Comput Geosci (2016) 20:341–353
DOI 10.1007/s10596-016-9566-3

ORIGINAL PAPER

A study on iterative methods for solving Richards' equation

Florian List[1] · **Florin A. Radu**[2]

Received: 29 July 2015 / Accepted: 7 October 2015 / Published online: 15 March 2016

Abstract This work concerns linearization methods for efficiently solving the Richards equation, a degenerate elliptic-parabolic equation which models flow in saturated/unsaturated porous media. The discretization of Richards' equation is based on backward Euler in time and Galerkin finite elements in space. The most valuable linearization schemes for Richards' equation, i.e. the Newton method, the Picard method, the Picard/Newton method and the $L-$scheme are presented and their performance is comparatively studied. The convergence, the computational time and the condition numbers for the underlying linear systems are recorded. The convergence of the $L-$scheme is theoretically proved and the convergence of the other methods is discussed. A new scheme is proposed, the $L-$scheme/Newton method which is more robust and quadratically convergent. The linearization methods are tested on illustrative numerical examples.

Keywords Richards' equation · Linearization schemes · Newton method · Picard method · Convergence analysis · Flow in porous media · Galerkin finite elements

✉ Florin A. Radu
Florin.Radu@math.uib.no

[1] Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Stuttgart, Germany

[2] Department of Mathematics, University of Bergen, Bergen, Norway

1 Introduction

There are plenty of societal relevant applications of multiphase flow in porous media, e.g. water and soil pollution, CO_2 storage, nuclear waste management, or enhanced oil recovery, to name a few. Mathematical modelling and numerical simulations are powerful, well-recognized tools for predicting flow in porous media and therefore for understanding and finally solving problems like the ones mentioned above. Nevertheless, mathematical models for multiphase flow in porous media involve coupled, non-linear partial differential equations on huge, complex domains and with parameters which may vary on multiple order of magnitudes. Moreover, typical for the type of applications we mentioned are long term time evolutions, recommending the use of implicit schemes which allow large time steps. Due to these, the design and analysis of appropriate numerical schemes for multiphase flow in porous media is a very challenging task. Despite of intensive research in the last decades, there is still a strong need for robust numerical schemes for multiphase flow in porous media.

In this work we consider a particular case of two-phase flow: flow of water in soil, including the region near the surface where the pores are filled with water and air (unsaturated zone). By considering that the pressure of air remains constant, i.e. zero, water flow through saturated/unsaturated porous media is mathematically described by Richards' equation

$$\partial_t \theta(\Psi) - \nabla \cdot (K(\theta(\Psi))\nabla(\Psi + z)) = f, \qquad (1)$$

which has been proposed by L.A. Richards in 1930 (see e.g. [6]). In Eq. 1, Ψ denotes the pressure head, θ the water content, K stands for the hydraulic conductivity of the medium, z for the height against the gravitational direction and f for a source term. Based on experimental results, different curves have been proposed for describing the dependency between K, θ and Ψ (see e. g. [6]), yielding the non-linear model (1). In the saturated zone, i.e. where the pores are filled only with water, we have θ and K constants and $\Psi \geq 0$. Whenever the flow is unsaturated, θ and K are non-linear, monotone functions and $\Psi < 0$. We point out that Richards' equation degenerates when $K(\theta(\Psi)) \rightarrow 0$ (slow diffusion case) or when $\theta' = 0$ (fast diffusion case). The regions of degeneracy depend on the saturation of the medium; therefore these regions are not known a-priori and may vary in time and space. In this paper we concentrate on the fast diffusion case, therefore Richards' equation will be a non-linear, degenerate elliptic-parabolic partial differential equation. Typically for this case is also the low regularity of the solution [1]. The non-linearities and the degeneracy make the design and analysis of numerical schemes for the Richards' equation very difficult.

The first choice for the temporal discretization is the backward Euler method. There are two reasons for this: the need of a stable discretization allowing large time steps and the low regularity of the solution which does not support any higher order scheme. As regards the spatial discretization there are much more options possible. Galerkin finite elements were used in [2, 3, 9, 22, 24, 32], often together with mass lumping to ensure a maximum principle [8]. Locally mass conservative schemes for Richards' equation were proposed and analysed in [10, 11] (finite volumes), in [16] (multipoint flux approximation) or [4, 5, 26, 29, 34, 35] (mixed finite element method). The analysis is performed mostly by using the Kirchhoff transformation (which combines the two main non-linearities in one) [1, 4, 26, 28, 34] or, alternatively, by restricting the generality, e.g. to the strictly unsaturated case [3, 29]. To deal with the low regularity of the solution, a time integration together with the use of Green's operator is usually necessary [4, 22].

