FAST AND ROBUST NUMERICAL SOLUTION OF THE RICHARDS EQUATION IN HOMOGENEOUS SOIL

HEIKO BERNINGER, RALF KORNHUBER, AND OLIVER SANDER

Abstract. We derive and analyse a solver-friendly finite element discretization of a time discrete Richards equation based on Kirchhoff transformation. It can be interpreted as a classical finite element discretization in physical variables with non-standard quadrature points. Our approach allows for nonlinear outflow or seepage boundary conditions of Signorini type. We show convergence of the saturation and, in the non-degenerate case, of the discrete physical pressure. The associated discrete algebraic problems can be formulated as discrete convex minimization problems and, therefore, can be solved efficiently by monotone multigrid methods. In numerical examples for two and three space dimensions we observe L^2-convergence rates of order $\mathcal{O}(h^2)$ and H^1-convergence rates of order $\mathcal{O}(h)$ as well as robust convergence behaviour of the multigrid method with respect to extreme choices of soil parameters.

1. Introduction

The Richards equation [7, 15, 32] serves as a model for the description of saturated–unsaturated groundwater flow and reads

$$n\,\theta(p)_t + \operatorname{div}\mathbf{v}(p) = 0\,, \qquad \mathbf{v}(p) = -K_h kr(\theta(p))(\nabla p - z) \tag{1.1}$$

in case of a homogeneous soil. Here, p is the unknown water or capillary pressure on $\Omega \times (0,T)$ for a time $T > 0$ and a domain $\Omega \subset \mathbb{R}^3$ inhibited by the porous medium. The porosity and the hydraulic conductivity of the soil $n : \Omega \to (0,1)$ and $K_h : \Omega \to \mathbb{R}^+$, respectively, may vary in space. The coordinate in the direction of gravity is denoted by z. The saturation $\theta : \mathbb{R} \to [\theta_m, \theta_M]$ with $\theta_m, \theta_M \in [0,1]$ is an increasing function of p which is constant $\theta(p) = \theta_M$—the case of full saturation and ellipticity of (1.1)—if p is sufficiently large. The relative permeability $kr : [\theta_m, \theta_M] \to [0,1]$ is an increasing function of θ with $kr(\theta_M) = 1$. It usually leads to a degeneracy in the elliptic-parabolic equation (1.1) by $kr(\theta) \to 0$ for $\theta \to \theta_m$ or even by $kr(\theta_m) = 0$ whereby it becomes an ODE.

The homogeneous character of (1.1) is due to the fact that neither $\theta(\cdot)$ nor $kr(\cdot)$ depend explicitly on $x \in \Omega$, i.e., these parameter functions are fixed on Ω and describe the relationships in a single soil only. Concrete forms of these functions have been given by Brooks and Corey [13] and van Genuchten [37]. We use the parameter functions according to Brooks and Corey which are constituted by the bubbling pressure $p_b < 0$ and the pore size distribution factor $\lambda > 0$ as the relevant soil parameters.

1991 *Mathematics Subject Classification.* MSC (2000): 65N30, 65N55, 76S05.

This work was supported by the BMBF–Programm “Mathematik für Innovationen in Industrie und Dienstleistungen”.

It is a long-standing problem in unsaturated porous media flow simulations that "most discretization approaches for Richards' equation lead to nonlinear systems that are large and difficult to solve" [21] and that "poor iterative solver performance ... [is] often reported" [25]. Apart from the degeneracy resulting from $kr(\theta) \to 0$ this is due to the fact that the parameter functions degenerate to step functions for extreme soil parameters. Therefore, it is necessary for robustness to refrain from linearizing the Richards equation (1.1) in the (iterative) solution process. To our knowledge there are no numerical approaches to the Richards equation in the literature that meet this requirement. For example, although several different discretizations are used in the papers by Wagner et al. [38], Fuhrmann [22], Schneid et al. [33] and Bastian et al. [6], all these authors apply Newton's method to the resulting finite-dimensional algebraic equations.

In this paper, we strive for robustness. Following Alt and Luckhaus [1] and Visintin [2], our approach is based on a Kirchhoff transformation of the physical pressure into a generalized pressure $u = \kappa(p)$. In this way the nonlinearity and the degeneracy are removed from the main part of the differential operator and only reappear as the inverse transformation κ^{-1} and its ill-conditioning, respectively. Incorporating Signorini-type boundary conditions occurring, e.g., around seepage faces at the bank of a lake, it turns out that the transformed problem can be formulated in a weak sense as a nonlinear parabolic variational inequality involving the monotonically increasing nonlinearity $u \mapsto \theta(\kappa^{-1}(u))$.

By an (otherwise implicit) time discretization in which the gravitational term is treated explicitly one obtains an elliptic variational inequality or, equivalently, a strictly convex minimization problem to be solved in each time step. The spatial discretization is carried out by piecewise linear finite elements. Upwinding of the gravitational (i.e., convective) part guarantees stability for sufficiently small time steps. In practical computations, however, this CFL condition does not seem severe. We prove H^1-convergence of the finite element approximations u_j to the generalized pressure.

The discretization in generalized variables u_j can be reinterpreted as a standard finite element discretization of the original Richards equation (1.1) in physical pressure p_j with numerical integration based on particular (solution dependent) quadrature points. More precisely, the physical approximations p_j and the retransformed generalized approximations $\kappa^{-1}(u_j)$ have the same nodal values. If the Richards equation is nondegenerate we obtain H^1-convergence of $\kappa^{-1}(u_j)$ and L^2-convergence of its piecewise linear interpolation p_j to the physical solution of the time discrete problem. Similar convergence results are obtained for the approximate saturation $\theta(\kappa^{-1}(u_j))$.

Our new approach pays off in two regards. First, the ill-conditioning inherent in the degenerate problem (i.e., the case that $kr(\cdot)$ can become arbitrarily small) is decoupled from the solution process by the Kirchhoff transformation and reoccurs only in the inverse transformation $u \mapsto p = \kappa^{-1}(u)$ after u has been determined. Secondly, there are powerful multigrid solvers [24, 29] at hand for the discrete minimization problems providing the approximations of the generalized pressure in each time step. These methods are based on successive minimization rather than linearization and, therefore, perform robustly even for extreme variations of the soil parameters.

Our approach is limited to homogeneous soils in the sense of spatially independent relative permeability kr (not hydraulic conductivity K_h) and saturation θ. The case of heterogeneous soils shall be treated in a forthcoming paper [11].

Outline. The paper is structured as follows. In Section 2 we first introduce the Brooks–Corey parameter functions and the Kirchhoff transformation. Then we give a weak formulation of a Signorini-type boundary value problem for the Richards equation as a variational inequality. In Section 3 we present our implicit–explicit time discretization and show that the resulting variational inequality is equivalent to a uniquely solvable convex minimization problem.

In Section 4 we introduce a finite element discretization of the convex minimization problem and prove H^1-convergence. There we also give a reinterpretation of the discretization in generalized variables as a certain finite element discretization of the untransformed problem in the physical pressure. Convergence results for the discrete saturation and the discrete physical pressure are derived, too.

In Section 5 we shortly indicate how existing monotone multigrid methods [24, 29] are applied to the spatial minimization problems. Finally, in Section 6 we present numerical discretization error studies for a semidiscrete problem in 2D with Brooks–Corey parameter functions. We determine numerically the asymptotic behaviour of the discretization error in transformed and in physical variables. In both cases we obtain the order of convergence $\mathcal{O}(h^2)$ for the L^2-norm and $\mathcal{O}(h)$ for the H^1-norm. By the way we illustrate and analyze the ill-conditioning of the inverse Kirchhoff transformation and its effect in the numerical calculations. In Section 7 we present numerical results for a dam problem in 3D. We observe good convergence rates within a large range of soil parameters, which illustrates the robust behaviour of our spatial solver.

2. Signorini-type problem and variational inequality for the Richards equation

The purpose of this first section is to develop a weak formulation of a boundary value problem for the Richards equation in which nonlinear outflow conditions around seepage faces, occurring, e.g., at the bank of a lake, are taken into account. We start by giving the concrete forms of the parameter functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$ according to Brooks and Corey which can be regarded as prototypes for these relationships and which we use in our numerical examples. Thereafter, we apply the Kirchhoff transformation to the Richards equation and then introduce our boundary value problem in a strong form involving Signorini-type boundary conditions to account for seepage faces. Finally, this problem is given a weak sense in the form of a variational inequality.

2.1. Brooks–Corey parameter functions.

Let $\theta_m, \theta_M \in [0,1]$, $\theta_m < \theta_M$, be the residual and the maximal saturation of water in a homogeneous soil, respectively. Furthermore, we assume the bubbling pressure $p_b < 0$ and the pore size distribution factor $\lambda > 0$ of the soil to be known. Then, according to Brooks and Corey [13], the saturation θ as a function of the pressure p is given by

$$\theta(p) = \begin{cases} \theta_m + (\theta_M - \theta_m)\left(\frac{p}{p_b}\right)^{-\lambda} & \text{for } p \le p_b \\ \theta_M & \text{for } p \ge p_b \end{cases} \tag{2.2}$$

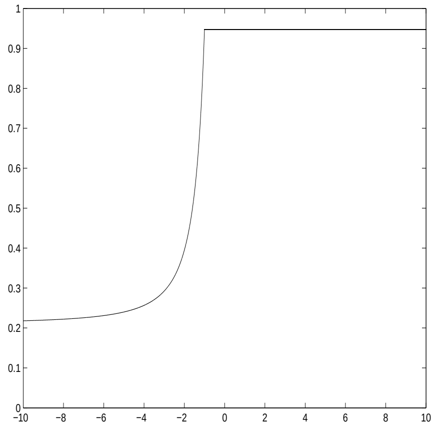

FIGURE 1. $p \mapsto \theta(p)$

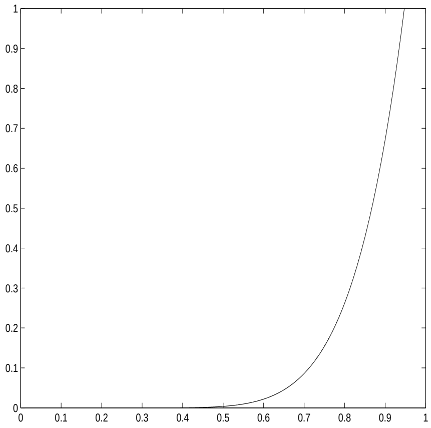

FIGURE 2. $\theta \mapsto kr(\theta)$

and (with results by Burdine [14]) the relative permeability kr as a function of the saturation reads

$$kr(\theta) = \left(\frac{\theta - \theta_m}{\theta_M - \theta_m} \right)^{3+\frac{2}{\lambda}} , \quad \theta \in [\theta_m, \theta_M] \,. \tag{2.3}$$

Typical shapes of these nonlinearities are depicted in Figures 1 and 2. As a consequence the composite function $kr(\theta(\cdot))$ in (1.1) has the form

$$kr(\theta(p)) = \begin{cases} \left(\frac{p}{p_b}\right)^{-2-3\lambda} & \text{for } \ p \le p_b \\ \quad 1 & \text{for } \ p \ge p_b \,. \end{cases} \tag{2.4}$$

We point out that although we use the Brooks–Corey functions in our numerical examples we do not restrict ourselves to this special case in the theory. In the following we collect the essential properties that the Brooks–Corey functions have in common with all other hydrologically reasonable parameter functions $\theta(\cdot)$ and $kr(\cdot)$ and that will be used in this paper.

