy tensor on this test case.

TABLE 1. Test 1, isotropic case $\Lambda = \text{Id}$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate
0.500	12	0.343E-3	–	0.548E-4	–	0.352E-2	–
0.250	37	0.218E-3	0.651	0.472E-4	0.215	0.197E-2	0.838
0.125	129	0.141E-3	0.629	0.329E-4	0.522	0.113E-2	0.801
0.063	481	0.769E-4	0.886	0.185E-4	0.844	0.607E-3	0.907
0.031	1857	0.399E-4	0.927	0.967E-5	0.912	0.306E-3	0.966
0.016	7297	0.202E-4	1.025	0.493E-5	1.019	0.154E-3	1.041
0.008	28 929	0.102E-4	0.989	0.249E-5	0.986	0.771E-4	0.996
0.004	115 201	0.512E-5	0.994	0.125E-5	0.993	0.386E-4	0.997

TABLE 2. Test 1, anisotropic case with $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 10^{-3}$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate
0.500	12	0.382E-3	–	0.581E-4	–	0.384E-2	–
0.250	37	0.368E-3	0.057	0.682E-4	−0.231	0.396E-2	−0.044
0.125	129	0.225E-3	0.710	0.475E-4	0.522	0.218E-2	0.861
0.063	481	0.120E-3	0.911	0.268E-4	0.838	0.112E-2	0.974
0.031	1857	0.621E-4	0.933	0.141E-4	0.904	0.522E-3	1.075
0.016	7297	0.315E-4	1.026	0.721E-5	1.012	0.260E-3	1.052
0.008	28 929	0.159E-4	0.990	0.365E-5	0.983	0.130E-3	1.003
0.004	115 201	0.796E-5	0.993	0.183E-5	0.992	0.647E-4	1.002

5.2. Test 2: Linear Fokker-Planck equation

In this test case, we study the behavior of our scheme on the problem (1.1) with the choice of nonlinearities $s(p) = \exp(p)$ and $\eta(s) = s$. The function s does not fulfill assumption (**A1**) since s is not constant on $\mathbb{R}_+$. Since s in injective, we can use $S = s(p)$ as a primary unknown, leading to the problem

$$\begin{cases} \partial_t S - \nabla\cdot(S\Lambda(\nabla \log(S) - \mathbf{e}_x)) = 0 & \text{in} \quad Q_{t_f}, \\ S\Lambda(\nabla \log(S) - \mathbf{e}_x)\cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0,T), \\ S_{|_{t=0}} = s_0 & \text{in} \quad \Omega, \end{cases} \tag{5.4}$$

that turns out to be the linear convection diffusion equation

$$\begin{cases} \partial_t S - \nabla\cdot(\Lambda(\nabla S - S\mathbf{e}_x) = 0 & \text{in} \quad Q_{t_f}, \\ \Lambda(\nabla S - S\mathbf{e}_x)\cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0,T), \\ S_{|_{t=0}} = s_0 & \text{in} \quad \Omega. \end{cases} \tag{5.5}$$

We compare the results obtained with the nonlinear CVFE scheme (5.2) with the following *linear scheme* where the convection is discretized thanks to centered fluxes:

$$\frac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL}\in\mathcal{E}_K} a_{KL}\left((s_K^{n+1} - s_L^{n+1}) + (x_K - x_L)\frac{s_K^{n+1} + s_L^{n+1}}{2}\right) = 0 \tag{5.6}$$

for all $K \in \mathcal{V}$ and for all $n \in \{0, \dots, N\}$.

The schemes (5.2) and (5.6) are compared on the following analytical solution built from a 1D case:

$$s_{\text{ex}}(x,y,t) = \exp\left(-\alpha t + \frac{x}{2}\right)\left(\pi\cos(\pi x) + \frac{1}{2}\sin(\pi x)\right) + \pi\exp\left(x - \frac{1}{2}\right) \text{ in } Q_{t_f},$$

where $\alpha = \Lambda_{xx}(\pi^2 + \frac{1}{4})$, and where the final time has been fixed to 0.05. This analytical solution is nonnegative and satisfies homogeneous Neumann boundary conditions.

In Tables 3 to 6, we report the $L^1(Q_{t_f})$, $L^2(Q_{t_f})$, and $L^\infty(Q_{t_f})$ on the variable S, *i.e.*,

$$err_{L^p} = \|s_{\mathcal{M},\Delta t} - s_{\text{ex}}\|_{L^p(Q_{t_f})} \quad \text{for } p = 1, 2, \infty$$

The numerical order of convergence of the linear scheme (5.6) is close to 2. However, the more the anisotropy ratio is important, the more we observe oscillations and undershoots (see in particular Tab. 6). The nonlinear scheme (5.2) preserves the positivity of the solution whatever the anisotropy, but this property has a cost. Indeed, the numerical diffusion introduced by the nonlinear scheme (5.2) becomes very important when the anisotropy ratio is large. This yields a loss of accuracy. The method (5.2) seems to be first order accurate, *i.e.*,

$$err_{L^p} \le C_p(\Lambda,\theta)h, \quad p \in \{1, 2, \infty\}, \tag{5.7}$$

but with constants $C_p(\Lambda,\theta)$ that strongly depend on the anisotropy ratio and of the regularity of the mesh.