The systems to be solved in each time step after temporal and spatial discretization are non-linear and one needs an efficient and robust algorithm to solve them. The main linearization methods which are used for the Richards equation are: Newton (also called Newton–Raphson in the literature) method, Picard method, modified Picard method, the $L-$scheme, and combination of them. The Newton method, which is quadratically convergent was very successfully applied to Richards' equation in e.g. [7, 8, 20, 23, 27]. The drawback of Newton's method is that it is only locally convergent and involves the computation of derivatives. Although the use of the solution of the last time step to start the Newton iterations improves considerably the robustness of Newton's method, in the degenerate case (saturated/unsaturated flow) the convergence is ensured only when a regularization step is applied and under additional constraints on the discretization parameters, see [27] for details. The Picard method is, although widely used, not a good choice when applied to Richards' equation as clearly shown in [8, 20]. In [8] is proposed an improvement of the Picard method, resulting in a new method called modified Picard. This method coincides with the Newton method in the case of a constant conductivity K or when applied to Richards' equation together with the Kirchhoff transformation. The modified Picard method is only linearly convergent, but more robust than Newton's method. An efficient combination of the modified Picard and the Newton method, the Picard/Newton method is proposed in [20]. For the sake of completeness we mention also the accelerated Picard method [21] for Richards' equation and the semi-smooth Newton method [19] and $L-$method [30] for two-phase flow in porous media, as valuable linearization methods.

The $L-$method is the only method which uses the monotonicity of $\theta(\cdot)$. It was proposed for Richards' equation in [25, 32, 36]. The method is robust and linearly convergent, and does not involve the computation of any derivatives. Moreover, the convergence rate does not depend on the mesh size. The linear systems arising after using the L-scheme are much better conditioned than the corresponding systems for Newton or modified Picard methods. Due to these, the L-scheme is in many situations even faster than the Newton method, although being only linear convergent. In the case of a constant K or when applied to Richards' equation together with the Kirchhoff transformation, one can improve the convergence of the $L-$method by adaptively computing L, this being the idea of the Jäger–Kačur method [15]. The choice of the Jacobian matrix for L would lead to Newton's method, therefore in this case all the three methods (Newton, modified Picard and $L-$scheme) will coincide. It is worth to mention that both the $L-$method and the modified Picard method can be seen as quasi-Newton (or Newton-like) methods. We refer to [18] for a comprehensive presentation of Newton's method and its variants.

In order to combine the robustness of the L-method with the speed of Newton's method, we propose in this paper a mixed L-scheme/Newton. The idea is the same as in the case of the modified Picard/Newton method in [20]: compute a few iterations with the robust scheme (now the L-scheme) before switching to Newton's method. The new mixed method performs best w.r.t. robustness and computational time from the all considered linearization schemes, as our numerical tests are clearly showing.

To summarize, this paper concentrates on linearization methods for Richards' equation, and its new contributions are:

- A comprehensive study on the most valuable linearization methods for Richards' equation: the Newton method, the modified Picard, the Picard/Newton method and the $L-$scheme. The study includes numerical convergence, CPU time and condition number of the resulting linear systems.
- The design of a new scheme based on the $L-$scheme and Newton's method, the $L-$scheme/Newton method which is robust and quadratically convergent.
- Provides the theoretical proof for the convergence of the $L-$scheme for Richards' equation (without using the Kirchhoff transformation) and discuss the convergence of modified Picard and Newton methods. The analysis furnishes new insights and helps towards a deeper understanding of the linearization schemes.

The present paper can be seen as a continuation of the works [8] and [20], and it is written in a similar spirit. We added in the study the $L-$schemes (including a new scheme combining Newton's method with the $L-$scheme) and we focus now on 2D numerical results (the two mentioned papers based their conclusions on 1D simulations). We present illustrative examples, with realistic parameters so that the computations are relevant for practical applications.

The paper is structured as follows. In the next section the variational formulations (continuous and fully discrete) of Richards' equation are presented together with the considered linearization schemes. In Section 3 we discuss the theoretical convergence of the methods. The next section concerns the numerical results. The paper is ending with a concluding section.

2 Linearization methods for Richards' equation

Throughout this paper we use common notations in the functional analysis. Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 1, 2$ or 3, having a Lipschitz continuous boundary $\partial\Omega$. We denote by $L^2(\Omega)$ the space of real valued square integrable functions defined on Ω, and by $H^1(\Omega)$ its subspace containing functions having also the first order derivatives in $L^2(\Omega)$. Let $H_0^1(\Omega)$ be the space of functions in $H^1(\Omega)$ which vanish on $\partial\Omega$. Further, we denote by $\langle\cdot,\cdot\rangle$ the inner product on $L^2(\Omega)$, and by $\|\cdot\|$ the norm of $L^2(\Omega)$. L_f stays for the Lipschitz constant of a Lipschitz continuous function $f(\cdot)$.