Lemma 2.1. *The Brooks–Corey functions θ and kr in (2.2) and (2.3) are nonnegative, bounded, monotonically increasing and continuous.*

2.2. **Kirchhoff transformation.** In the following we assume $n = K_h = 1$ for simplicity and thus deal with the Richards equation in the form

$$\theta(p)_t - \operatorname{div}\Big(kr(\theta(p))\nabla\,(p - z)\Big) = 0\,. \tag{2.5}$$

An essential tool for our approach is Kirchhoff's transformation which is well known in the literature on the Richards equation, see for example Alt et al. [2] or Eymard et al. [20]. It is defined by

$$\kappa : p \mapsto u := \int_0^p kr(\theta(q))\, dq \tag{2.6}$$

where the new variable u shall be called *generalized pressure.* If we take the chain rule into account which provides

$$\nabla u = kr(\theta(p))\nabla p \tag{2.7}$$

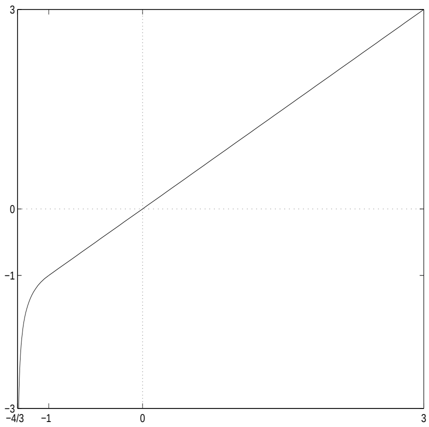

FIGURE 3. $u \mapsto \kappa^{-1}(u)$

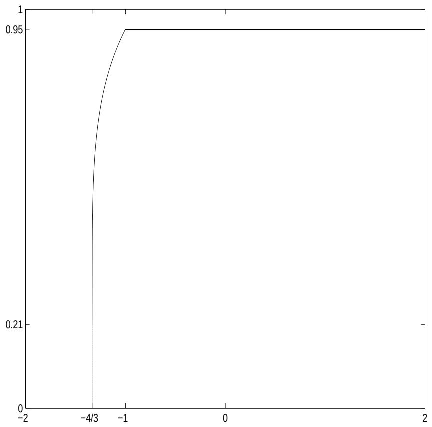

FIGURE 4. $u \mapsto M(u)$

and denote the saturation as a function of u by

$$M(u) := \theta(\kappa^{-1}(u)) \tag{2.8}$$

the transformed Richards equation (2.5) reads

$$M(u)_t - \operatorname{div}\Big(\nabla u - kr(M(u))e_z\Big) = 0\,. \tag{2.9}$$

Hence, we obtain a semilinear equation from the quasilinear equation (2.5). Here, we denote by $e_z = \nabla z$ the unit vector in the direction of gravity. Observe that we have $u = p$ due to $kr(\theta_M) = 1$ in case of full saturation, i.e., for $p \geq p_b$, and $p \leq 0 \Leftrightarrow u \leq 0$ since $kr(\cdot) \geq 0$. Furthermore, we point out that the image $\kappa(\mathbb{R})$ can be a strict subset (u_c, ∞) of $\mathbb{R}$ and, consequently, κ^{-1} and M may be defined on (u_c, ∞) only. This is indeed the case for the parameter functions according to Brooks and Corey. Obviously, the critical pressure $u_c < 0$ corresponds to $p = -\infty$ in this case and, therefore, $M(u_c) = \theta_m$ is a sensible definition. Typically, M has unbounded derivatives and κ^{-1} is ill-conditioned around u_c, compare Figures 3 and 4. Both these singularities disappear, however, and the functions M and κ^{-1} are Lipschitz continuous on $\mathbb{R}$ in the nondegenerate case

$$kr(\cdot) \geq c \quad \text{for a } \ c > 0\,. \tag{2.10}$$

For the Brooks–Corey functions this can be obtained, e.g., by redefining $kr(\cdot)$ by the function $\max(kr(\cdot), c)$.

As in Lemma 2.1 we state the decisive properties of M, κ and κ^{-1}, which are easy to prove.

Lemma 2.2. *If θ and kr satisfy the properties in Lemma 2.1, the function M defined by (2.8) is nonnegative, bounded, monotonically increasing and continuous. Furthermore, the function $\kappa : \mathbb{R} \to \mathbb{R}$ is monotonically increasing and in $C^1(\mathbb{R})$.*

If θ and kr are chosen according to Brooks and Corey in (2.2) and (2.3) then, with some $u_c < 0$, the function M is defined on $[u_c, \infty)$ and is Hölder continuous whereas κ^{-1} is a continuous function defined on (u_c, ∞) with $\kappa^{-1}(u) \to -\infty$ for $u \downarrow u_c$. In the nondegenerate case (2.10) with $kr \in L^\infty(\theta(\mathbb{R}))$, both κ and κ^{-1} are Lipschitz continuous functions on $\mathbb{R}$, and if, in addition, θ is Lipschitz continuous (as in the Brooks–Corey case), so is M.

Remark 2.3. *We point out that the results of this paper hold in case of the Brooks–Corey parameter functions—with* $\max(kr(\cdot), c)$ *instead of* $kr(\cdot)$ *wherever nondegeneracy (2.10) is assumed. However, they apply to more general cases and only depend on properties listed in Lemmas 2.1 and 2.2. In order to make the theory as transparent as possible, we will always make clear which properties of the functions are needed for a result to hold.*

2.3. **Signorini-type boundary value problem.** Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. For a time $t \in (0, T]$ we assume that a decomposition of $\partial\Omega$ into submanifolds $\gamma_D(t)$, $\gamma_N(t)$ and $\gamma_S(t)$ with a positive Hausdorff measure as well as functions $u_D(t) \in H^{1/2}(\gamma_D(t))$ and $f_N(t) \in L^2(\gamma_N(t))$ are given. Then the unknown function u and the flux

$$\mathbf{v} = -(\nabla u - kr(M(u))e_z) = -kr(\theta(p))\nabla\,(p - z)$$

shall satisfy the boundary conditions

$$u = u_D(t) \quad \text{on } \gamma_D(t) \tag{2.11}$$

$$\mathbf{v}\cdot\mathbf{n} = f_N(t) \quad \text{on } \gamma_N(t) \tag{2.12}$$

$$u \le 0\,, \quad \mathbf{v}\cdot\mathbf{n} \ge 0\,, \quad u\cdot(\mathbf{v}\cdot\mathbf{n}) = 0 \quad \text{on } \gamma_S(t)\,. \tag{2.13}$$

Dirichlet boundary conditions (2.11), i.e., a prescribed water pressure $p_D(t) = \kappa^{-1}(u_D(t))$, and Neumann boundary conditions (2.12), i.e., a prescribed water flux $f_N(t)$, are well known in porous media flow problems.

The boundary conditions (2.13), which are less common and sometimes called outflow conditions [34], describe the situation on and close to seepage faces, which can be found, e.g., at the bank around lakes or in dam problems, see e.g. [16]. There, inflow of water does not occur and we have $p = 0$ if water flows out ($\mathbf{v}\cdot\mathbf{n} > 0$) and $p \le 0$ if there is no outflow ($\mathbf{v}\cdot\mathbf{n} = 0$).

Strangely enough, these quite natural boundary conditions do not seem to have been given a name by hydrologists. Although the name "seepage face boundary condition" can be found in the literature (see, e.g., [17]), this can be misleading since the seepage face itself is only the part of $\gamma_S(t)$ where outflow actually happens, i.e., where we have the boundary condition $p = 0$. The actual problem is to find the extent of the seepage face on $\gamma_S(t)$, since its boundary is a free boundary and part of the solution that has to be determined, often within an iterative process (cf. [8]), e.g., by switching between Neumann and Dirichlet boundary conditions [17]. With regard to this problem, our approach turns out to be quite elegant and appropriate, because the boundary conditions (2.13) constitute just another obstacle in the convex minimization problem that we obtain in Section 3.

Mathematically, the nonlinear complementarity conditions (2.13) are exactly the ones known from Signorini problems in mechanics (see, e.g., Kikuchi and Oden [26]), and this starts to be acknowledged in porous media research (see, e.g., [36]). Therefore, we call them *Signorini-type boundary conditions* here and refer to the corresponding boundary value problem (2.9)–(2.13) as a *Signorini-type problem* for the (Kirchhoff–transformed) Richards equation.

2.4. **Variational inequality.** Just as variational equalities in Sobolev spaces turn out to be appropriate weak formulations of boundary value problems, we obtain a variational inequality on a convex subset of the space $H^1(\Omega)$ as a weak formulation of the Signorini-type problem (2.9)–(2.13) for the Richards equation. This is justified by an equivalence result which holds if all functions in (2.9)–(2.13) and $\partial\Omega$ are smooth enough and which relates (2.9)–(2.13) to a variational inequality on a convex subset of the space $C^2(\overline{\Omega})$ of twice continuously differentiable functions on $\overline{\Omega}$, see [9, Prop. 1.5.3].

Before we turn to our problem formulation we give some notation and specifications. Let $\gamma \subset \partial\Omega$ be a submanifold with a positive Hausdorff measure. We call

$$tr_\gamma : H^1(\Omega) \to H^{1/2}(\gamma)$$

the corresponding trace operator. With the decomposition of $\partial\Omega$ and the functions $u_D(t)$ and $f_N(t)$ given above we set

$$\mathcal{K}(t) := \{v \in H^1(\Omega) : v \geq u_c \,\wedge\, tr_{\gamma_D(t)}v = u_D(t) \,\wedge\, tr_{\gamma_S(t)}v \leq 0\}\,, \tag{2.14}$$

where $\geq$, $=$ and $\leq$ have to be understood to hold up to null sets on Ω, $\gamma_D(t)$ and $\gamma_S(t)$, respectively. It is not hard to show that $\mathcal{K}(t)$ is a nonempty, closed and convex subset of $H^1(\Omega)$ if $u_D(t)$ is chosen to be compatible with the other conditions constituting $\mathcal{K}(t)$, see [9, Prop. 1.5.5].

Now, relaxing the assumptions from the introduction, let $kr : M(\mathbb{R}) \to \mathbb{R}$ be a bounded Borel function and $M : [u_c, \infty) \to \mathbb{R}$ be bounded, monotonically increasing and continuous. Then we say that $u \in L^2(0,T;H^1(\Omega))$, with the property $M(u)_t \in L^2(\Omega)$ a.e. on $(0,T]$, is a weak solution of (2.9)–(2.13) at the time $t \in (0,T]$ if $u(t) \in \mathcal{K}(t)$ and

$$\begin{aligned} (2.15)\quad &\int_\Omega M(u(t))_t\,(v-u(t))\,dx + \int_\Omega \nabla u(t)\,\nabla(v-u(t))\,dx \geq \\ &\int_\Omega kr(M(u(t)))e_z\nabla(v-u(t))\,dx - \int_{\gamma_N(t)} f_N(t)\,(v-u(t))\,d\sigma \quad \forall v \in \mathcal{K}(t)\,. \end{aligned}$$

We remark that it is possible to relate this variational inequality to a corresponding one given in the physical pressure $p(t)$ for the original Richards equation (2.5) with the boundary value problem (2.9)–(2.13) retransformed in physical variables. More concretely, $u(t)$ solves (2.9)–(2.13) if $p(t)$ solves the corresponding variational inequality, and, in case of $kr(\cdot) \geq c > 0$ and $\gamma_S(t) = \emptyset$, both formulations are equivalent. For the analysis one needs the chain rule (2.7) in a weak form and an interpretation of the Kirchhoff transformation (2.6) as a superposition operator on $H^1(\Omega)$ and on trace spaces, see [9, Sec. 1.5.4] and [10].

3. Implicit–explicit time discretization and convex minimization

In the following we give our implicit–explicit time discretization of the variational inequality (2.15). Our aim in this section is to derive an equivalent uniquely solvable convex minimization problem from the resulting variational inequality.

3.1. Time discretization.

For simplicity and without loss of generality we set $f_N(t) = 0$ in (2.15). Let $0 = t_0 < t_1 < \ldots < t_N = T$ be a partition of $[0,T]$ with the time step sizes $\tau_n := t_n - t_{n-1}$, $n \in \{1,\ldots,N\}$, and set $u^0 = u(0) \in H^1(\Omega)$ as the given initial condition for (2.9). Then our time discretized version of (2.15) reads: Find $u^n \in \mathcal{K}(t_n)$, successively for $n = 1, \dots, N$, with

(3.16)
$$\begin{aligned}&\int_\Omega M(u^n)\,(v-u^n)\,dx + \tau_n \int_\Omega \nabla u^n \nabla(v-u^n)\,dx \geq \\ &\int_\Omega M(u^{n-1})\,(v-u^n)\,dx + \tau_n \int_\Omega kr(M(u^{n-1}))e_z\nabla(v-u^n)\,dx \quad \forall v \in \mathcal{K}(t_n)\,.\end{aligned}$$

This amounts to a time discretization of (2.15) in which the main part of the equation is treated implicitly whereas the term arising from gravity is treated explicitly.