TABLE 3. Test 2, nonlinear scheme (5.2), with an isotropic tensor $\Lambda = \mathrm{Id}$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.328E-01	–	0.820E-02	–	0.232E+00	–	0
0.250	37	0.306E-01	0.0979	0.798E-02	0.0389	0.239E+00	−0.0466	0
0.125	129	0.198E-01	0.6320	0.508E-02	0.6519	0.153E+00	0.6477	0
0.063	481	0.109E-01	0.8674	0.276E-02	0.8911	0.841E-01	0.8722	0
0.031	1857	0.570E-02	0.9130	0.143E-02	0.9237	0.441E-01	0.9101	0
0.016	7297	0.292E-02	1.0152	0.729E-03	1.0214	0.226E-01	1.0123	0
0.008	28 929	0.147E-02	0.9845	0.368E-03	0.9893	0.114E-01	0.9831	0
0.004	115 201	0.741E-03	0.9923	0.185E-03	0.9937	0.575E-02	0.9913	0

TABLE 4. Test 2, linear scheme (5.6) with an isotropic tensor $\Lambda = \mathrm{Id}$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.294E-01	–	0.372E-02	–	0.484E+00	–	0
0.250	37	0.829E-02	1.8267	0.198E-02	1.6352	0.166E+00	1.5428	0
0.125	129	0.218E-02	1.9286	0.349E-03	1.8389	0.426E-01	1.9639	0
0.063	481	0.548E-03	2.0138	0.859E-04	1.9863	0.108E-01	2.0069	0
0.031	1857	0.137E-03	1.9521	0.216E-04	1.9310	0.274E-02	1.9310	0
0.016	7297	0.343E-04	2.0956	0.542E-05	2.0675	0.697E-03	2.0675	0
0.008	28 929	0.858E-05	1.9998	0.135E-05	1.9994	0.178E-03	1.9720	0
0.004	115 201	0.214E-05	2.0000	0.339E-06	1.9998	0.453E-04	1.9719	0

TABLE 5. Test 2: nonlinear scheme (5.2) with an anisotropic tensor $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 20$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.179E+00	–	0.488E-01	–	1.022E+00	–	0
0.250	37	0.166E+00	0.1080	0.462E-01	0.0792	0.959E+00	0.0930	0
0.125	129	0.118E+00	0.4947	0.318E-01	0.5396	0.744E+00	0.3659	0
0.063	481	0.746E-01	0.6685	0.197E-01	0.7008	0.504E+00	0.5689	0
0.031	1857	0.439E-01	0.7498	0.113E-01	0.7755	0.309E+00	0.6880	0
0.016	7297	0.243E-01	0.8904	0.621E-02	0.9118	0.177E+00	0.8416	0
0.008	28 929	0.130E-01	0.9087	0.327E-02	0.9229	0.964E-01	0.8793	0
0.004	115 201	0.672E-02	0.9481	0.169E-02	0.9571	0.506E-01	0.9304	0

TABLE 6. Test 2, linear scheme (5.6) with an anisotropic tensor: $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 20$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.566E-01	–	0.115E-01	–	0.376E+00	–	0
0.250	37	0.222E-01	1.3523	0.427E-02	1.4290	0.250E+00	0.5878	0
0.125	129	0.613E-02	1.8553	0.119E-02	1.8469	0.883E-01	1.5036	−2.1867E-03
0.063	481	0.155E-02	2.0021	0.300E-03	2.0053	0.247E-01	1.8621	−9.3704e-04
0.031	1857	0.390E-03	1.9506	0.755E-04	1.9468	0.647E-02	1.8859	−2.6687e-04
0.016	7297	0.976E-04	2.0948	0.189E-04	2.0952	0.168E-02	2.0358	−6.9729e-05
0.008	28 929	0.244E-04	1.9997	0.472E-05	1.9997	0.437E-03	1.9470	−1.7741e-05
0.004	115 201	0.610E-05	1.9999	0.118E-05	1.9999	0.113E-03	1.9495	−4.4696e-06

TABLE 7. Nonlinear scheme, with an isotropic tensor: $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 1$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.174E-02	–	0.215E-03	–	0.284E-01	–	0
0.250	37	0.238E-02	-0.4509	0.224E-03	−0.0573	0.559E-01	−0.9751	0
0.125	129	0.168E-02	0.5062	0.160E-03	0.4883	0.305E-01	0.8754	0
0.063	481	0.100E-02	0.7489	0.889E-04	0.8544	0.237E-01	0.3645	0
0.031	1857	0.609E-03	0.7049	0.486E-04	0.8522	0.174E-01	0.4369	0
0.016	7297	0.359E-03	0.7994	0.259E-04	0.9509	0.114E-01	0.6459	0
0.008	28 929	0.206E-03	0.8043	0.136E-04	0.9315	0.734E-02	0.6301	0
0.004	115 201	0.115E-03	0.8445	0.703E-05	0.9511	0.460E-02	0.6751	0