We consider to solve the Richards Eq. 1 on $(0, T] \times \Omega$, with T denoting the final time and with homogeneous Dirichlet boundary conditions and an initial condition given by $\Psi(0,\mathbf{x}) = \Psi^0(\mathbf{x})$ for all $\mathbf{x} \in \Omega$. We will use linear Galerkin finite elements for this study, but the linearization methods considered can be applied to any discretization method. We restrict the formulations and analysis to homogeneous Dirichlet boundary conditions just for the sake of simplicity, the extension to more general boundary conditions being straightforward (the numerical examples in Section 4 involve general boundary conditions). The continuous Galerkin formulation of Eq. 1 reads as:

Find $\Psi \in H_0^1(\Omega)$ such that there holds

$$\langle \partial_t\theta(\Psi), \phi\rangle + \langle K(\theta(\Psi))\left(\nabla\Psi + \mathbf{e_z}\right), \nabla\phi\rangle = \langle f, \phi\rangle, \qquad (2)$$

for all $\phi \in H_0^1(\Omega)$, with $\mathbf{e_z} := \nabla z$. Results concerning the existence and uniqueness of solutions of Eq. 2 can be found in several papers, e.g. [1].

For the discretization in time we let $N \in \mathbb{N}$ be strictly positive, and define the time step $\tau = T/N$, as well as $t_n = n\tau$ ($n \in \{1, 2, \ldots, N\}$). Furthermore, $\mathcal{T}_h$ is a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh diameter. Here we assume $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$, hence Ω is assumed polygonal. The Galerkin finite element space is given by

$$V_h := \left\{ v_h \in H_0^1(\Omega) |\, v_{h|_T} \in \mathcal{P}_1(T), T \in \mathcal{T}_h \right\}, \qquad (3)$$

where $\mathcal{P}_1(T)$ denotes the space of linear polynomials on any simplex T. For details about this finite element space and the implementation we refer to standard books, like e.g. [18].

By using the backward Euler method in time and the linear Galerkin finite elements defined above in space, the fully discrete variational formulation of Eq. 2 at time t_n reads as:

Let $n \in \{1, \ldots, N\}$ and $\Psi_h^{n-1} \in V_h$ be given. Find $\Psi_h^n \in V_h$ such that there holds

$$\left\langle \theta\left(\Psi_h^n\right) - \theta\left(\Psi_h^{n-1}\right), v_h \right\rangle + \tau\langle K\left(\theta\left(\Psi_h^n\right)\right)\left(\nabla\Psi_h^n + \mathbf{e_z}\right), \nabla v_h\rangle = \tau\langle f^n, v_h\rangle, \quad (4)$$

for all $v_h \in V_h$. At the first time step we take $\Psi_h^0 = P_h\Psi^0 \in V_h$, with $P_h : H_0^1(\Omega) \to V_h$ being the standard projection. We assume in the next that the fully discrete schemes above have a unique solution and we refer to [2, 3, 9, 24] for a proof.

At this point, dealing with the doubly non-linear character of Richards' equation due to the relations $K(\theta)$ and $\theta(\Psi)$ is essential. We will briefly present in the following the main linearization methods used to solve the non-linear problem (4): the Newton method, the modified Picard method (called simply Picard's method below, when does not exist a possibility of confusion) and the $L-$schemes.

We denote the discrete solution at time level n (which is now fixed) and iteration $j \in \mathbb{N}$ by $\Psi_h^{n,j}$ henceforth. The iterations are always starting with the solution at the last time step, i.e. $\Psi_h^{n,0} = \Psi_h^{n-1}$. The Newton method to solve (4) reads as:

Let $\Psi_h^{n-1}, \Psi_h^{n,j-1} \in V_h$ be given. Find $\Psi_h^{n,j} \in V_h$, so that

$$\begin{aligned}&\left\langle \theta\left(\Psi_h^{n,j-1}\right), v_h\right\rangle + \left\langle \theta'\left(\Psi_h^{n,j-1}\right)\left(\Psi_h^{n,j} - \Psi_h^{n,j-1}\right), v_h\right\rangle\\ &+\tau\left\langle K\left(\Psi_h^{n,j-1}\right)\left(\nabla\Psi_h^{n,j} + \mathbf{e_z}\right), \nabla v_h\right\rangle\\ &+\tau\left\langle K'\left(\Psi_h^{n,j-1}\right)\left(\nabla\Psi_h^{n,j-1} + \mathbf{e_z}\right)\left(\Psi_h^{n,j} - \Psi_h^{n,j-1}\right), \nabla v_h\right\rangle\\ &= \tau\left\langle f^n, v_h\right\rangle + \left\langle \theta\left(\Psi_h^{n-1}\right), v_h\right\rangle, \end{aligned} \tag{5}$$

holds for all $v_h \in V_h$. Newton's method is quadratically, but only locally convergent. As mentioned above, although $\Psi_h^{n,0} := \Psi_h^{n-1}$ might be an appropriate choice, failure of Newton's method can occur (see [27] and the numerical examples given below).

The modified Picard method was proposed by [8] and reads:

Let $\Psi_h^{n-1}, \Psi_h^{n,j-1} \in V_h$ be given. Find $\Psi_h^{n,j} \in V_h$, so that

$$\begin{aligned}&\left\langle \theta\left(\Psi_h^{n,j-1}\right), v_h\right\rangle + \left\langle \theta'\left(\Psi_h^{n,j-1}\right)\left(\Psi_h^{n,j} - \Psi_h^{n,j-1}\right), v_h\right\rangle\\ &+\tau\left\langle K\left(\Psi_h^{n,j-1}\right)\left(\nabla\Psi_h^{n,j} + \mathbf{e_z}\right), \nabla v_h\right\rangle = \tau\langle f^n, v_h\rangle\\ &+\left\langle \theta\left(\Psi_h^{n-1}\right), v_h\right\rangle, \end{aligned} \tag{6}$$

holds for all $v_h \in V_h$. The modified Picard method was shown to perform much better than the classical Picard method [8, 20]. The idea is to discretize the time non-linearity quadratically, whereas the non-linearity in K is linearly approximated. The method is therefore linearly convergent. The method still involves the computation of derivatives and in the degenerate case might also fail to converge (see the numerical examples in Section 4).

The $L-$method was proposed for Richards' equation by [25, 32, 36] and it is the only method which exploits the monotonicity of $\theta(\cdot)$. The $L-$scheme to solve the non-linear problem (4) reads:

Let $\Psi_h^{n-1}, \Psi_h^{n,j-1} \in V_h$ and $L > 0$ be given. Find $\Psi_h^{n,j} \in V_h$, so that

$$\begin{aligned}&\left\langle \theta\left(\Psi_h^{n,j-1}\right), v_h\right\rangle + L\left\langle \Psi_h^{n,j} - \Psi_h^{n,j-1}, v_h\right\rangle\\ &+\tau\left\langle K\left(\Psi_h^{n,j-1}\right)\left(\nabla\Psi_h^{n,j} + \mathbf{e_z}\right), \nabla v_h\right\rangle\\ &= \tau\langle f^n, v_h\rangle + \left\langle \theta\left(\Psi_h^{n-1}\right), v_h\right\rangle, \end{aligned} \tag{7}$$

holds for all $v_h \in V_h$. To ensure the convergence of the scheme, the constant L should satisfy $L \geq L_\theta (:= \sup_\Psi |\theta'(\Psi)|)$ (see Section 3 for details). The $L-$scheme is robust and linearly convergent. Furthermore, the scheme does not involve the computation of any derivatives. The key element in the $L-$scheme is the addition of a stabilization term $L\left\langle \Psi_h^{n,j} - \Psi_h^{n,j-1}, v_h\right\rangle$, which together with the monotonicity of $\theta(\cdot)$ will ensure the convergence of the scheme.

Remark 1 We note that the scheme presented above is slightly different from the one considered in [32], where $K\left(\Psi_h^{n-1}\right)$ was taken instead of $K\left(\Psi_h^{n,j-1}\right)$. Moreover, the schemes [25, 36] are considering the Kirchhoff transformation, which is not the case in the present work.

Remark 2 It is to be seen that in the case of a constant K, the methods (5) and (6) coincide. Moreover, if L is replaced by the Jacobian matrix in Eq. 7 one obtains again the modified Picard scheme (6).

Any of the linearization methods presented above leads to a system of linear equations for $\Psi_h^{n,j}$ (more precisely, the unknown will be the vector $\mathbf{d}_h^{n,j}$ with the components of $\Psi_h^{n,j}$ in a basis of V_h). The derivatives of the water content and the hydraulic conductivity in case of the modified Picard scheme and Newton's method can be computed analytically or by a perturbation approach as suggested by [12] and occurring integrals are approximated by a quadrature formula.