We proceed with some notation and abbreviations. For a given $n \in \{1, \dots, N\}$, we set $\mathcal{K} := \mathcal{K}(t_n)$ and also $\gamma_D := \gamma_D(t_n)$, $\gamma_S := \gamma_S(t_n)$ and $\gamma_N := \gamma_N(t_n)$ as well as $u_D := u_D(t_n)$. We choose a $w \in H^1(\Omega)$ such that $tr_{\gamma_D} w = u_D$. Then we define the space

$$H^1_{\gamma_D}(\Omega) := \{v \in H^1(\Omega) : tr_{\gamma_D} v = 0\}$$

and the translated convex set

$$\mathcal{K}_{\gamma_D} := \mathcal{K} - w = \{v \in H^1_{\gamma_D}(\Omega) : v \geq u_c - w \,\wedge\, tr_{\gamma_S} v \leq -tr_{\gamma_S} w\}\,. \tag{3.17}$$

Furthermore, we denote $u = u^n$. The left hand side in (3.16) is given by a continuous linear functional ℓ on $\mathcal{K} \subset H^1(\Omega)$ defined as

$$\ell(v) := \int_\Omega M(u^{n-1})\,v\,dx + \tau_n \int_\Omega kr(M(u^{n-1}))e_z \nabla v\,dx \quad \forall v \in H^1(\Omega)\,. \tag{3.18}$$

We abbreviate the norm in the Sobolev space $H^1(\Omega)$ by

$$\|v\|_1 = \left(\int_\Omega |v|^2 + |\nabla v|^2\,dx\right)^{1/2} \quad \forall v \in H^1(\Omega)\,.$$

Let $\gamma_D \subset \partial\Omega$ have a positive Hausdorff measure. Then the continuous symmetric bilinear form $a(\cdot,\cdot)$ on $H^1(\Omega)$ given by

$$a(v,w) := \tau_n \int_\Omega \nabla v \nabla w\,dx \quad \forall v, w \in H^1(\Omega) \tag{3.19}$$

is coercive on $H^1_{\gamma_D}(\Omega)$, i.e., there is a $c > 0$ such that

$$a(v,v) \geq c\|v\|_1^2 \quad \forall v \in H^1_{\gamma_D}(\Omega)\,.$$

With this notation we can write (3.16) more compactly as the variational inequality

$$u \in \mathcal{K} : \quad \int_\Omega M(u)(v-u)\,dx + a(u, v-u) - \ell(v-u) \geq 0 \quad \forall v \in \mathcal{K}\,. \tag{3.20}$$

3.2. **Convex minimization.** The variational inequality (3.20) is equivalent to a convex minimization problem. In what is to come we sketch the reasoning to derive this fact and refer to [9, Sec. 2.3.2–2.3.4] for proofs and further details. We start with a primitive $\Phi : [u_c, \infty) \to \mathbb{R}$ of M defined as

$$\Phi(z) := \int_0^z M(s)\,ds \quad \forall z \in [u_c, \infty) \tag{3.21}$$

which gives rise to a functional $\phi : \mathcal{K} \to \mathbb{R}$ by

$$\phi(v) := \int_\Omega \Phi(v(x))\,dx \quad \forall v \in \mathcal{K}\,. \tag{3.22}$$

Since M is monotonically increasing, Φ is convex, and since M is bounded, Φ is Lipschitz continuous. Therefore, ϕ is a well-defined convex and Lipschitz continuous functional with the property

$$|\phi(v)| \leq C\|v\|_1 \quad \forall v \in \mathcal{K}$$

for a $C > 0$. Furthermore, since M is continuous, Φ is differentiable with $\Phi' = M$. This has analogous consequences on the differentiability of ϕ which we turn to now.

Recall that for a function $F : S \to \mathbb{R}$ on a subset $S \subset V$ of a normed space V the one-sided limit

$$\partial_v F(u) := \lim_{h \downarrow 0} \frac{F(u + hv) - F(u)}{h}\,, \quad u, u + hv \in S\,,$$

if it exists, is the directional derivative of F at u in the direction of $v \in V$.

Since Φ is convex and differentiable, one can interchange differentiation with the integral in (3.22) and obtain

Lemma 3.1. *For any $u, v \in \mathcal{K}$ the directional derivative $\partial_{v-u}\phi(u)$ exists and can be written as*

$$\partial_{v-u}\phi(u) = \int_\Omega \Phi'(u(x))(v(x) - u(x))\,dx = \int_\Omega M(u(x))(v(x) - u(x))\,dx\,.$$

Now, it is well known that the quadratic functional $\mathcal{J} : H^1_{\gamma_D}(\Omega) \to \mathbb{R}$ defined by

$$\mathcal{J}(v) := \frac{1}{2}a(v, v) - \ell(v) \quad \forall v \in H^1_{\gamma_D}(\Omega) \tag{3.23}$$

is strictly convex, continuous and coercive, i.e., for any sequence $(u_n) \subset H^1_{\gamma_D}(\Omega)$ with $\|u_n\|_1 \to \infty$ we have $\mathcal{J}(u_n) \to \infty$. Moreover, $\mathcal{J}$ is Fréchet–differentiable in $u \in H^1_{\gamma_D}(\Omega)$ with the derivative

$$\mathcal{J}'(u)(v) = \partial_v \mathcal{J}(u) = a(u, v) - \ell(v) \quad \forall v \in H^1_{\gamma_D}(\Omega)\,.$$

Consequently, the functional $F : \mathcal{K} \to \mathbb{R}$ defined by

$$F(v) := \phi(v) + \mathcal{J}(v) \quad \forall v \in \mathcal{K} \tag{3.24}$$

(and extended by $+\infty$ to $H^1_{\gamma_D}(\Omega)\backslash\mathcal{K}$) is strictly convex, continuous and coercive, and $\partial_{v-u}F(u)$ exists for any $u, v \in \mathcal{K}$. Altogether, we conclude that (3.20) has the following form: Find $u \in \mathcal{K}$ such that

$$\partial_{v-u}F(u) \geq 0 \quad \forall v \in \mathcal{K}\,.$$

The next result provides the link to convex minimization.

Lemma 3.2. *Let V be a real vector space, $K \subset V$ a convex set and $F : K \to \mathbb{R}$ a convex functional whose directional derivative $\partial_{v-u}F(u)$ exists for all $u, v \in K$. Then*

$$u \in K : \quad \partial_{v-u}F(u) \geq 0 \quad \forall v \in K$$

is equivalent to

$$u \in K : \quad F(u) \leq F(v) \quad \forall v \in K\,.$$

The functional F defined in (3.24) is strictly convex, proper (i.e., F does not assume $-\infty$ and is not identically $+\infty$), continuous and coercive. Therefore, we can apply a well-known general existence and uniqueness result for convex minimization problems in reflexive Banach spaces (see, e.g., Ekeland and Temam [19, p. 35]) and obtain the main result of this section.

Theorem 3.3. *Let $\mathcal{K} \subset H^1(\Omega)$, $a(\cdot,\cdot)$ and $\ell(\cdot)$ be defined as in (2.14), (3.19) and (3.18), respecively. If $M : [u_c, \infty) \to \mathbb{R}$ is bounded, monotonically increasing and continuous and $kr : M(\mathbb{R}) \to \mathbb{R}$ a bounded Borel function, then the variational inequality (3.20) has a unique solution. More specifically, it is equivalent to the minimization problem*

$$u \in \mathcal{K} : \quad \mathcal{J}(u) + \phi(u) \le \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K} \tag{3.25}$$

with $\mathcal{J}$ and ϕ as defined in (3.23) and (3.22), respectively.

Finally, we remark that Theorem 3.3 can be generalized to the case of nonnegative and bounded porosity $n(\cdot)$ and a hydraulic conductivity $K_h(\cdot)$ satisfying

$$c \le K_h(\cdot) \le C \quad \text{with some } \; c, C > 0\,. \tag{3.26}$$

It also applies to the case $M : \mathbb{R} \to \mathbb{R}$, in particular to the nondegenerate case (2.10).

4. Finite element discretization

In this section we present a finite element discretization of (3.25), which extends the results in Kornhuber [28, pp. 36–43] to our more general boundary conditions (see also Glowinski [23, pp. 12–15]). We give a reinterpretation as a certain finite element discretization of the problem in physical variables, thus making clear that our discretization in the transformed variables is not artificial. We obtain convergence of the discrete generalized solutions u_j to the continuous solution in the H^1-norm, which entails H^1-convergence of the corresponding saturation $M(u_j)$ and L^2-convergence of its piecewise linear interpolation. In case of nondegenerate $kr(\cdot) \ge c > 0$, we can also prove H^1-convergence of the retransformed pressure $\kappa^{-1}(u_j)$ as well as L^2-convergence of its piecewise linear interpolation.

4.1. Discretized problem in generalized variables.

For the sake of presentation we consider the two-dimensional case of a polygonal domain $\Omega \subset \mathbb{R}^2$. Let $\mathcal{T}_j$, $j \in \mathbb{N}_0$, be a partition of Ω into triangles $t \in \mathcal{T}_j$ with minimal diameter of order $\mathcal{O}(2^{-j})$. We assume the triangulation $\mathcal{T}_j$ to be conforming in the sense that the intersection of two different triangles in $\mathcal{T}_j$ is either empty or consists of a common edge or a common vertex. The set of all vertices of the triangles in $\mathcal{T}_j$ is denoted by $\mathcal{N}_j$.

For a consistent discretization, the triangulation should resolve the parts of the boundary corresponding to different boundary conditions. Therefore, we require that each intersection point of two closures (in $\mathbb{R}^2$) of the subsets γ_D, γ_N and γ_S of the boundary $\partial\Omega$ is contained in $\mathcal{N}_j$. Furthermore, we assume γ_D and $\gamma_S \cup \gamma_D$ to be closed and we define $\mathcal{N}_j^D := \mathcal{N}_j \cap \gamma_D$ as well as $\mathcal{N}_j^S := \mathcal{N}_j \cap \gamma_S$.

We choose the finite element space $\mathcal{S}_j \subset H^1(\Omega)$ as the subspace of all continuous functions in $H^1(\Omega)$ which are linear on each triangle $t \in \mathcal{T}_j$. Analogously, we define $\mathcal{S}_j^D \subset H^1_{\gamma_D}(\Omega)$. $\mathcal{S}_j$ and $\mathcal{S}_j^D$ are spanned by the nodal bases

$$\Lambda_j := \{\lambda_q^{(j)} : q \in \mathcal{N}_j\} \quad \text{and} \quad \Lambda_j^D := \{\lambda_q^{(j)} : q \in \mathcal{N}_j \backslash \mathcal{N}_j^D\}\,,$$

respectively, which for $\mathcal{S}_j^D$ is only guaranteed because of our special choice of $\mathcal{N}_j^D$ containing all intersection points of parts of $\partial\Omega$ adjacent to γ_D.

For the definition of the finite dimensional analogue of $\mathcal{K}$ we assume that the Dirichlet boundary condition u_D is continuous in each node $q \in \mathcal{N}_j^D$, $j \in \mathbb{N}_0$, such that writing $u_D(q)$ makes sense in these nodes. Then it is natural to define the convex set $\mathcal{K}_j \subset \mathcal{S}_j$ by

(4.27)
$$\mathcal{K}_j := \left\{ v \in \mathcal{S}_j : v(q) \geq u_c \, \forall q \in \mathcal{N}_j \;\wedge\; v(q) = u_D(q) \, \forall q \in \mathcal{N}_j^D \;\wedge\; v(q) \leq 0 \, \forall q \in \mathcal{N}_j^S \right\}$$

which, as a subset of the finite dimensional space $\mathcal{S}_j$, is clearly nonempty and closed.

Furthermore, we approximate the integral in the definition (3.22) of ϕ by a quadrature formula arising from $\mathcal{S}_j$-interpolation of the integrand $\Phi(v)$. In this way, we arrive at the discrete functional $\phi_j : \mathcal{S}_j \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\phi_j(v) := \sum_{q \in \mathcal{N}_j} \Phi(v(q)) \, h_q \quad \forall v \in \mathcal{S}_j \tag{4.28}$$

with the positive weights

$$h_q := \int_\Omega \lambda_q^{(j)}(x) \, dx \, .$$

The following properties of the discrete functionals ϕ_j and, in particular, their relation to the continuous counterpart ϕ in (3.22) are crucial (see [9, p. 80/81] for a proof).

Lemma 4.1. *Provided M in (3.21) is monotonically increasing and bounded, the functional ϕ_j is convex and Lipschitz continuous on its domain*

$$\operatorname{dom} \phi_j = \{ v \in \mathcal{S}_j : v(q) \geq u_c \ \ \forall q \in \mathcal{N}_j \}$$

with a Lipschitz constant independent of $j \geq 0$. Furthermore, ϕ_j is lower semicontinuous and proper and it admits an estimate

$$\phi_j(v) \geq C \|v\|_1 \quad \forall v \in \mathcal{S}_j$$

with a constant $C \in \mathbb{R}$ independent of $j \geq 0$. Moreover, for $v_j \in \mathcal{S}_j$, $j \geq 0$, and $v \in H^1(\Omega)$ we have

$$v_j \rightharpoonup v, \; j \to \infty \implies \liminf_{j \to \infty} \phi_j(v_j) \geq \phi(v)$$

where $v_j \rightharpoonup v$ denotes the weak convergence of v_j to v in $H^1(\Omega)$.

With the definitions from above, our discrete version of the minimization problem (3.25) reads

$$u_j \in \mathcal{K}_j : \;\; \mathcal{J}(u_j) + \phi_j(u_j) \leq \mathcal{J}(v) + \phi_j(v) \quad \forall v \in \mathcal{K}_j \, . \tag{4.29}$$

Since $\mathcal{K}_j$, $\mathcal{J}$ and ϕ_j have the same properties as $\mathcal{K}$, $\mathcal{J}$ and ϕ in Theorem 3.3, now in the subspace $\mathcal{S}_j$ of the Hilbert space $H^1(\Omega)$, we obtain

Theorem 4.2. *The discrete minimization problem (4.29) has a unique solution.*

Note that we do not alter the quadratic functional $\mathcal{J}$ for the discretization. In practice one needs to apply a quadrature rule if $K_h(\cdot)$ is a space-dependent function. Moreover, with regard to stability in the numerical treatment of the discretized problem it is necessary to use an upwind technique for the gravitational term in the linear functional (3.18). In the finite element context this can be achieved by adding an artificial viscosity term to the discretized convection, cf. [11].