5.3. Test 3: Porous medium equation with drift

In this third test case, we set $s(p) = p/2$ if $p \geq 0$ and $\eta(s) = s$. Choosing $S = s(p)$ as a primary variable, we obtain the degenerate parabolic equation

$$\partial_t S - \nabla \cdot (\Lambda(2|S|\nabla S - S\mathbf{e}_x)) = 0 \qquad \text{in} \quad Q_{t_f},$$

or equivalently

$$\partial_t S - \nabla \cdot (\Lambda(\nabla\varphi(S) - S\mathbf{e}_x)) = 0 \qquad \text{in} \quad Q_{t_f}, \qquad \text{where } \varphi(S) = |S|S. \tag{5.8}$$

The function s_{ex} defined by

$$s_{\text{ex}}(x, y, t) = \max(\beta t - x, 0), \ \forall((x, y), t) \in Q_{t_f}, \tag{5.9}$$

with $\beta = 2\Lambda_{xx}$ satisfies the equation (5.8). As in 1, we complement (5.8) by Dirichlet boundary conditions and an initial condition prescribed by (5.9). The final time t_f has been set to 0.05.

The nonlinear scheme (5.2) is adapted to the case of Dirichlet boudary conditions: (5.2a) is assumed to hold only for $K \in \mathcal{V}_{\text{int}} = \{K \in \mathcal{V} \mid \boldsymbol{x}_K \notin \partial\Omega\}$. The equations (5.2b) and (5.2c) are enforced for all $K \in \mathcal{V}$, and (5.2d) is enforced for all $\sigma_{KL} \in \mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E} \mid \sigma \not\subset \partial\Omega\}$. In order to close the system, we impose $s_K^{n+1} = s_{\text{ex}}(\boldsymbol{x}_K, t_{n+1})$ for all K such that $\boldsymbol{x}_K \in \partial\Omega$.

The numerical results obtained thanks to our scheme are compared with those obtained thanks to a so-called *quasilinear scheme* where (5.2a) has been replaced by

$$\frac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} a_{KL} \left(\varphi(s_K^{n+1}) - \varphi(s_L^{n+1}) + (x_K - x_L) \frac{s_K^{n+1} + s_L^{n+1}}{2} \right) = 0. \tag{5.10}$$

The analytical solution s_{ex} defined by (5.9) belongs to $C([0, t_f), H^{3/2-\epsilon}(\Omega))$ for all $\epsilon > 0$. Therefore, we expect for the quasilinear scheme (5.10) a convergence order close to 1.5 in the $L^2(Q_{t_f})$ norm, as observed in Tables 8 and 10.

We observe that, as expected, that the nonlinear scheme (5.2) has a smaller order of convergence (less than 1) when Λ is isotropic, *cf.* Table 5. Here again, as in Test 2, the accuracy is strongly affected by the anisotropy. The numerical diffusion introduced by the scheme increases with the anisotropy ratio. But the solutions to the scheme (2.8) do not present undershoots (up to the precision of the nonlinear solver), on the contrary to the solutions to the quasilinear scheme (5.10), *cf.* Table 10. In order to illustrate the overdiffusive behavior of the nonlinear scheme (5.2) as well as the undershoots produced by the quasilinear scheme (5.10), we present in Figure 5 the snapshots of both numerical solutions at time $t = t_f$.

TABLE 8. Quasilinear scheme, with an isotropic tensor: $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 1$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.990E-03	–	0.120E-03	–	0.166E-01	–	0
0.250	37	0.148E-02	−0.5805	0.129E-03	−0.1076	0.383E-01	−1.2026	0
0.125	129	0.825E-03	0.8427	0.720E-04	0.8424	0.176E-01	1.1253	0
0.063	481	0.356E-03	1.2265	0.268E-04	1.4434	0.106E-01	0.7307	0
0.031	1857	0.151E-03	1.2052	0.998E-05	1.3912	0.582E-02	0.8507	0
0.016	7297	0.581E-04	1.4499	0.320E-05	1.7193	0.296E-02	1.0232	−1.3853e-18
0.008	28 929	0.214E-04	1.4403	0.950E-06	1.7531	0.149E-02	0.9921	−6.9053e-17
0.004	115 201	0.711E-05	1.4722	0.270E-06	1.8125	0.743E-03	1.0008	−2.1592e-18