For stopping the iterations, we adopt a general criterion for convergence given by

$$\left\|\mathbf{d}_h^{n,j} - \mathbf{d}_h^{n,j-1}\right\| \leq \varepsilon_a + \varepsilon_r \left\|\mathbf{d}_h^{n,j}\right\|, \tag{8}$$

with the Euclidean norm $\|\cdot\|$ and some constants $\varepsilon_a > 0$ and $\varepsilon_r > 0$. The tolerances ε_a and ε_r in criterion (8) are both taken as 10^{-5} in all numerical simulations in this paper as proposed by [20]. We refer to [14] for possible improvements of the stopping criterion.

The Newton method is the only method out of the proposed three which is second order convergent. Nevertheless, it is not that robust as the other, linearly convergent, methods. In order to increase the robustness of Newton's method one can perform first a few (modified) Picard iterations, this being the combined Picard/Newton scheme proposed in [20]. The Picard/Newton method is shown to perform better than both the Newton and the modified Picard method [20]. We propose in this paper also a combination of the $L-$scheme with the Newton method, the

$L-$scheme/Newton method. The mixed methods are based upon the idea to harness the robustness of the $L-$scheme or the modified Picard scheme initially and to switch to Newton's method e.g. if

$$\left\|\mathbf{d}_h^{n,j} - \mathbf{d}_h^{n,j-1}\right\| \leq \delta_a + \delta_r \left\|\mathbf{d}_h^{n,j}\right\|, \tag{9}$$

is satisfied for $\delta_a, \delta_r > 0$, similar to criterion (8). However, an appropriate choice of the parameters δ_a, δ_r is intricate and heavily dependent on the problem, for which reason a switch of the method after a fixed number of iterations may be an alternative. As shown in Section 4 this new method incorporating the $L-$scheme seems to perform best with respect to computing time and robustness.

3 Convergence results

In this section we will rigorously analyse the convergence of the $L-$scheme and discuss the convergence of Newton's and modified Picard's method. We denote by

$$e^{n,j} = \Psi_h^{n,j} - \Psi_h^n, \tag{10}$$

the error at iteration j. A scheme is convergent if $e^{n,j} \to 0$, when $j \to \infty$.

The following assumptions on the coefficient functions and the discrete solution are defining the framework in which we can prove the convergence of the $L-$scheme.

(A1) The function $\theta(\cdot)$ is monotonically increasing and Lipschitz continuous.

(A2) The function $K(\cdot)$ is Lipschitz continuous and there exist two constants K_m and K_M such that $0 < K_m \leq K(\theta) \leq K_M < \infty, \ \forall \theta \in \mathbb{R}$.

(A3) The solution of problem (4) satisfies $\|\nabla \Psi_h^n\|_\infty \leq M < \infty$, with $\|\cdot\|_\infty$ denoting the $L^\infty(\Omega)$-norm.

We can now state the central theoretical result of this paper.

Theorem 1 *Let $n \in \{1, \ldots, N\}$ be given and assume (A1) - (A3) hold true. If the constant L and the time step are chosen such that (16) below is satisfied, the $L-$scheme (7) converges linearly, with a rate of convergence given by*

$$\sqrt{\frac{L}{L + \frac{K_m \tau}{C_\Omega^2}}}. \tag{11}$$

Proof By subtracting (4) from (7) we obtain for any $v_h \in V_h$ and any $j \geq 1$

$$\begin{aligned}
&\left\langle \theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right), v_h \right\rangle + L\langle e^{n,j} - e^{n,j-1}, v_h\rangle \\
&+\tau\left\langle K\left(\Psi_h^{n,j-1}\right)\nabla\Psi_h^{n,j} - K\left(\Psi_h^n\right)\nabla\Psi_h^n, \nabla v_h\right\rangle \\
&+\tau\left\langle\left(K\left(\Psi_h^{n,j-1}\right) - K\left(\Psi_h^n\right)\right)\mathbf{e_z}, \nabla v_h\right\rangle = 0.
\end{aligned}$$