4.2. **Interpretation in physical space: discrete Kirchhoff transformation.** Now we give a reinterpretation of (4.29) in terms of discrete physical variables. It turns out that (4.29) can be understood as a finite element discretization of problem (2.5), written in physical variables, where a particular quadrature rule with κ-dependent quadrature points for $kr(\theta(p))$ is applied.

By Lemma 3.2 the discrete minimization problem (4.29) is equivalent to the variational inequality

$$u_j \in \mathcal{K}_j : \sum_{q\in\mathcal{N}_j} M(u_j(q))\,(v(q)-u_j(q))\,h_q + a(u_j, v-u_j) - \ell(v-u_j) \geq 0 \quad \forall v \in \mathcal{K}_j \tag{4.30}$$

if $M : [u_c, \infty) \to \mathbb{R}$ is continuous. Obviously, (4.30) can be regarded as the corresponding discretization of the original variational inequality (3.20).

It is clear that in case of $u_j(q) = u_c$ for a $q \in \mathcal{N}_j$ we have $\kappa^{-1}(u_j(q)) = -\infty$, which is a physically unrealistic situation. Note that the somewhat unnatural condition $v \geq u_c$ instead of $v > u_c$ in (2.14) and, correspondingly, in (4.27) is necessary to guarantee the existence of a solution to the minimization problem by the closedness of the convex sets $\mathcal{K}$ and $\mathcal{K}_j$, respectively, and does not occur in the original physical problem. Therefore, we assume

$$u_j(q) > u_c \quad \forall q \in \mathcal{N}_j \tag{4.31}$$

from now on, which entails real-valuedness of $\kappa^{-1}(u_j)$ and allows the

Definition 4.3. *Let $I_{\mathcal{S}_j} : H^1(\Omega) \cap C(\overline{\Omega}) \to \mathcal{S}_j$ be the piecewise linear interpolation operator defined by $(I_{\mathcal{S}_j} v)(q) = v(q)$ $\forall q \in \mathcal{N}_j$ for $v \in H^1(\Omega) \cap C(\overline{\Omega})$. With the assumption (4.31) we call*

$$I_{\mathcal{S}_j}\kappa : \mathcal{S}_j \to \mathcal{S}_j$$

the discrete Kirchhoff transformation *on $\mathcal{S}_j$. Furthermore, assuming (4.31) we call*

$$p_j := I_{\mathcal{S}_j}\kappa^{-1}(u_j)$$

the discrete physical pressure *corresponding to problem (4.29).*

We are now going to investigate what kind of discretization of the untransformed problem corresponds to the discrete pressure variable p_j. To this end we impose the condition

$$\kappa \in C^1(\mathbb{R})$$

on the Kirchhoff transformation (2.6) which means that $kr \circ \theta$ is continuous. The latter is satisfied for the Brooks–Corey parameter functions in (2.2) and (2.3).

First, by (2.8) we clearly have

$$M(u_j(q)) = \theta(p_j(q)) \quad \forall q \in \mathcal{N}_j\,.$$

Accordingly, the linear term $\ell(\cdot)$ arising from the solution of the previous time step on the right hand side in (2.15) is retransformed in discrete physical variables. The remaining problem is to see how the bilinear form

$$a(u_j, w) = \int_\Omega \nabla u_j \nabla w \, dx\,, \quad w = v - u_j\,,\ v \in \mathcal{K}_j\,, \tag{4.32}$$

looks like in physical variables. For the continuous problem (2.5) the reformulation is provided by the chain rule (2.7) in a weak sense, consult [9, Sec. 1.5.4] or [10]. For the discrete problem we need a discrete counterpart of (2.7) and argue as follows with the help of the mean value theorem.

First, we consider the integral in (4.32) only on a triangle $t \in \mathcal{T}_j$. Recall that the transformation from the reference triangle

$$T \subset \mathbb{R}^2 \quad \text{with the vertices} \quad a = (0,0)\,, \ b = (1,0)\,, \ c = (0,1) \tag{4.33}$$

onto the triangle t is given by an affine map

$$F_t : \xi \mapsto x = B_t \xi + b_t$$

acting on $\mathbb{R}^2$ with a nonsingular matrix $B_t \in \mathbb{R}^{2\times 2}$ and a vector $b_t \in \mathbb{R}^2$. Transformed functions on the reference element shall be denoted by

$$\hat{v}(\xi) := v(F_t(\xi)) = v(x) \quad \forall x \in t \ \forall v \in H^1(\Omega) \cap C(\overline{\Omega})\,.$$

By the chain rule we can write

$$\nabla_\xi \, \hat{v}(\xi) = \nabla_x \, v(x) \, B_t \quad \forall x = F_t(\xi) \in t \ \forall v \in H^1(\Omega) \cap C(\overline{\Omega})\,.$$

Now, with the Eucledian norm $|\cdot|$ in $\mathbb{R}^2$ and (4.33) the first component in $\nabla_\xi \, \hat{u}_j$ is given by

$$\big(\nabla_\xi \, \hat{u}_j\big)_1 = \frac{\hat{u}_j(b) - \hat{u}_j(a)}{|b-a|} = \frac{\kappa(\hat{p}_j(b)) - \kappa(\hat{p}_j(a))}{\hat{p}_j(b) - \hat{p}_j(a)} \cdot \frac{\hat{p}_j(b) - \hat{p}_j(a)}{|b-a|}\,.$$

The range of the affine function $\hat{p}_j$ on the edge between a and b is the interval with the endpoints $\hat{p}_j(a)$ and $\hat{p}_j(b)$. Since κ is bijective and continuously differentiable on this interval, there exists a unique point $\bar{\xi}_1$ on the edge between a and b with the property

$$\big(\nabla_\xi \, \hat{u}_j\big)_1 = \kappa'(\hat{p}_j(\bar{\xi}_1)) \ \frac{\hat{p}_j(b) - \hat{p}_j(a)}{|b-a|} = kr(\theta(\hat{p}_j(\bar{\xi}_1))) \ \big(\nabla_x \, \hat{p}_j\big)_1\,. \tag{4.34}$$

Analogously, we can find a point $\bar{\xi}_2$ on the edge of T between the vertices a and c with the corresponding property. Altogether, with the transformation onto the reference triangle, the reformulation in physical variables and the transformation back onto t, we obtain

$$\nabla u_j = D_t(p_j) \nabla p_j \quad \text{on } t \tag{4.35}$$

with the diagonal matrix

$$D_t(p_j) = \begin{pmatrix} kr(\theta(p_j(\bar{x}_1))) & 0 \\ 0 & kr(\theta(p_j(\bar{x}_2))) \end{pmatrix}$$

and points

$$\bar{x}_1 = F_t(\bar{\xi}_1) \quad \text{and} \quad \bar{x}_2 = F_t(\bar{\xi}_2) \tag{4.36}$$

situated on edges of t.

We point out that since p_j is affine and $\kappa : \mathbb{R} \to (u_c, \infty)$ is bijective, the points $\bar{x}_1$ and $\bar{x}_2$ are uniquely defined by the properties (4.34) and (4.36). Therefore, the discrete counterpart (4.35) of the chain rule (2.7) also holds with the discrete Kirchhoff–transformed

$$u_j = I_{\mathcal{S}_j} \kappa(p_j)$$

if $p_j \in \mathcal{S}_j$ is known. In other words, just as the chain rule (2.7) corresponds to the Kirchhoff transformation in $H^1(\Omega)$, property (4.35) can be regarded as a discrete chain rule which corresponds to the discrete Kirchhoff transformation in $\mathcal{S}_j$.

Now we introduce the nonlinear form

$$b(p_j, v) := \sum_{t \in \mathcal{T}_j} \int_t D_t(p_j) \nabla p_j \nabla v \, dx \,, \quad p_j, v \in \mathcal{S}_j \,. \tag{4.37}$$

Then, with discrete Dirichlet boundary data p_D on $\mathcal{N}_j^D$ and the discrete closed and convex set

$$\mathcal{K}_j^0 := \left\{ v \in \mathcal{S}_j : v(q) = p_D(q) \ \ \forall q \in \mathcal{N}_j^D \wedge v(q) \le 0 \ \ \forall q \in \mathcal{N}_j^S \right\}$$

we consider the discrete problem

$$p_j \in \mathcal{K}_j^0 : \sum_{q \in \mathcal{N}_j} \theta(p_j(q)) \, (v(q) - p_j(q)) \, h_q + b(p_j, v - p_j) - \ell(v - p_j) \ge 0 \quad \forall v \in \mathcal{K}_j^0 \tag{4.38}$$

in physical variables. Note that in case of $\gamma_S = \emptyset$ we have

$$\mathcal{K}_j - u_j = \left\{ v \in \mathcal{S}_j : v(q) \ge -\varepsilon \ \ \forall q \in \mathcal{N}_j \ \wedge \ v(q) = 0 \ \ \forall q \in \mathcal{N}_j^D \right\}$$

with an $\varepsilon > 0$ due to the strict inequality in (4.31), so that by linearity the corresponding set of test functions $v - u_j$ in (4.30) can be chosen as the space $\mathcal{S}_j^D$ which is equal to $\mathcal{K}_j^0 - p_j$. On the other hand, with the assumption $kr(\theta(\mathbb{R})) \subset (0, 1]$ we have

$$p \le \kappa(p) \quad \forall p \in \mathbb{R} \,,$$

that is

$$\mathcal{K}_j - u_j \subset \mathcal{K}_j^0 - p_j \,.$$

In general, these sets of test functions in (4.30) and (4.38), respectively, are not equal. However, with these ingredients we can now prove the following discrete counterpart of Theorem 1.5.18 in [9], with arguments as given there for the continuous case.

Theorem 4.4. *Let $\theta : \mathbb{R} \to \mathbb{R}$ and $kr : \theta(\mathbb{R}) \to (0, 1]$ be bounded, monotonically increasing and continuous, while $\kappa : \mathbb{R} \to \mathbb{R}$ is defined by (2.6). In addition, let $p_D = \kappa^{-1}(u_D)$ on $\mathcal{N}_j^D$. Then $u_j = I_{\mathcal{S}_j} \kappa(p_j)$ solves (4.30) if p_j solves (4.38). Conversely, $p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j)$ solves (4.38) if u_j solves (4.30) with (4.31) in case of $\gamma_S(t) = \emptyset$. If $kr \ge c$ holds for a $c > 0$ and $\gamma_S(t) = \emptyset$, then (4.30) and (4.38) are equivalent in the sense that u_j satisfies (4.30) if and only if $p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j)$ satisfies (4.38).*

Our discretization (4.38) of problem (3.20), retransformed in physical variables, involves a quadrature formula with special quadrature points for the term

$$\int_\Omega kr(\theta(p)) \nabla p \nabla (v - p) \, dx \tag{4.39}$$

which is given by (4.37). This quadrature is uniquely defined and could even be explicitly formulated in terms of the given function $\kappa : \mathbb{R} \to \mathbb{R}$. Even though one would not use this quadrature in practical calculations, one would certainly be forced to use some quadrature for (4.39). At the end of this section we will prove that the quadrature (4.37) is as good as any appropriately chosen quadrature in the sense that it leads to a convergent discretization, see Theorem 4.12.

4.3. **Convergence of the generalized pressure.** Now we address the convergence of our finite element solutions from (4.29) to the solution of the continuous problem (3.25). The derivation of our results is largely based on the arguments given in Kornhuber [28, pp. 38–42] for the case of homogeneous Dirichlet boundary conditions on all of $\partial\Omega$. We need to take special care of the inhomogeneous Dirichlet values and the Signorini-type boundary conditions defined only on parts of $\partial\Omega$.

As in [28] the convergence results depend on the assumption that the corresponding sequence of triangulations has a decreasing mesh size

$$h_j := \max_{t \in \mathcal{T}_j} \operatorname{diam} t \to 0 \quad \text{for} \quad j \to \infty \, . \tag{4.40}$$

In addition, we assume that the sequence of triangulations

$$(\mathcal{T}_j)_{j \geq 0} \ \text{ is } \textit{shape regular} \tag{4.41}$$

which means that the minimal interior angle of all triangles contained in $\cup_{j\geq 0}\mathcal{T}_j$ is bounded from below by a positive constant.