TABLE 9. Nonlinear scheme, with an anisotropic tensor: $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.672E-02	–	0.983E-03	–	0.829E-01	–	0
0.250	37	0.664E-02	0.0178	0.102E-02	−0.0551	0.101E00	−0.2802	0
0.125	129	0.552E-02	0.2663	0.862E-03	0.2439	0.831E-01	0.2774	0
0.063	481	0.441E-02	0.3286	0.647E-03	0.4191	0.699E-01	0.2526	0
0.031	1857	0.345E-02	0.3471	0.458E-03	0.4876	0.625E-01	0.1586	0
0.016	7297	0.260E-02	0.4284	0.310E-03	0.5954	0.514E-01	0.2946	0
0.008	28 929	0.189E-02	0.4608	0.200E-03	0.6241	0.410E-01	0.3266	0
0.004	115 201	0.132E-02	0.5141	0.125E-03	0.6794	0.318E-01	0.3666	0

TABLE 10. Quasilinear scheme, with an anisotropic tensor: $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$.

h	$\#\mathcal{V}$	err_{L^2}	Rate	err_{L^1}	Rate	err_{L^∞}	Rate	$S_{\min}$
0.500	12	0.976E-02	–	0.159E-01	–	0.111E+00	–	−5.4034E-02
0.250	37	0.722E-02	0.4337	0.110E-02	0.5325	0.108E+00	0.0424	−3.5579E-02
0.125	129	0.414E-02	0.8015	0.583E-03	0.9103	0.589E-01	0.8736	−2.5825E-02
0.063	481	0.179E-02	1.2215	0.198E-03	1.5786	0.419E-01	0.4968	−1.1696E-02
0.031	1857	0.779E-03	1.1765	0.746E-03	1.3747	0.220E-01	0.9062	−5.8549E-03
0.016	7297	0.336E-02	1.2698	0.262E-04	1.5806	0.118E-01	0.9376	−2.9309E-03
0.008	28 929	0.140E-03	1.2662	0.876E-05	1.5822	0.636E-02	0.8980	−1.4663E-03
0.004	115 201	0.565E-04	1.3073	0.282E-05	1.6351	0.333E-02	0.9341	−7.3339E-04

5.4. Decay of discrete free energy

Let us denote by $\mathfrak{M}(Q_{t_f})$ the set of the measurable functions mapping Q_{t_f} to $\mathbb{R}$. The *free energy* functional [26] $\mathfrak{E} : \mathfrak{M}(Q_{t_f}) \to \mathbb{R} \cup \{+\infty\}$, defined by

$$\mathfrak{E}(p) = \int_\Omega \Big(\Gamma(p) + s(p)\mathbf{g} \cdot \boldsymbol{x} \Big) \, \mathrm{d}\boldsymbol{x}, \qquad \forall p \in \mathfrak{M}(Q_{t_f}), \tag{5.11}$$

consists in the sum of the capillary energy (1.6), and the gravitational energy. We have formally the classical energy/dissipation property (1.11), and in particular $t \longmapsto \mathfrak{E}(p)(t)$ is decreasing. The discrete counterpart of the *free energy* is

$$\mathfrak{E}(p^n_{\mathcal{M}}) = \sum_{K \in \mathcal{V}} m_K \big[\Gamma(p^n_K) + g s(p^n_K) x_K \big].$$

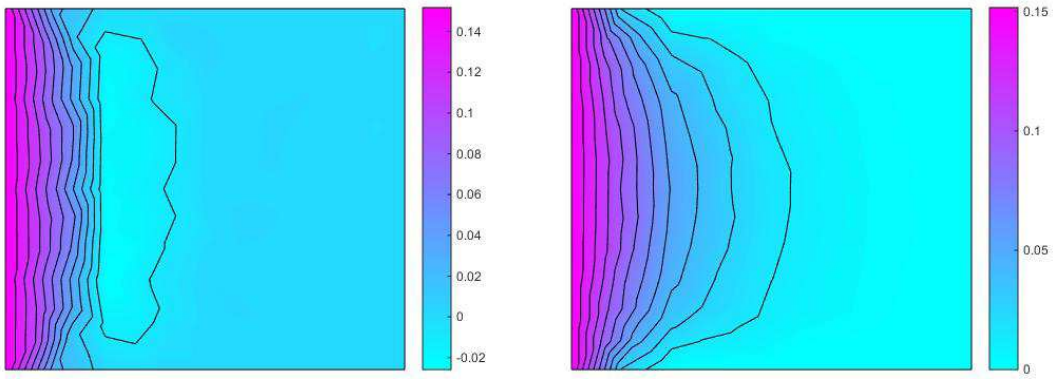

FIGURE 5. Test 3: 2nd mesh and anisotropic tensor $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$. Discrete solutions $s_{\mathcal{M},\Delta t}(\cdot, t_f)$ and their iso-values. *Left:* Quasilinear scheme (5.10). *Right:* Nonlinear scheme (5.2).