By testing the above with $v_h = e^{n,j}$ and doing some algebraic manipulations one gets

$$\begin{aligned}
&\left\langle \theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right), e^{n,j-1}\right\rangle \\
&+\left\langle \theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right), e^{n,j} - e^{n,j-1}\right\rangle \\
&+\frac{L}{2}\left\|e^{n,j}\right\|^2 + \frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|^2 - \frac{L}{2}\left\|e^{n,j-1}\right\|^2 \\
&+\tau\left\langle K\left(\Psi_h^{n,j-1}\right)\nabla e^{n,j}, \nabla e^{n,j}\right\rangle \\
&+\left\langle\left(K\left(\Psi_h^{n,j-1}\right) - K\left(\Psi_h^n\right)\right)\nabla\Psi_h^n, \nabla e^{n,j}\right\rangle \\
&+\tau\left\langle\left(K\left(\Psi_h^{n,j-1}\right) - K\left(\Psi_h^n\right)\right)\mathbf{e_z}, \nabla e^{n,j}\right\rangle = 0,
\end{aligned} \tag{12}$$

or, equivalently

$$\begin{aligned}
&\left\langle \theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right), e^{n,j-1}\right\rangle + \frac{L}{2}\left\|e^{n,j}\right\|^2 \\
&+\tfrac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|^2 + \tau\left\langle K\left(\Psi_h^{n,j-1}\right)\nabla e^{n,j}, \nabla e^{n,j}\right\rangle \\
&= \frac{L}{2}\left\|e^{n,j-1}\right\|^2 - \left\langle \theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right), e^{n,j} - e^{n,j-1}\right\rangle \\
&-\left\langle\left(K\left(\Psi_h^{n,j-1}\right) - K\left(\Psi_h^n\right)\right)\nabla\Psi_h^n, \nabla e^{n,j}\right\rangle \\
&-\tau\left\langle\left(K\left(\Psi_h^{n,j-1}\right) - K\left(\Psi_h^n\right)\right)\mathbf{e_z}, \nabla e^{n,j}\right\rangle.
\end{aligned} \tag{13}$$

By using now the monotonicity of $\theta(\cdot)$, its Lipschitz continuity (A1), the boundedness (from below) and Lipschitz continuity of $K(\cdot)$, i.e. (A2), the boundedness of $\nabla\Psi_h^n$, and the Young and Cauchy-Schwarz inequalities one obtains from Eq. 13

$$\begin{aligned}
&\frac{1}{L_\theta}\left\|\theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right)\right\|^2 + \frac{L}{2}\left\|e^{n,j}\right\|^2 \\
&+\frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|^2 + \tau K_m\left\|\nabla e^{n,j}\right\|^2 \\
&\leq \frac{L}{2}\left\|e^{n,j-1}\right\|^2 + \frac{1}{2L}\left\|\theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right)\right\|^2 \\
&+\frac{L}{2}\left\|e^{n,j} - e^{n,j-1}\right\|^2 + \frac{\tau(M+1)^2 L_K^2}{2K_m} \\
&\quad\left\|\left(\theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right)\right)\right\|^2 + \frac{\tau K_m}{2}\left\|\nabla e^{n,j}\right\|^2.
\end{aligned} \tag{14}$$

After some obvious simplifications, the inequality (14) becomes

$$L\|e^{n,j}\|^2 + \tau K_m\|\nabla e^{n,j}\|^2 + \left(\frac{2}{L_\theta} - \frac{1}{L} - \frac{\tau(M+1)^2 L_K^2}{K_m}\right)\left\|\theta\left(\Psi_h^{n,j-1}\right) - \theta\left(\Psi_h^n\right)\right\|^2 \leq L\|e^{n,j-1}\|^2. \tag{15}$$

Finally, by choosing $L > 0$ and the time step τ such that

$$\frac{2}{L_\theta} - \frac{1}{L} - \frac{\tau(M+1)^2 L_K^2}{K_m} \geq 0 \tag{16}$$

and by using the Poincare inequality (recall that $e^{n,j} \in H_0^1(\Omega)$)

$$\|e^{n,j}\| \leq C_\Omega\|\nabla e^{n,j}\|, \tag{17}$$

from Eq. 15 follows the convergence of the scheme (7)

$$\|e^{n,j}\|^2 \leq \frac{L}{L + \frac{K_m \tau}{C_\Omega^2}}\|e^{n,j-1}\|^2. \tag{18}$$

□

We continue with some important remarks concerning the result above and the implications to the convergence of the Newton and modified Picard methods.

Remark 3 In the case of a constant hydraulic conductivity K (or if we refer to Richards' equation after Kirchhoff's transformation and without gravity, see e.g. [25]), the condition for convergence of the $L-$scheme (7) simply becomes

$$L \geq \frac{L_\theta}{2} \tag{19}$$

and there is **no restriction** on the time step size. Furthermore, the assumptions (A2) and (A3) are not necessary in this case.

Remark 4 The rate of convergence (11) depends on K_m, τ and L, but it is independent of the mesh size. Smaller L or larger time steps are resulting in a faster convergence. We also emphasize that larger hydraulic conductivities will imply a faster convergence as well.