It will turn out that we can only guarantee convergence if the Dirichlet boundary data u_D can be considered as the trace of a uniformly continuous function $w \in H^1(\Omega)$ on γ_D, i.e., if we have

$$u_D = tr_{\gamma_D} w \quad \text{for a} \quad w \in H^1(\Omega) \cap C(\overline{\Omega}) \, . \tag{4.42}$$

It is well known that if u_D is continuous on γ_D (which we assumed to be closed), then it can be extended to a continuous function on the closure of Ω (see [39, p. 498]). We require that there exists such an extension $w \in H^1(\Omega)$. Moreover, for the proof of convergence we will assume that the piecewise linear interpolations w_j of w in $\mathcal{S}_j$ approximate w in $H^1(\Omega)$, i.e., we require

$$w_j := I_{\mathcal{S}_j} w \to w \ \text{ for } \ j \to \infty \ \text{ in } \ H^1(\Omega). \tag{4.43}$$

In general, according to the interpolation theory in Ciarlet [18, pp. 122–124], the latter can only be guaranteed if w is regular enough. To check the assumptions stated there, we recall that the Sobolev embedding theorem (see [12, p. 1.52]) provides the compact embedding

$$H^k(t) \hookrightarrow C(\overline{t}) \quad \Longleftrightarrow \quad k > \frac{d}{2}$$

for polyhedra $t \subset \mathbb{R}^d$ and $k \in \mathbb{N}$. Now, with $t \in \mathcal{T}_j$ and $d = 2$ in our case, we obtain (4.43) (with order $\mathcal{O}(h_j)$) for $w \in H^2(\Omega)$ from results in [18, pp. 122–124], provided (4.40) and (4.41) hold. Consequently, we could also replace (4.42) and (4.43) by $u_D = tr_{\gamma_D} w$ with the condition $w \in H^2(\Omega)$ or a corresponding condition for $d > 2$.

In [28, pp. 38/39] it is proved that for $\tilde{\mathcal{K}} = \{v \in H_0^1(\Omega) : v \geq u_c\}$ the subset $C_0^\infty(\Omega) \cap \tilde{\mathcal{K}}$ is dense in $\tilde{\mathcal{K}}$. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, for given $v \in \tilde{\mathcal{K}}$ there is always a sequence $(v_k)_{k\geq 0} \subset C_0^\infty(\Omega)$ with $v_k \to v$ for $k \to \infty$. In order to ensure $v_k \in \tilde{\mathcal{K}}$, however, regularizations of v with mollifiers are considered. It is a nontrivial task to extend this result to more general settings like our convex set $\mathcal{K}_{\gamma_D}$ in $H^1_{\gamma_D}(\Omega)$. The technique can be refined (see [23, pp. 36–38]) to generalize the result to continuous obstacles on $\overline{\Omega}$ which are nonnegative in a neighbourhood of $\gamma_D = \partial\Omega$ as $u_c - w$ is in our case (3.17) with property (4.42). Furthermore, an exercise in [23, pp. 38/39] suggests that the latter result can be extended to $H^1_{\gamma_D}(\Omega)$ for sufficiently smooth $\gamma_D \subset \partial\Omega$ if one uses the density of

$$C^\infty_{\gamma_D}(\overline{\Omega}) := \big\{ v \in C^\infty(\overline{\Omega}) : v = 0 \text{ in a neighbourhood of } \gamma_D \big\}$$

in $H^1_{\gamma_D}(\Omega)$. As to the boundary conditions of Signorini's type, one finds a proof for the density of $C^\infty(\overline{\Omega}) \cap \bar{\mathcal{K}}$ in the convex set $\bar{\mathcal{K}} = \{v \in H^1(\Omega) : tr_{\partial\Omega}\, v \geq 0\}$ in [23, p. 61]. Since we do not want to go into more details here, it seems to be in order to require that

$$C^\infty_{\gamma_D}(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D} \text{ is dense in } \mathcal{K}_{\gamma_D} \tag{4.44}$$

as an additional condition for our translated convex set $\mathcal{K}_{\gamma_D}$ in (3.17). Property (4.44) provides an essential density argument by which one can generalize the convergence proof in [28, pp. 41/42] and obtain Theorem 4.6 below.

As a necessary ingredient for convergence, the following lemma provides the consistency of the discrete functionals ϕ_j. The proof is essentially the same as in [28, pp. 38–40] for the homogeneous case. However, the adaptation of the proof to our generalized case naturally leads to the assumptions given above on the extension w of u_D and the interpolating w_j. We refer to [9, p. 85] for details.

Lemma 4.5. *We assume (4.40), (4.41) and M in (3.21) to be bounded and monotonically increasing. If $v \in C^\infty(\overline{\Omega})$ and $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$ for $j \geq 0$, then we have the convergence*

$$v_j \to v \;\; \textit{in} \;\; H^1(\Omega) \quad \textit{and} \quad \phi_j(v_j) \to \phi(v) \quad \textit{for } j \to \infty\,. \tag{4.45}$$

Assuming in addition (4.42) and (4.43), the assertion (4.45) also holds for $v = w + \tilde{v} \in w + C^\infty(\overline{\Omega})$ and $v_j = I_{\mathcal{S}_j} v = I_{\mathcal{S}_j} w + I_{\mathcal{S}_j} \tilde{v} = w_j + \tilde{v}_j$, $j \geq 0$.

Now, with the consistency result in Lemma 4.5 and the further properties of ϕ_j in Lemma 4.1 one can prove convergence, adapting the ideas in [28, pp. 41/42] to our inhomogeneous case, see [9, pp. 86–88].

Theorem 4.6. *We assume (4.40)–(4.44) and the conditions imposed in Theorem 3.3 except for the continuity of M. Then the solutions u_j of the discrete minimization problem (4.29) converge to the solution u of (3.25) in the sense that*

$$u_j \to u \;\; \textit{in} \;\; H^1(\Omega) \quad \textit{and} \quad \phi_j(u_j) \to \phi(u) \quad \textit{for } j \to \infty\,.$$

We point out that the proof of Theorem 4.6 carries over to the case (3.26) of space-dependent hydraulic conductivity $K_h(\cdot)$ and also applies to $M : \mathbb{R} \to \mathbb{R}$. If, in addition, we have a positive and bounded porosity $n(\cdot)$ in (1.1) and use a discretization of the accordingly altered ϕ in (3.22) by adapting the weights in (4.28) properly, Theorem 4.6 holds, too (compare [9, pp. 89–91]).

4.4. **Convergence of the saturation and the physical pressure.** The last part of this section is devoted to what can be inferred from Theorem 4.6 on the behaviour of the saturation $M(u_j)$ and the retransformed pressure $\kappa^{-1}(u_j)$ as well as their piecewise linear interpolations $I_{\mathcal{S}_j} M(u_j)$ and $p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j)$ for $j \to \infty$. The L^2-convergence results for the saturation hold in quite general situations including the Brooks–Corey model. For the physical variables we only obtain convergence results in case of uniformly bounded p_j, $j \geq 0$, which is guaranteed if the Richards equation is nondegenerate in the sense of (2.10). Although we only prove

L^2-convergence of $I_{\mathcal{S}_j}M(u_j)$ and p_j for $j \to \infty$, we show H^1-convergence for the iterates $M(u_j)$ and $\kappa^{-1}(u_j)$, which can also be evaluated on a discrete level.

Recall that the real functions M and κ^{-1} induce superposition operators by composition $M \circ u$ and $\kappa^{-1} \circ u$ (consult, e.g., [3]). We start with a few results about superposition operators induced on Lebesgue spaces.

Lemma 4.7. *Let $\Omega \subset \mathbb{R}^d$ be bounded and $M : \mathbb{R} \to \mathbb{R}$ uniformly continuous and bounded. Then the corresponding superposition operator M_Ω acts on $L^2(\Omega)$ and is continuous.*

Sketch of the proof. M_Ω acts on $L^2(\Omega)$ because Ω and M are bounded. For the proof of continuity let $(u_n)_{n\geq 0} \subset L^2(\Omega)$ with $u_n \to u$ for $n \to \infty$ in $L^2(\Omega)$. One can split Ω in

$$\Omega^n_{>\varepsilon} = \{x \in \Omega : |u(x) - u_n(x)|^2 > \varepsilon\}$$

and $\Omega^n_{\leq\varepsilon} = \Omega \backslash \Omega^n_{>\varepsilon}$ for an $\varepsilon > 0$ and derive $|\Omega^n_{>\varepsilon}| \to 0$ for $n \to \infty$ with the Lebesgue measure $|\cdot|$ from the convergence $u_n \to u$ in $L^2(\Omega)$. Using the uniform continuity of M on $\Omega^n_{\leq\varepsilon}$ and its boundedness on $\Omega^n_{>\varepsilon}$, one can show $M(u_n) \to M(u)$ in $L^2(\Omega)$. □

Lemma 4.8. *Let $\Omega \subset \mathbb{R}^d$ be bounded. If $M : \mathbb{R} \to \mathbb{R}$ is Hölder continuous with respect to the exponent $\alpha \in (0,1]$, then M induces a superposition operator*

$$M_\alpha : L^2(\Omega) \to L^{2/\alpha}(\Omega)$$

which is also Hölder continuous with respect to α.

Sketch of the proof. One can show that $|M(u(\cdot))|^{2/\alpha}$ is integrable by considering $|M(u(\cdot))| \leq |M(u(\cdot)) - M(u(x_0))| + |M(u(x_0))|$ for an $x_0 \in \Omega$, using the inequality

$$(a+b)^q \leq 2^{q-1}(a^q + b^q) \tag{4.46}$$

for $a, b \geq 0$ and $q \geq 1$ (consult, e.g., [27, p. 161]) with $q = 2/\alpha$ and then the Hölder continuity of M. The claimed Hölder continuity of M_α is straightforward. □

With the continuous embedding $i : L^{2/\alpha}(\Omega) \hookrightarrow L^2(\Omega)$ for bounded $\Omega \subset \mathbb{R}^d$, it follows that $i \circ M_\alpha : L^2(\Omega) \to L^2(\Omega)$ is Hölder continuous with respect to α which improves the continuity result in Lemma 4.7 for Hölder continuous M. We point out that the saturation M in (2.8) from the Brooks–Corey parameter functions is indeed Hölder continuous. Now, in order to deduce

$$M(u_j) \to M(u) \quad \text{in } L^2(\Omega) \text{ for } j \to \infty$$

from $u_j \to u$ for $j \to \infty$ with the help of Theorem 4.6, it is enough to assume the properties of M (in Lemmas 4.7 or 4.8) only on the union of the ranges of the functions u_j, $j \geq 0$, and u.

The situation is more convenient in the nondegenerate case (2.10) since here the convergence properties of the generalized pressure are inherited by the saturation and the retransformed pressure.

Theorem 4.9. *In the nondegenerate case $kr(\cdot) \geq c > 0$, and with the assumptions of Theorem 4.6, we have the convergence*

$$M(u_j) \to M(u) \quad \text{in } H^1(\Omega) \text{ for } j \to \infty$$

and

$$\kappa^{-1}(u_j) \to \kappa^{-1}(u) \quad \text{in } H^1(\Omega) \text{ for } j \to \infty\,. \tag{4.47}$$

For the proof we can use the following remarkable result on superposition operators on $H^1(\Omega)$ (the converse of which is also true for $d \geq 2$, see [30]).

Lemma 4.10. *If $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, then the corresponding superposition operator acts on $H^1(\Omega)$ and is continuous.*

With respect to discrete solutions one will certainly be interested in the convergence behaviour of the $\mathcal{S}_j$-interpolations of $M(u_j)$ and $\kappa^{-1}(u_j)$, in particular, since the latter is the discrete physical pressure from the finite element discretization (4.38). With regard to the convergence of these inexact evaluations of $M(u_j)$ and $\kappa^{-1}(u_j)$, respectively, we first state the following

Lemma 4.11. *Let $f : \mathbb{R} \to \mathbb{R}$ be Hölder continuous with respect to the exponent $\alpha \in (0, 1]$. Then, for linear finite element functions $u_j \in \mathcal{S}_j$, $j \geq 0$, satisfying $u_j \to u$ in $H^1(\Omega)$ for $j \to \infty$, we have*

$$f(u_j) - I_{\mathcal{S}_j} f(u_j) \to 0 \quad \text{in } L^2(\Omega) \text{ for } j \to \infty\,. \tag{4.48}$$

Proof. For any point x contained in a triangle $t \in \mathcal{T}_j$ with the vertices q_1, q_2, q_3 there are $\vartheta_i \in [0, 1]$, $i = 1, 2, 3$, with $\sum_{i=1}^3 \vartheta_i = 1$ such that

$$I_{\mathcal{S}_j} f(u_j)(x) = \sum_{i=1}^{3} \vartheta_i\, f(u_j(q_i))\,.$$

Therefore, using binomial formulas such as (4.46) and the Hölder continuity of f with the Hölder constant C_α, we can estimate

$$\begin{aligned} |f(u_j(x)) - I_{\mathcal{S}_j} f(u_j)(x)|^2 &\leq \left(\sum_{i=1}^{3} \vartheta_i\, |f(u_j(x)) - f(u_j(q_i))| \right)^2 \\ &\leq 3 \sum_{i=1}^{3} |f(u_j(x)) - f(u_j(q_i))|^2 \leq 3\, C_\alpha^2 \sum_{i=1}^{3} |u_j(x) - u_j(q_i)|^{2\alpha}\,. \end{aligned} \tag{4.49}$$