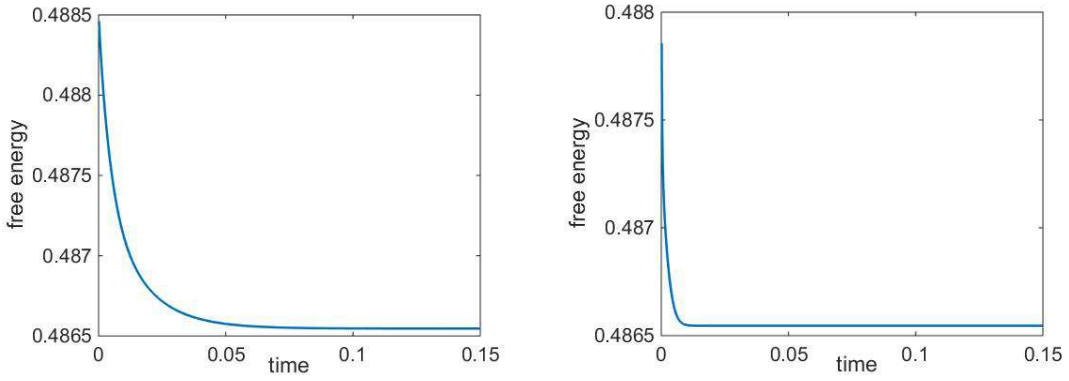

FIGURE 6. Evolution of the free energy along time, for $\Lambda_{xx} = 1, \Lambda_{yy} = 1$ (*on the left*) and $\Lambda_{xx} = 1, \Lambda_{yy} = 100$ (*on the right*).

We have not succeeded to prove the decay of the discrete free energy contrarily to [11]. Let us provide a numerical evidence of this energy/dissipation property. Define the nonlinearities

$$s(p) = \begin{cases} \dfrac{1}{1+p^2} & \text{if} \quad p < 0, \\ 1 & \text{if} \quad p \geq 0, \end{cases} \qquad \eta(s) = s^2,$$

and set $\mathbf{g} = \mathbf{e}_x$, and

$$p_0 = \begin{cases} -\dfrac{x-y}{2} & \text{if} \quad x - y < 0, \\ -\tan\left(\dfrac{\mathrm{e}^{x-y}-1}{\mathrm{e}^{x-y}+1}\right) & \text{if} \quad x - y \geq 0. \end{cases}$$

We solve the scheme (2.8) and we remark (*cf.* Fig. 6) that $(\mathfrak{E}(p^n_{\mathcal{M}}))_{n\geq 0}$ is decreasing. As already noticed on the previous test cases, the scheme (2.8) suffers from an excessive numerical diffusion, in particular when the anisotropy ratio is high. The origins of faster convergence towards the equilibrium in the anisotropic case illustrated by Figure 6 are twofold. The anisotropy favors the convergence towards the equilibrium at the continuous level. But the additional numerical diffusion introduced by the scheme also accelerates this convergence.

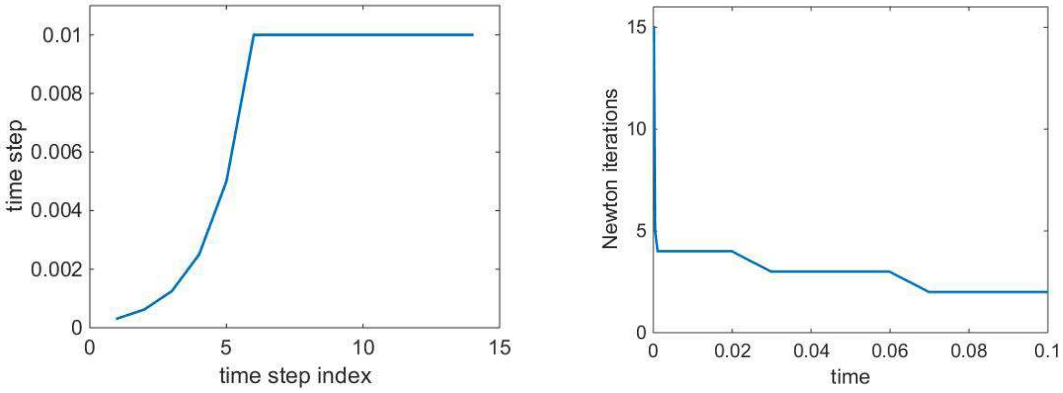

FIGURE 7. Adaptive time step (*left*) and Newton iterations (*right*).

5.5. Newton-Raphson method and adaptive time stepping

In order to illustrate the behavior of Newton's method, we consider again Test 2 of Section 5.2 solved with the nonlinear scheme (5.2) on the 5th mesh in the anisotropic case $\Lambda_{yy} = 20$. The final time t_f for the simulation is set to 0.1. The maximal time step $\Delta t_{\max}$ mesh is fixed to 0.01, and the initial time step is chosen equal to $\Delta t_{\max}$. We observe on Figure 7 that the Newton's method fails to converge a first step. Four successive time step refinements are required. But there is no need to further refine the time step in the next steps. The time step increases until it reaches the maximal value $\Delta t_{\max}$.