Remark 5 In the general case, the optimal choice is $L = L_\theta$ and $\tau = \frac{K_m}{L_\theta(M+1)^2 L_K^2}$. The restriction on the time step size (after choosing $L = L_\theta$) is $\tau \leq \frac{K_m}{L_\theta(M+1)^2 L_K^2}$, which is relatively mild because it does not involve the mesh size or any regularization parameter.

Remark 6 The convergence of the $L-$scheme is global, i.e. independent of the initial choice. Nevertheless, it is obviously beneficial if one starts the iterations with the solution of the last time step.

Remark 7 The convergence of the modified Picard method and of the Newton method is studied in [27] for the case of constant K or for Richards' equation after Kirchhoff's transformation and without gravity. A regularization step is in this case necessary to ensure the convergence. The corresponding convergence condition to Eq. 16 will look like

$$\tau \leq C\epsilon^3 h^d, \tag{20}$$

with ϵ denoting the regularization parameter and h the mesh size, d the spatial dimension and C a constant not depending on the discretization parameters. A similar condition is derived also for the Jäger–Kačur scheme, see [27]. We remark that the condition (20) is much more restrictive than the condition (16). The proofs in [27] are done for mixed finite element based discretizations, but the proof for Galerkin finite elements is similar. The condition (20) is derived by using some inverse estimates and it is in practice quite pessimistic. Nevertheless, we emphasize the fact that the convergence is ensured only when doing a regularization step (reflected by the ϵ in Eq. 20) and this is what one sees in practice as well (see Section 4).

Remark 8 One can extend the convergence proof in [27] for Newton's and modified Picard's methods to the general case of a non-linear K and saturated/unsaturated flow. Under a similar assumption (A2) for the modified Picard and an assumption involving also the Lipschitz continuity of the derivative of K for Newton's method one can show the convergence of the methods. The modified Picard will be linearly convergent, whereas Newton's method will converge quadratically. The condition of convergence will be similar to Eq. 20 for both methods. From the theoretical point of view, only a quantitatively increased robustness for the Picard method comparing with Newton's method should be expected, i.e. when e.g. the mesh size becomes smaller if one of the method fails then also the other one (see Fig. 2, where Newton's methods is not converging and modified Picard converges, but increasing the number of elements leads to divergence for the modified Picard or Picard/Newton methods as well). This is not the case with the $L-$scheme, which is clearly the most robust out of the considered methods, see Section 4.

Remark 9 By using error estimates derived as mentioned in the remark above, one can construct an indicator to predict the convergence of Newton's method. Based on this, one can design an adaptive algorithm for using the $L-$scheme only when necessary. Nevertheless, because the $L-$iterations are so cheap and the resulting linear systems are (much) better conditioned, it seems that the $L-$scheme/Newton is almost that fast as the Newton method. In Example 1 in Section 4 we even experienced that the $L-$scheme/Newton was faster than the Newton method. Therefore, we simply recommend the use of the $L-$scheme/Newton with a fixed number of $L-$iterations (4-5), without any indicator predictions. It the case of convergence failing, one should as a response automatically increase the number of $L-$iterations. We never experienced the need of more than 11 $L-$iterations in order to guarantee the convergence of the $L-$scheme/Newton.

4 Numerical results

In this section, numerical results in two spatial dimensions are presented. The considered linearization schemes: the Newton method, the modified Picard method, Picard/Newton, the $L-$scheme and the $L-$scheme/Newton are comparatively studied. We focus on convergence, computational time and the condition number of the underlying linear systems. We consider two main numerical examples, both based on realistic parameters. The first one was developed by us, the second is a benchmark problem from [31]. Different conditions are created by varying the parameters. The sensitivity of the schemes w.r.t. the mesh size h is particularly studied. All computations were performed on a Schenker XMG notebook with an Intel Core i7-3630GM processor.

The relationships $K(\Psi)$ and $\theta(\Psi)$ for both examples are provided by the van Genuchten–Mualem model, namely

$$\theta(\Psi) = \begin{cases} \theta_R + (\theta_S - \theta_R)\left[\frac{1}{1+(-\alpha\Psi)^n}\right]^{\frac{n-1}{n}}, & \Psi \leq 0, \\ \theta_S, & \Psi > 0, \end{cases}$$
$$K(\Psi) = \begin{cases} K_S \theta(\Psi)^{\frac{1}{2}}\left[1 - \left(1 - \theta(\Psi)^{\frac{n}{n-1}}\right)^{\frac{n-1}{n}}\right]^2, & \Psi \leq 0, \\ K_S, & \Psi > 0, \end{cases} \tag{21}$$

in which θ_S and K_S denote the water content respectively the hydraulic conductivity when the porous medium is fully saturated, θ_R is the residual water content and α and n are model parameters related to the soil properties. We compute the derivatives of K and θ analytically whenever they arise. The evaluation of integrals is executed by applying a quadrature formula accurate for polynomials up to a degree of 4.