Using the mean value theorem

$$|u_j(x) - u_j(q_i)| \leq |\nabla u_j| |x - q_i|$$

on the triangle t (with the Euclidean norm $|\cdot|$ on $\mathbb{R}^d$) while considering that $|\nabla u_j|$ is constant on t, we can go on estimating the last term in (4.49) to obtain

$$|f(u_j(x)) - I_{\mathcal{S}_j} f(u_j)(x)|^2 \leq 9\, C_\alpha^2\, |\nabla u_j|^{2\alpha}\, h_j^{2\alpha}$$

with h_j as in (4.40). Now, integration over Ω provides

$$\begin{aligned} \int_\Omega |f(u_j(x)) - I_{\mathcal{S}_j} f(u_j)(x)|^2\, dx &\leq \sum_{t \in \mathcal{T}_j} \int_t |f(u_j(x)) - I_{\mathcal{S}_j} f(u_j)(x)|^2\, dx \\ &\leq 9\, C_\alpha^2\, h_j^{2\alpha} \int_\Omega (|\nabla u_j|^2 + 1)\, dx\,. \end{aligned}$$

Since $(u_j)_{j \geq 0}$ converges in $H^1(\Omega)$, the last integral is uniformly bounded and, therefore, this whole last term tends to 0 as $j \to \infty$ due to (4.40). □

Note that in the proof we also obtained the order of convergence $\mathcal{O}(h_j^\alpha)$ for (4.48). We remark that due to the Sobolev embedding theorem, Lemmas 4.7, 4.8 as well as 4.11 also hold in one space dimension if $L^2(\Omega)$ and $L^{2/\alpha}(\Omega)$ are replaced by $(C(\overline{\Omega}), \|\cdot\|_\infty)$. In this case the assertions of Lemmas 4.7 and 4.11 are already satisfied for any uniformly continuous $M : \mathbb{R} \to \mathbb{R}$. Again, Lemma 4.11 can also be applied to $M : [u_c, \infty) \to \mathbb{R}$ for our $u_j \in \mathcal{K}_j$, $j \geq 0$, and $u \in \mathcal{K}$. Now, as a consequence of Lemmas 4.7, 4.8, 4.10 and 4.11 we obtain

Theorem 4.12. *We assume the conditions in Theorem 4.6 to be satisfied. Then we have the convergence*

$$\theta_j(p_j) := I_{\mathcal{S}_j} M(u_j) \to \theta(p) = M(u) \quad \textit{in } L^2(\Omega) \textit{ for } j \to \infty$$

of the discrete saturation if $M : [u_c, \infty) \to \mathbb{R}$ (or $M : \mathbb{R} \to \mathbb{R}$) is Hölder continuous, or bounded and uniformly continuous. In the nondegenerate case $kr(\cdot) \geq c > 0$ we also have the convergence

$$p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j) \to p = \kappa^{-1}(u) \quad \textit{in } L^2(\Omega) \textit{ for } j \to \infty \tag{4.50}$$

of the discrete physical pressure.

It is clear that if $kr(\cdot)$ can become arbitrarily small, leading to a singularity of κ^{-1} as in Figure 3, we do not even know if $\kappa^{-1}(u_j)$, $\kappa^{-1}(u) \in L^1(\Omega)$, $j \geq 0$. However, if (4.30) and (3.20) lead to physically realistic solutions p_j and $p = \kappa^{-1}(u)$, respectively, then these solutions should be uniformly bounded since arbitrarily big physical pressures in porous media are unnatural. Then, u_j, $j \geq 0$, and u are uniformly bounded away from u_c which is the same situation as in the nondegenerate case. Therefore, we can conclude (4.47) and (4.50) here, too.

Note that in the proof of Theorem 4.11 we also obtained the order of convergence $\mathcal{O}(h_j^\alpha)$ for (4.48). Therefore, altogether we get

$$\|I_{\mathcal{S}_j} M(u_j) - M(u)\|_{L^2(\Omega)} \leq C_\alpha(h_j^\alpha + \|u_j - u\|_1^\alpha)$$

and

$$\|I_{\mathcal{S}_j} \kappa^{-1}(u_j) - \kappa^{-1}(u)\|_{L^2(\Omega)} \leq C_1(h_j + \|u_j - u\|_1)$$

for the situation in Theorem 4.12. Thus, the convergence $p_j \to p$ in the L^2-norm is of order $\mathcal{O}(h_j)$ if the convergence $u_j \to u$ in the H^1-norm has this order.

Although we cannot prove H^1-convergence of the discrete saturation $\theta_j(p_j)$ and the discrete physical pressure p_j we present a numerical example for a (degenerate!) Brooks–Corey case in Section 6 in which much more can be observed. In this example we obtain numerically the order $\mathcal{O}(h_j^2)$ for both $u_j \to u$ and $p_j \to p$ in the L^2-norm and the order $\mathcal{O}(h_j)$ for both $u_j \to u$ and $p_j \to p$ in the H^1-norm. This is the best one can expect since it is already optimal for linear cases.

5. Monotone multigrid

In this section we give a short description of the algebraic solver that we use for the discrete minimization problems (4.29)—the monotone multigrid method. The standard reference for the numerical treatment of problems like (4.29) by monotone multigrid methods is [29]. Therefore, we restrict ourselves to a short presentation of such methods and mention the special situation given by the use of the Brooks–Corey functions. Moreover, we note that the algebraic theory allows Φ in (4.28) to depend explicitly on $q \in \mathcal{N}_j$.

The smoother used in the monotone multigrid is the nonlinear Gauss–Seidel method. Starting with a given iterate, this method minimizes the convex functional

$$F_j := \mathcal{J}(\cdot) + \phi_j(\cdot) \tag{5.51}$$

successively in the directions of the nodal basis functions $\lambda_q^{(j)}$ for $q \in \mathcal{N}_j \backslash \mathcal{N}_j^D$. By considering the subdifferential of (5.51) and thanks to the pointwise structure of the discrete convex functional (4.28), this leads to successive one-dimensional problems of finding the zero of the functions

$$M(\cdot)\, h_q + g_q(\cdot)$$

where $g_q(\cdot)$ are affine functions which already occur in the linear Gauss–Seidel method. Here, $M(\cdot)$ has to be extended to a monotone graph in u_c (and in 0 if $q \in \mathcal{N}_j^S$). The smoother on the fine grid guarantees convergence of the multigrid. More concretely, with the assumptions in Theorem 3.3 one can prove global convergence of the nonlinear Gauss–Seidel method to the discrete solution u_j of the minimization problem (4.29), even without imposing continuity on M.

To speed up the convergence, a coarse grid correction is carried out after presmoothing with the nonlinear Gauss–Seidel method. In the smoothed iterate $u_j^* \in \mathcal{S}_j$, given in the nodal basis, we skip all coordinates that assume the value $u_c < -1$ or $p_b = -1$ or else 0 if $q \in \mathcal{N}_j^S$, thus restricting u_j^* to the resulting subspace $\mathcal{S}_j^\circ \subset \mathcal{S}_j$. Then the coarse grid correction is restricted to $\mathcal{S}_j^\circ$, which is the largest subspace $V \subset \mathcal{S}_j$ spanned by nodal basis functions, such that F_j restricted to V is C^2 around the corresponding restriction of u_j^* on V. On this space we can approximate F_j by a quadratic model as in a standard Newton method. In order to ensure that the coarse grid correction does not lead to phase changes, we impose u_c and $p_b = -1$ as bound constraints. (On Signorini nodes, 0 is additionally imposed as an upper bound.) On this quadratic constraint minimization problem one iteration of a monotone multigrid method for quadratic problems is then performed [29].

Additional damping, locally for each coarse grid direction, ensures that the iterates provided by the coarse grid correction lead to further decreasing energy F_j and thus convergence. In the following numerical examples we use this way to construct coarse grid corrections with this property. For an alternative approach based on a non-smooth Newton method see [24].

Under some technical and non-degeneracy conditions one can prove that the coarse grid corrections of the monotone multigrid eventually become Newton multigrid steps applied to smooth problems for which convergence rates can be derived if an appropriate norm depending on u_j is considered. Concretely, one can prove that these asymptotic convergence rates only degenerate very mildly with j. In addition, these rates only depend on the initial triangulation and on the ellipticity constant of the bilinear form. In particular, they do not depend on the slope of M and do not change if the parameter functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$ and, consequently, $u \mapsto M(u)$ degenerate. In Section 7 we will demonstrate this robustness of the monotone multigrid performance with respect to extremely varying soil parameters.

6. Numerical example in 2D: discretization error

This section is devoted to adding a quantitative flavour to Theorems 4.6, 4.9 and 4.12. These results only state the convergence of $u_j \to u$, $\kappa^{-1}(u_j) \to p$ and $p_j \to p$ for $j \to \infty$, respectively, while leaving the order of convergence open. Here we want to determine the order numerically for an example in two space dimensions.

We consider the function

$$(x, y) \mapsto \tilde{p}(x, y) = 0.1 - 10\,(x^2 + y^2)$$

on the quadrilateral $[0,2] \times [0,1]$ and set

$$f := n\,\theta(\tilde{p}) - \operatorname{div}\Big(K_h kr(\theta(\tilde{p}))\nabla\tilde{p}\Big).$$

One can regard $\tilde{p}$ as a stationary solution of a corresponding time-discretized Richards equation (1.1) without gravity with the time step size $\tau = 1$. For simplicity, we use $\lambda = 1.0$ and $p_b = -0.1\,[m]$ as the Brooks–Corey parameters in this model problem. These, and the other parameters in Table 1, are in a realistic range of sandy soils (see [31]).

n	θ_m	θ_M	λ	p_b	K_h
0.38	0.21	0.95	1.0	$-0.1\,[m]$	$2 \cdot 10^{-3}\,[m/s]$

TABLE 1. Soil parameters for the model problem

Now we approximate $\tilde{p}$ by solving the discretized equation

$$n\,\theta(p) - \operatorname{div}\Big(K_h kr(\theta(p))\nabla p\Big) = f$$

in the finite element space $\mathcal{S}_j$ as described above and determine discrete solutions u_j and p_j for $j = 1, \ldots, 11$ with monotone multigrid. We choose Dirichlet boundary conditions on $\partial\Omega$ and $I_{\mathcal{S}_j}\tilde{p}$ as the initial iterate. We start with a uniform coarse triangular grid for $j = 1$ with 15 nodes and obtain the higher levels by uniform refinement. This leads to 8, 394, 753 nodes on the finest level.

The exact solution is a paraboloid directed downwards, and we have full saturation $\theta(\tilde{p}) = \theta_M$ on a disc around the origin with the radius $\sqrt{0.02} \approx 0.14$ only, so that a large part of the domain is dominated by the nonlinear nature of the problem. Also note that due to the choice of the domain the problem is not radially symmetric.

As one can see in Figures 5 and 6, we observe an order of convergence $\mathcal{O}(h_j^2)$ for both $u_j \to \tilde{u} = \kappa(\tilde{p})$ and $p_j \to \tilde{p}$ as $j \to \infty$, if we measure the convergence in the L^2-norm. Figures 7 and 8 show that with the H^1-norm we only obtain an order of convergence $\mathcal{O}(h_j)$, which one might expect from the result for the L^2-norm, and which is optimal even for linear problems.

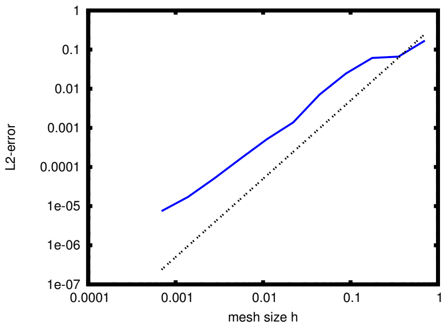

FIGURE 5. L^2-error in u (dotted line: $\mathcal{O}(h^2)$)

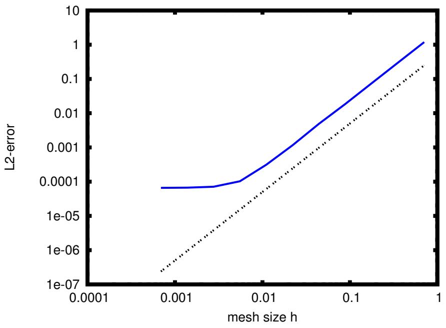

FIGURE 6. L^2-error in p (dotted line: $\mathcal{O}(h^2)$)

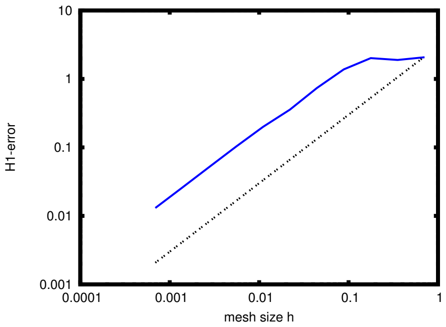

FIGURE 7. H^1-error in u (dotted line: $\mathcal{O}(h)$)

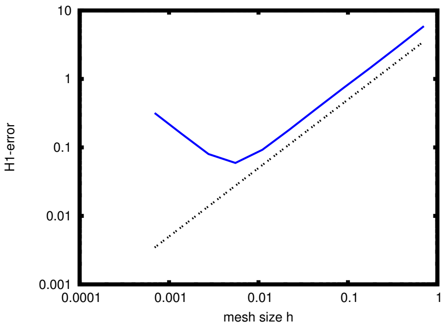

FIGURE 8. H^1-error in p (dotted line: $\mathcal{O}(h)$)

The anomalous behaviour of the curves corresponding to the physical pressure for small mesh sizes can be completely explained by the ill-conditioning of the inverse Kirchhoff transformation $\kappa^{-1} : (u_c, \infty) \to \mathbb{R}$ around u_c. This means that small differences $u_j - u$ of values in the neighbourhood of u_c are magnified by κ^{-1} to yield large differences $p_j - p$. By the concrete form of κ^{-1} we can even substantiate this phenomenon quantitatively, and thus confirm the behaviour of the curves in Figures 6 and 8. This shall be done in the following.