6. Conclusion

We proposed and analyzed a nonlinear energy stable scheme for solving the Richards equation. Moreover, the definition of the scheme only rely on physical quantities and not on artificial ones like for instance the Kirchhoff transform. We were able to carry out a full convergence analysis based on compactness arguments. Contrarily to classical schemes, this new nonlinear scheme produces no undershoot. As far as we know, our scheme is the first one to ensure that the discrete solution remains in the physical range even in the case of strong anisotropy.

However, it appears in the numerical simulations that in the case of strong anisotropy ratio, the scheme introduces an excessive numerical diffusion that makes its convergence very slow. This shall motivate the design of some new more robust schemes (for instance based on [12]) that preserve the main advantages of the scheme studied in this paper, namely the formulation in physical variables, the preservation of the physical range, and then control of the physical energy.

Appendix A.

A.1. Some inequalities of Sobolev's type

Lemma A.1. *Let $q \geq 1$, and let $u \in W^{1,q}(\Omega)$ be such that*

$$u \geq 0 \quad \text{and} \quad \lambda(\{u = 0\}) \geq \alpha > 0, \tag{A.1}$$

where λ denotes the 2-dimensional Lebesgue measure. Define $q^\star = 2q/(2-q)$ if $q < 2$ and $q^\star = +\infty$ if $q \geq 2$, then, for all $r \leq q^\star$ if $q \neq 2$ and $r < \infty$ if $q = 2$, there exists C depending only on Ω, r, and α such that

$$\|u\|_{L^r(\Omega)} \leq C \|\nabla u\|_{L^q(\Omega)^2}.$$

Proof. Define the mean $\langle u \rangle$ value of u by

$$\langle u \rangle = \frac{1}{\lambda(\Omega)} \int_\Omega u(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \geq 0.$$

Due to the properties (A.1) of u, one has

$$\int_\Omega |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} = \int_{\{u=0\}} \langle u \rangle \, \mathrm{d}\boldsymbol{x} + \int_{\{u>0\}} |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} \geq \alpha \langle u \rangle.$$

On the other hand, thanks to Poincaré's inequality (see, *e.g.*, [1]), one has

$$\int_\Omega |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} \leq \frac{\mathrm{diam}(\Omega)}{2} \int_\Omega |\nabla u| \, \mathrm{d}\boldsymbol{x} \leq \frac{\mathrm{diam}(\Omega)}{2} \lambda(\Omega)^{\frac{q-1}{q}} \|\nabla u\|_{L^q(\Omega)}.$$

Therefore, we get that

$$\langle u \rangle \leq \frac{\mathrm{diam}(\Omega)}{2\alpha} \lambda(\Omega)^{\frac{q-1}{q}} \|\nabla u\|_{L^q(\Omega)}.$$

Combining this estimate with Sobolev's inequality (see, *e.g.*, [2]) yields

$$\|u\|_{L^r(\Omega)} \leq \|u - \langle u \rangle\|_{L^r(\Omega)} + \lambda(\Omega) \langle u \rangle \leq C \|\nabla u\|_{L^q(\Omega)^d}$$

where C depends only the prescribed quantities. □

In the next Lemma, we prove a discrete Sobolev inequality. Note that the proof takes advantage of the existence of a conformal $V_\mathcal{T}$, leading to a much simpler proof than in [18] or [6].

Lemma A.2. *Let $\mathcal{T}$ and $\mathcal{M}$ be a primal and a dual discretizations of Ω as prescribed in Section 2.1.1. Let $(u_K)_{K \in \mathcal{V}}$ be an arbitrary element of $\mathbb{R}^{\#\mathcal{V}}$, and denote by*

$$\langle u_\mathcal{M} \rangle = \frac{1}{\lambda(\Omega)} \int_\Omega u_\mathcal{M} \, \mathrm{d}\boldsymbol{x} = \frac{1}{\lambda(\Omega)} \int_\Omega u_\mathcal{T} \, \mathrm{d}\boldsymbol{x}.$$

Then there exists C depending only on r, q Ω, and $\theta_\mathcal{T}$ such that

$$\|u_\mathcal{M} - \langle u_\mathcal{M} \rangle\|_{L^r(\Omega)} \leq C \int_\Omega |\nabla u_\mathcal{T}|^q \, \mathrm{d}\boldsymbol{x}, \qquad \forall r \in [1, \infty), \forall q \geq \min\left(1, \frac{2r}{2+r}\right).$$

Proof. Since $u_\mathcal{T}$ is Lipschitz continuous, the classical Sobolev inequality (*cf.* [2]) gives that

$$\|u_\mathcal{T} - \langle u_\mathcal{M} \rangle\|_{L^r(\Omega)} \leq C \int_\Omega |\nabla u_\mathcal{T}|^q \, \mathrm{d}\boldsymbol{x}, \qquad \forall r \in [1, \infty), \forall q \geq \min\left(1, \frac{2r}{2+r}\right).$$

It only remains to use (2.1) to conclude the proof. □

With that discrete Sobolev inequality at hand (*cf.* Lem. A.2), we can now easily adapt the proof of Lemma A.1 to the discrete setting, leading to the following statement, whose proof is left to the reader.