Concretely, with regard to the constant asymptotic behaviour of the L^2-error in p in Figure 6 note that if $\bar{u}$ is an approximation of $\tilde{u}$ up to the numerical accuracy of

$$|\bar{u}(x,y) - \tilde{u}(x,y)| = 10^{-16} \quad \text{on } \Omega$$

the square of this error is only given up to an accuracy of

$$0.01 \int_\Omega |\kappa^{-1}(\bar{u}(x,y)) - \kappa^{-1}(\tilde{u}(x,y))|^2 \, dx \, dy = 10^{-34} \int_\Omega |(\kappa^{-1})'(u(x,y))|^2 \, dx \, dy \tag{6.52}$$

with suitable $u(x,y)$ between $\bar{u}(x,y)$ and $\tilde{u}(x,y)$. (The factor 0.01 enters (6.52) because we treat u in the unit $|p_b|$, whereas the unit for p is given by $[m]$, i.e., meters of a water column.) Now, we have

$$(\kappa^{-1})'(u) = \frac{1}{\kappa'(\kappa^{-1}(u))} = \frac{1}{\kappa'(p)} = \frac{1}{kr(\theta(p))} = (-10\,p)^5$$

for $p \leq p_b = -0.1$ due to (2.4) and the choice of $\lambda = 1$. If we insert this into (6.52) for $p = \tilde{p}$ we can get an estimation of the numerical accuracy for the square of the L^2-error in p by considering the integral only on the right half of the quadrilateral Ω where we have $x^2 + y^2 \geq 1$. Therefore, we obtain the estimate

$$10^{-34} \int_0^1 \int_1^2 (100\,(x^2+y^2) - 1)^{10} \, dx \, dy \\ \geq 10^{-34}\, 99^{10} \int_0^1 \int_1^2 (x^2+y^2)^{10} \, dx \, dy \approx 5 \cdot 10^{-5}$$

for the numerical accuracy that we can expect for the L^2-error in p. In fact, the L^2-error in p on levels 9, 10 and 11 is already around $7 \cdot 10^{-5}$ as one can see in Figure 6.

Consequently, the H^1-error in p raises from level 9 to 10 and from level 10 to 11 by a factor of 2, since the numerical accuracy of these terms is given by the numerical accuracy of the L^2-error in p divided by the horizontal mesh size $h_j/\sqrt{2}$. For example, on the 11th level, with $h_{11}/\sqrt{2} = 2^{-11}$ and the numerical accuracy of $7 \cdot 10^{-5}$ for the L^2-error we obtain 0.14 as an estimate for the numerical accuracy of the H^1-error in p on level 11. In fact, here we obtain the H^1-error 0.32 as one can see in Figure 8. This explains the asymptotic behaviour displayed in that graphics and even confirms its order of magnitude numerically.

We find this example instructive since it illustrates how strongly the ill-conditioning of the inverse Kirchhoff transformation can influence practical problems. We point out that this ill-conditioning is part of the problem, i.e., a measure for the degeneracy of the Richards equation (1.1), and has to be dealt with in one way or the other within any solution process. The advantage of our approach is the separation of this ill-conditioning from the solution process. The ill-conditioning occurs only once, in form of the inverse Kirchhoff transformation, after the solution has already been obtained in generalized variables.

7. Numerical example in 3D: robustness of the solver

In this last section we illustrate that our discretization is solver friendly in the sense that our monotone multigrid method sketched in Section 5 exhibits a good convergence behaviour as applied to the resulting discrete minimization problems (4.29). Moreover, the convergence speed turns out to be robust with respect to soil parameters. As for the preceeding example, the implementation has been performed in the numerics environment DUNE [5] using the grid manager from UG [4]. For the visualization of the results we made use of the toolbox AMIRA [35].

7.1. Dam problem.

We consider the following problem with the coarse grid of a dam (consisting of prisms and hexahedra) as displayed in Figure 9. The dam has a constant width on the bottom and a constant maximal height, both equal to 9.81 $[m]$. Its length is 4 times this value. It is assumed to be filled with (homogeneous) sand only. The material parameters are listed in Table 2 which we obtained from [31].

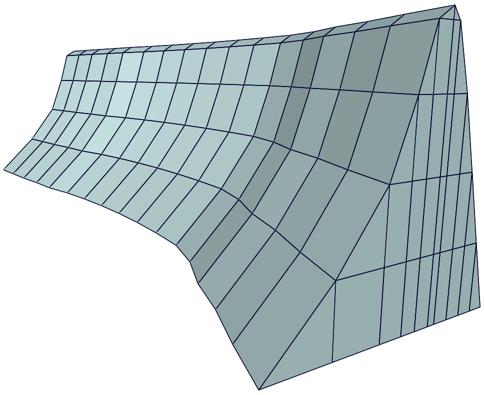

Figure 9. Coarse (prism) grid of the dam

n	θ_m	θ_M	λ	p_b	K_h
0.437	0.0458	1.0	0.694	$-0.0726\,[m]$	$6.54 \cdot 10^{-5}\,[m/s]$

Table 2. Soil parameters of sand for the dam problem

Starting with the constant initial pressure $p_0 = -10\,[m]$, which by (2.2) and the parameters in Table 2 corresponds to an almost dry dam with the initial saturation $\theta_0 = 0.0771$, we solve the Richards equation with gravity. We impose mixed boundary conditions of Dirichlet, Neumann and Signorini type. More concretely, a constant sea level of the maximal height $9.81\,[m]$ of the dam on the front side (left in Figure 9) leads to Dirichlet boundary conditions by hydrostatic pressure. The small faces of the dam as well as its bottom side are assumed to be impermeable giving rise to homogeneous Neumann boundary conditions. Finally, on the back side water may (and eventually will) flow out so that we have an outflow or Signorini-type condition (2.13) there.

As a result, water infiltrates until a fully saturated dam with an overall nonnegative pressure, i.e., a stationary state, is reached. With the time step size $\tau = 2.5\,[s]$ this takes 106 time steps. See Figures 10–17 for the evolution of the wetting front ($p = p_b$) on the left and colour plots of the physical pressure (between -10 and 9.81) on a vertical cut through the dam (situated at about a third of the dam length from the left small face).

The space discretization is carried out by first order Lagrangian finite elements (compare Figure 9 and 17 for the coarse grid). We have four refinement levels with $216,849$ nodes on the finest level. The monotone multigrid starts with the function obtained by nested iteration and stops as soon as the relative distance of succeeding iterates u^{k-1}, u^k in the H^1-seminorm $|\cdot|_1$ satisfies

$$\frac{|u^k - u^{k-1}|_1}{|u^{k-1}|_1} < 10^{-13}\,. \tag{7.53}$$

Let u^n be the last iterate. Then for each time step we calculate the multigrid convergence rate as the geometric mean of the rates

$$\frac{|u^k - u^n|_1}{|u^{k-1} - u^n|_1}\,, \quad k = 1, \ldots, n-1\,, \tag{7.54}$$

setting the rate equal to 0 if $n \leq 2$. Figure 18 shows the multigrid convergence rates for the different spatial problems over the time steps $j = 1, \ldots, 106$. As a result, the maximal rate for all time steps is $\rho_{max} = 0.35$ and the average rate is about $\rho_{av} = 0.23$ in this example, which we find a quite good performance of the multigrid solver. We point out that these rates do not differ much from the rate $\rho_{lin} = 0.21$ in the linear case, which is a Darcy problem that has to be solved at the end of the evolution when a stationary state is reached and the dam is fully saturated.

7.2. **Robustness with respect to soil parameters.** In the following, we illustrate the robust behaviour of the multigrid solver with respect to the soil parameters which enter the nonlinearities, i.e., the (negative) bubbling pressure p_b and the pore size distribution factor λ in the Brooks–Corey case. Concretely, we fix the time step size $\tau = 2.5\,[s]$ and the initial condition $\theta_0 = 0.0771$ as well as the parameters given

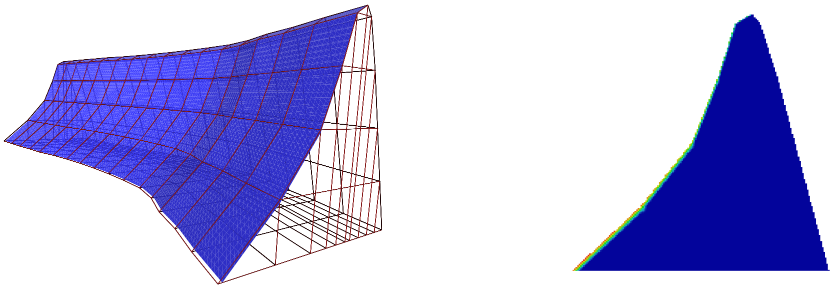

Figure 10. $t = 0\,s$

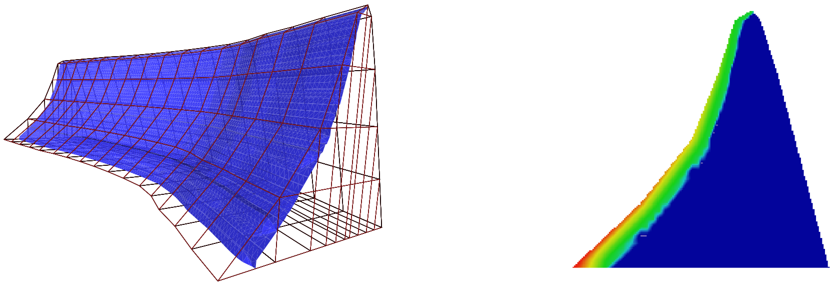

Figure 11. $t = \tau = 2.5\,s$

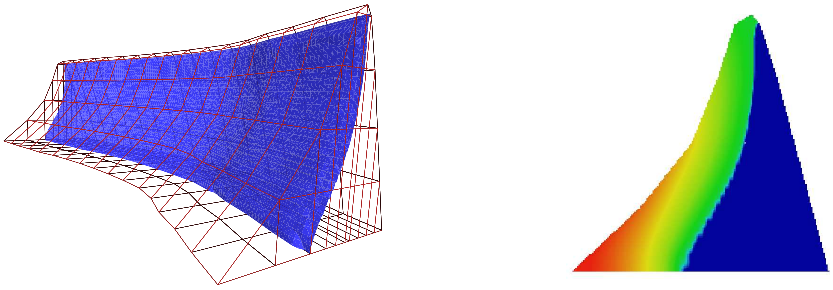

Figure 12. $t = 10\,\tau = 25\,s$

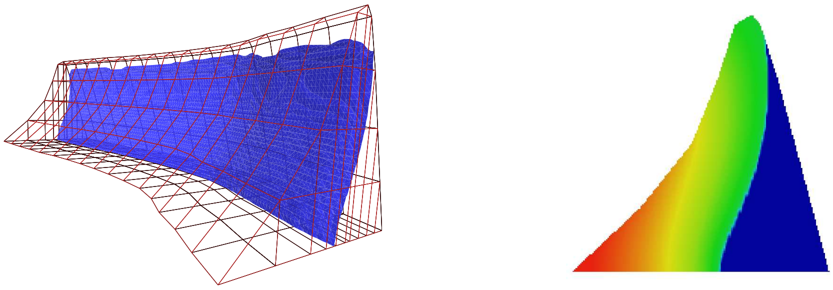

Figure 13. $t = 20\,\tau = 50\,s$

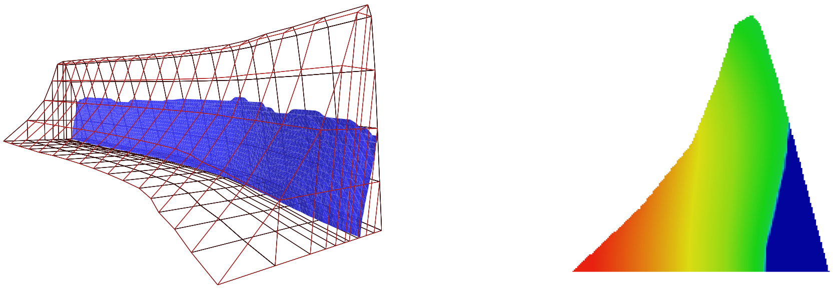

FIGURE 14. $t = 40\,\tau = 100\,s$

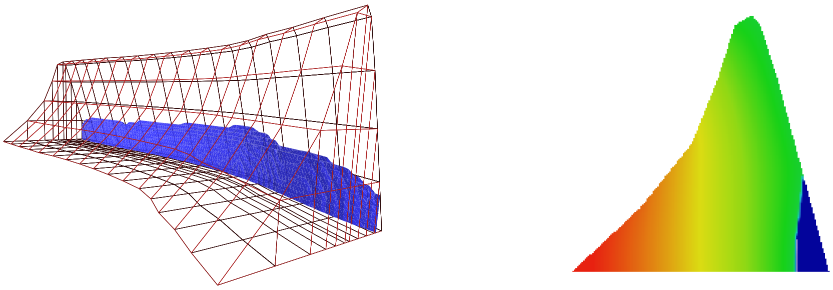

FIGURE 15. $t = 60\,\tau = 150\,s$

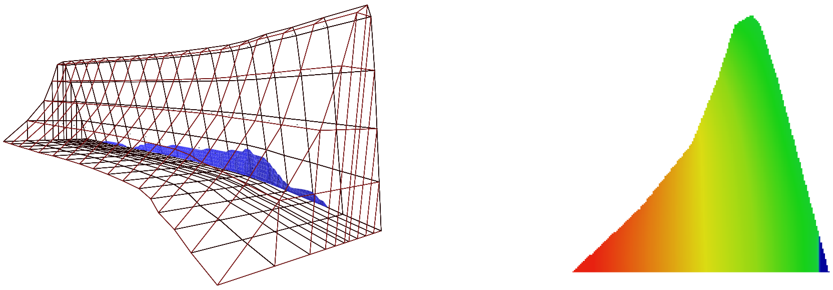

FIGURE 16. $t = 80\,\tau = 200\,s$

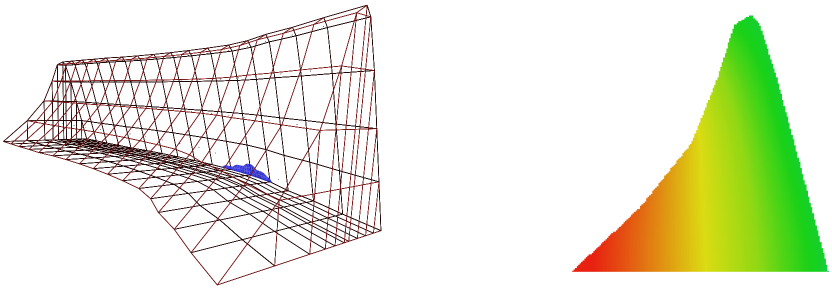

FIGURE 17. $t = 100\,\tau = 250\,s$

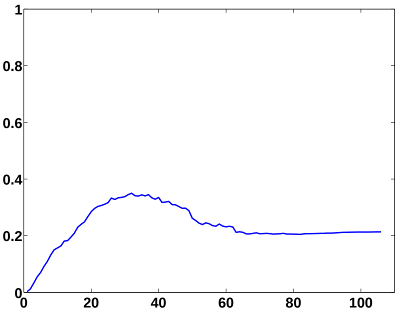

FIGURE 18. Multigrid convergence rates over the time steps $j = 1, \dots, 106$ for evolution with soil parameters as in Table 2

in Table 2 apart from p_b or λ. We vary $-p_b$ or λ, respectively, within a large range of the decimal powers between 10^{-10} and 10^{10} and, in addition, on the intervals $[0.01, 0.1]$, $[0.1, 1]$ and $[1, 10]$, each subdivided in 10 subintervals with equal length, which represent a hydrologically realistic range (compare [31, Table 5.3.2]). We have computed the evolution for each case until a stationary state with a fully saturated dam has been reached. This takes between 2 and 115 time steps. For each time step we calculated the multigrid convergence rates according to (7.53) and (7.54) as above. Then we determined the maximum ρ_{max} and the average ρ_{av} of these rates for each evolution. Observe that the saturation $\theta(p)$ as a function $M(u) = \theta(\kappa^{-1}(u))$ of u degenerates to step functions for $\lambda \to 0$, $\lambda \to \infty$ or $p_b \to 0$. As a consequence, variation of λ and $-p_b$ over 20 orders of magnitude requires considerable care to obtain a numerically stable implementation of $M(u)$. For example, already for $\lambda = 10^{-4}$ the interval $|u - u_c| < 10^{-200}$ covers $0 - 95\%$ of full saturation.

Figures 19 and 20 show the maximal and, as a dashed line, the average convergence rates ρ_{max} and ρ_{av} per evolution for varying λ and $-p_b$, respectively. In light of the preceeding remarks, we cannot rule out that the oscillations occurring

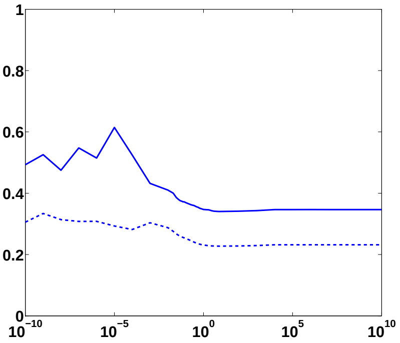

FIGURE 19. ρ_{max} and ρ_{av} vs. variation of λ

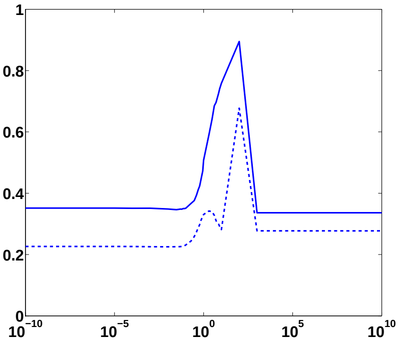

FIGURE 20. ρ_{max} and ρ_{av} vs. variation of $-p_b$

in Figure 19 for (completely unphysical) values $\lambda < 10^{-4}$ are due to numerical instabilities. In Figure 20 one can see a peak with unusually big convergence rates of about 0.9 for (unphysical) values $-p_b \approx 10^2\,[m]$. It seems that for these cases nested iteration does not provide an initial iterate which is accurate enough to enter the fast asymptotic regime of monotone multigrid convergence immediately (cf. [29]). Nevertheless, our extensive numerical experiments reveal that for a wide variation of soil parameters λ and p_b the monotone multigrid solver exhibits good convergence rates which are often comparable to the linear self-adjoint case.

References

[1] H.W. Alt and S. Luckhaus. Quasilinear elliptic–parabolic differential equations. *Math. Z.*, 183:311–341, 1983.

[2] H.W. Alt, S. Luckhaus, and A. Visintin. On nonstationary flow through porous media. *Ann. Math. Pura Appl.*, 136:303–316, 1984.

[3] J. Appell and P.P. Zabrejko. *Nonlinear superposition operators.* Cambridge University Press, 1990.

[4] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuß, H. Rentz–Reichert, and C. Wieners. UG – A flexible software toolbox for solving partial differential equations. *Comput. Vis. Sci.*, 1(1):27–40, 1997.

[5] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. *Computing*, 82(2-3):121–138, 2008.

[6] P. Bastian, O. Ippisch, F. Rezanezhad, H.J. Vogel, and K. Roth. Numerical simulation and experimental studies of unsaturated water flow in heterogeneous systems. In W. Jäger, R. Rannacher, and J. Warnatz, editors, *Reactive Flows, Diffusion and Transport*, pages 579–598. Springer, 2005.

[7] J. Bear. *Dynamics of Fluids in Porous Media.* Dover Publications, 1988.

[8] H. Beaugendre, A. Ern, T. Esclaffer, E. Gaume, I. Ginzburg, and C. Kao. A seepage face model for the interaction of shallow water tables with the ground surface: Application of the obstacle-type method. *J. Hydrol.*, 329: 258–273, 2006.

[9] H. Berninger. *Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation.* PhD thesis, Freie Universität Berlin, 2008.

[10] H. Berninger. Non-overlapping domain decomposition for the Richards equation via superposition operators. In *Domain Decomposition Methods in Science and Engineering XVIII*, LNCSE. Springer, 2009.

[11] H. Berninger, R. Kornhuber, and O. Sander. Solution of the Richards equation in heterogeneous soil. To appear.

[12] F. Brezzi and G. Gilardi. Functional spaces. In H. Kardestuncer and D.H. Norrie, editors, *Finite Element Handbook*, chapter 2 (part 1), pages 1.29–1.75. Springer, 1987.

[13] R.J. Brooks and A.T. Corey. Hydraulic properties of porous media. Technical Report Hydrology Paper No. 3, Colorado State University, Civil Engineering Department, Fort Collins, 1964.

[14] N.T. Burdine. Relative permeability calculations from pore-size distribution data. *Petr. Trans., Am. Inst. Mining Metall. Eng.*, 198:71–77, 1953.

[15] G. Chavent and J. Jaffré. *Dynamics of Fluids in Porous Media.* Elsevier Science, 1986.

[16] M. Chipot and A. Lyaghfouri. The dam problem for non-linear Darcy's laws and non-linear leaky boundary conditions. *Math. Methods Appl. Sci.*, 20(12): 1045–1068, 1997.

[17] T.M. Chui and D.L. Freyberg. The use of COMSOL for integrated hydrological modeling. In *Proceedings of the COMSOL Conference 2007 Boston*, pages 217–223.

[18] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North–Holland, 1978.

[19] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems.* North–Holland, 1976.

[20] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. *Comput. Geosci.*, 3(3–4):259–294, 1999.

[21] M.W. Farthing, C.E. Kees, T.S. Coffey, C.T. Kelley, and C.T. Miller. Efficient steady-state solution techniques for variably saturated groundwater flow. *Adv. Water Resour.*, 26(8):833–849, 2003.

[22] J. Fuhrmann. On numerical solution methods for nonlinear parabolic problems. In R. Helmig, W. Jäger, W. Kinzelbach, P. Knabner, and G. Wittum, editors, *Modeling and Computation in Environmental Sciences, First GAMM-Seminar at ICA Stuttgart*, pages 170–180. Vieweg, 1997.

[23] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer, 1984.

[24] C. Gräser, U. Sack, and O. Sander. Truncated nonsmooth Newton multigrid methods for convex minimization problems. In *Domain Decomposition Methods in Science and Engineering XVIII*, LNCSE. Springer, 2009.

[25] C.E. Kees, M.W. Farthing, S.E. Howington, E.W. Jenkins, and C.T. Kelley. Nonlinear multilevel iterative methods for multiscale models of air/water flow in porous media. In P.J. Binning, P.K. Engesgaard, H.K. Dahle, G.F. Pinder, and W.G. Gray, editors, *Proceedings of Computational Methods in Water Resources XVI*, 8 pages, Copenhagen, Denmark, June 2006.

[26] N. Kikuchi and J.T. Oden. *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods.* SIAM, 1988.

[27] G. Koethe. *Topologische lineare Räume*, volume I. Springer, 2nd edition, 1966.

[28] R. Kornhuber. *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems.* Teubner, 1997.

[29] R. Kornhuber. On constrained Newton linearization and multigrid for variational inequalities. *Numer. Math.*, 91:699–721, 2002.

[30] M. Marcus and V.J. Mizel. Every superposition operator mapping one Sobolev space into another is continuous. *J. Funct. Anal.*, 33:217–229, 1979.

[31] W.J. Rawls, L.R. Ahuja, D.L. Brakensiek, and A. Shirmohammadi. Infiltration and soil water movement. In D.R. Maidment, editor, *Handbook of Hydrology*, chapter 5. McGraw–Hill, 1993.

[32] L.A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1:318–333, 1931.

[33] E. Schneid, P. Knabner, and F. Radu. A priori error estimates for a mixed finite element discretization of the Richards' equation. *Numer. Math.*, 98(2): 353–370, 2004.

[34] B. Schweizer. Regularization of outflow problems in unsaturated porous media with dry regions. *J. Differ. Equations*, 237(2):278–306, 2007.

[35] D. Stalling, M. Westerhoff, and H.-C. Hege. Amira: A highly interactive system for visual data analysis. In C. Hansen and C. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.

[36] M.I.J. van Dijke and S.E.A.T.M. van der Zee. Analysis of oil lens removal by extraction through a seepage face. *Comput. Geosci.*, 2(1):47–72, 1998.

[37] M.T. van Genuchten. A closed–form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.*, 44:892–898, 1980.

[38] C. Wagner, G. Wittum, R. Fritsche, and H.-P. Haar. Diffusions–Reaktionsprobleme in ungesättigten porösen Medien. In K.-H. Hoffmann et al., editor, *Mathematik: Schlüsseltechnologie für die Zukunft. Verbundprojekte zwischen Universität und Industrie.*, pages 243–253. Springer, 1997.

[39] D. Werner. *Funktionalanalysis*. Springer, 5th edition, 2005.

Heiko Berninger, Ralf Kornhuber, Oliver Sander, Freie Universität Berlin, Institut für Mathematik, Arnimallee 6, 14195 Berlin, Germany

E-mail address: berninger|kornhuber|sander@math.fu-berlin.de