Lemma A.3. *Let $(v_K)_{K \in \mathcal{V}}$, and let $v_\mathcal{M}$ and $v_\mathcal{T}$ the corresponding discrete functions belonging to $X_\mathcal{M}$ and $V_\mathcal{T}$ respectively. Assume that*

$$v_\mathcal{M} \geq 0 \quad \text{and} \quad \lambda_d(\{v_\mathcal{M} = 0\}) \geq \alpha > 0,$$

Define $q^\star = qd/(d-q)$ if $q < d$ and $q^\star = +\infty$ if $q \geq d$, then, for all finite *$r \leq q^\star$, there exists C depending only on Ω, $\theta_\mathcal{T}$, r, and α such that*

$$\|v_\mathcal{M}\|_{L^r(\Omega)} \leq C \|\nabla v_\mathcal{T}\|_{L^q(\Omega)^d}.$$

A.2. Uniqueness of the weak solution

Proposition A.4. *Under assumptions* (**A1**)–(**A4**)*, there exists a unique weak solution to the problem* (1.1) *in the sense of Definition* 1.4.

Proof. First, define the full Kirchhoff transform $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(p) = \int_0^p \eta(s(a))\mathrm{d}a, \qquad \forall p \in \mathbb{R}.$$

It follows from assumptions (**A1**) and (**A2**) that φ is Lipschitz continuous, increasing, and fulfills $p\varphi(p) > 0$ for all $p \neq 0$. Since η is assumed to be bounded, one has

$$\nabla\varphi(p) = \sqrt{\eta(s(p))}\nabla\xi(p) \in L^2(Q_{t_f})^d$$

for any $p : Q_{t_f} \to \mathbb{R}$ such that $\xi(p) \in L^2((0,T); H^1(\Omega))$ (thus in particular for any weak solution). Therefore, any weak solution p satisfies

$$\iint_{Q_{t_f}} s(p)\partial_t\psi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega s_0\psi(\cdot,0)\mathrm{d}\boldsymbol{x} + \iint_{Q_{t_f}} (\eta(s(p))\mathbf{g} - \nabla\varphi(p)) \cdot \Lambda\nabla\psi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0 \tag{A.2}$$

for all $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_f))$. Mimicking Otto's uniqueness proof for degenerate parabolic-elliptic problems [31], we obtain that, given two weak solutions p and $\widehat{p}$ corresponding to the same initial data s_0, one has

$$\int_\Omega |s(p(\boldsymbol{x},t)) - s(\widehat{p}(\boldsymbol{x},t))|\mathrm{d}\boldsymbol{x} \leq 0 \quad \text{for a.e. } t \geq 0, \tag{A.3}$$

hence $s(p) = s(\widehat{p})$. Moreover, the mass being conserved, it follows from assumption (**A4**) that

$$0 < \int_\Omega s(p(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} = \int_\Omega s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \overline{s}_0 \,\mathrm{meas}(\Omega) < \mathrm{meas}(\Omega) \quad \text{for a.e. } t \geq 0.$$

Therefore, defining

$$\mathcal{U}(t) := \{\boldsymbol{x} \in \Omega \mid s(p(\cdot,t)) < 1\} \quad \text{for a.e. } t \geq 0,$$

one has

$$\mathrm{meas}\,(\mathcal{U}(t)) \geq (1 - \overline{s}_0)\mathrm{meas}(\Omega) > 0 \quad \text{for a.e. } t \geq 0. \tag{A.4}$$

Since s is increasing on $[p_\star, 0]$, one gets that $p(\cdot,t) = \widehat{p}(\cdot,t)$ on $\mathcal{U}(t)$ for a.e. $t \geq 0$.

Subtracting the weak formulation (A.2) corresponding to p to the one for $\widehat{p}$ then yields

$$\iint_{Q_{t_f}} \nabla\big(\varphi(p) - \varphi(\widehat{p})\big) \cdot \Lambda\nabla\psi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0, \quad \forall\psi \in C_c^\infty(\overline{\Omega} \times [0, t_f)),$$

and thus for all ψ in $L^2((0,T); H^1(\Omega))$ thanks to a density argument. Choosing $\psi = \varphi(p) - \varphi(\widehat{p})$ and using assumption (**A3**) yields

$$\|\nabla\big(\varphi(p(\cdot,t)) - \varphi(\widehat{p}(\cdot,t))\big)\|_{L^2(\Omega)^d} = 0 \quad \text{for a.e. } t \geq 0.$$

The function $\varphi(p) - \varphi(\widehat{p})$ is identically equal to 0 on $\mathcal{U}(t)$, we can apply Lemma A.1 to infer that

$$\|\varphi(p(\cdot,t)) - \varphi(\widehat{p}(\cdot,t))\|_{L^2(\Omega)} = 0 \quad \text{for a.e. } t \geq 0.$$

Since φ is increasing, one obtains that $p = \widehat{p}$ a.e. in Q_{t_f}. □

References

[1] G. Acosta and R.G. Durán, An optimal Poincaré inequality in L^1 for convex domains. *Proc. Amer. Math. Soc.* **132** (2004) 195–202.

[2] R.A. Adams and J.J.F. Fournier, Sobolev spaces. Academic press (Elsevier), 2nd edition (2003).

[3] H.W. Alt and S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183** (1983) 311–341.

[4] 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.

[5] 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.

[6] M. Bessemoulin−Chatard, C. Chainais−Hillairet and F. Filbet, On discrete functional inequalities for some finite volume schemes. *IMA J. Numer. Anal.* **35** (2015) 1125–1149.

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

[8] K. Brenner, D. Hilhorst and H.C. Vu Do, A gradient scheme for the discretization of Richards equation. In Finite volumes for complex applications. VII. Elliptic, parabolic and hyperbolic problems. Vol. 78 of *Springer Proc. Math. Stat.* Springer, Cham (2014) 537–545.

[9] R.H. Brooks and A.T. Corey, Hydraulic properties of porous media and their relation to drainage design. *Trans. ASAE* **7** (1964) 0026–0028.

[10] C. Cancès and C. Guichard, Entropy-diminishing CVFE scheme for solving anisotropic degenerate diffusion equations. In Finite volumes for complex applications. VII. Methods and theoretical aspects. Vol. 77 of *Springer Proc. Math. Stat.* Springer, Cham (2014) 187–196.

[11] 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.

[12] C. Cancès and C. Guichard, Numerical analysis of a robust free energy-diminishing Finite Volume scheme for degenerate parabolic equations with gradient structure. *Found. Comput. Math.* **17** (2017) 1525–1584.

[13] C. Cancès and M. Pierre, An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.* **44** (2012) 966–992.

[14] K. Deimling, Nonlinear functional analysis. Springer Verlag, Berlin (1985).

[15] J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci.* **24** (2014) 1575–1619.

[16] 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.

[17] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods. In Vol. VII of *Handbook of numerical analysis, Vol. VII.* North-Holland, Amsterdam (2000) 713–1020.

[18] R. Eymard, T. Gallouët and R. Herbin, Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.* **30** (2010) 1009–1043.

[19] 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. *Math. Models Methods Appl. Sci.* **11** (2001) 1505–1528.

[20] R. Eymard, M. Gutnic and D. Hilhorst, The finite volume method for Richards equation. *Comput. Geosci.* **3** (2000) 259–294.

[21] P.A. Forsyth, A control volume finite element approach to NAPL groundwater contamination. *SIAM J. Sci. Statist. Comput.* **12** (1991) 1029–1057.

[22] P.A. Forsyth and M.C. Kropinski, Monotonicity considerations for saturated–unsaturated subsurface flow. *SIAM J. Sci. Comput.* **18** (1997) 1328–1354.

[23] T. Gallouët and J.-C. Latché, Compactness of discrete approximate solutions to parabolic PDEs – application to a turbulence model. *Commun. Pure Appl. Anal.* **11** (2012) 2371–2391.

[24] R. Herbin and F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V.* ISTE, London (2008) 659–692.

[25] U. Hornung and W. Messing, Poröse medien: methoden und simulation. Verlag Beiträge zur Hydrologie (1984).

[26] R. Jordan, D. Kinderlehrer and F. Otto, Free energy and the Fokker-Planck equation. In Special Issue: *16th Annual International Conference of the Center for Nonlinear Studies, Los Alamos, 1996. Phys. D* **107** (1997) 265–271.

[27] R.A. Klausen, F.A. Radu and G.T. Eigestad, Convergence of MPFA on triangulations and for Richards' equation. *Int. J. Numer. Meth. Fl.* **58** (2008) 1327–1351.

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

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

[30] Y. Mualem, A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resour. Res.* **12** (1976) 513–522.

[31] F. Otto, L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differ. Equ.* **131** (1996) 20–38.

[32] I.S. Pop, Error estimates for a time discretization method for the Richards' equation. *Comput. Geosci.* **6** (2002) 141–160.

[33] I.S. Pop, M. Sepúlveda, F.A. Radu and O.P. Vera Villagrán, Error estimates for the finite volume discretization for the porous medium equation. *J. Comput. Appl. Math.* **234** (2010) 2135–2142.

[34] F. 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.

[35] F.A. Radu, I.S. Pop and P. Knabner, Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations. In *Numer. Math. Adv. Appl.* Springer, Berlin (2006) 1192–1200.

[36] L.A. Richards, Capillary conduction of liquids through porous mediums. *J. Appl. Phys.* **1** (1931) 318–333.

[37] J. Simon, Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl.* **146** (1987) 65–96.

[38] 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.

[39] R.L. Zarba, E.T. Bouloutas and M. Celia, General mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.* **26** (1990) 1483–1496.