Remark 10 The use of automatic differentiation might speed up the Newton method, but the concerns regarding the robustness will remain. This and the fact that most of the codes for solving Richards' equation do not have implemented automatic differentiation, were the reasons to compute the derivatives as mentioned above.

4.1 Example 1

This example deals with injection and extraction in the vadose zone Ω_{vad} located above the groundwater zone Ω_{gw}. The composite flow domain is $\Omega = \Omega_{\text{vad}} \cup \Omega_{\text{gw}}$ defined as $\Omega_{\text{vad}} = (0, 1) \times (-3/4, 0)$ and $\Omega_{\text{gw}} = (0, 1) \times (-1, -3/4]$. We choose the van Genuchten parameters $\alpha = 0.95$, $n = 2.9$, $\theta_S = 0.42$, $\theta_R = 0.026$ and $K_S = 0.12$ in parametrization (21). The choice $n > 2$ implies Lipschitz continuity of both θ and K. Constant Dirichlet conditions $\Psi \equiv \Psi_{\text{vad}}$ on the surface $\Gamma_D = (0, 1) \times \{0\}$ and no-flow Neumann conditions on $\Gamma_N = \partial\Omega \setminus \Gamma_D$ are imposed. The initial pressure height distribution is discontinuous at the transition of the groundwater to the vadose zone and is given by $\Psi^0 \equiv \Psi_{\text{vad}}$ on Ω_{vad} and $\Psi^0 = \Psi^0(z) = -z - 3/4$ on Ω_{gw}. We investigate two initial pressure heights in the vadose zone, $\Psi_{\text{vad}} \in \{-3, -2\}$. In the vadose zone, we select a source term taking both positive and negative values given by $f = f(x, z) = 0.006\cos(4/3\pi z)\sin(2\pi x)$ on Ω_{vad}, whereas we have $f \equiv 0$ in the saturated zone Ω_{gw}.

We examine the numerical solutions after the first time step for $\tau = 1$. A regular mesh is employed, consisting of right-angled triangles whose legs are of length $h = \Delta x = \Delta z$ for $h \in \left\{\frac{1}{10}, \frac{1}{20}, \frac{1}{30}, \frac{1}{40}, \frac{1}{50}, \frac{1}{60}, \frac{1}{70}, \frac{1}{80}\right\}$ (the mesh size is actually $h\sqrt{2}$). The parameters regulating the switch for the mixed methods are taken as $\delta_a = 2$ and $\delta_r = 0$. The computation using the $L-$scheme was carried out with parameter L slightly greater than $L_\theta = \sup_\Psi \theta'(\Psi) = 0.2341$ for the given van Genuchten parametrization, to be specific $L = 0.25$. However, as pointed out in the analysis, when the influence of the non-linear K is not that big (see Remark 3), a constant L bigger than $\frac{L_\theta}{2}$ is enough for the convergence. According to our experience, this is the limit relevant for the practice. Hence, we performed another computation with parameter $L = 0.15$. For the mixed $L-$scheme/Newton we chose $L = 0.15$ as well.

The results for Example 1 are presented in Figs. 1, 2, 3, 4, 5, and 6 and discussed in detail below.

4.1.1 Convergence

In case of higher initial moisture in the vadose zone, that is $\Psi_{\text{vad}} = -2$, convergence was observed for all methods and all investigated meshes. For the choice $\Psi_{\text{vad}} = -3$, Newton's method failed on each mesh, the modified Picard scheme exhibited convergence only for $h \geq \frac{1}{40}$, whereas both parametrizations of the $L-$scheme converged on all meshes. This is consistent with the theoretical findings in Section 3, in particular with Remark 8.

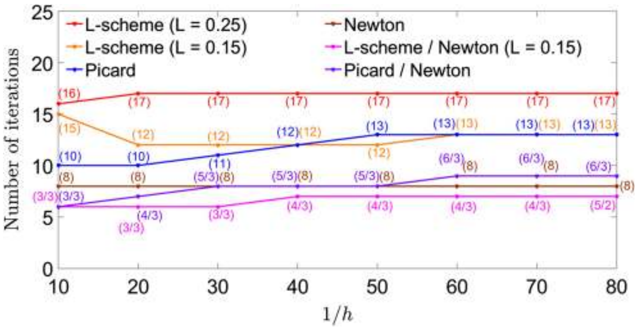

Fig. 1 Numbers of iterations for several mesh sizes, $\Psi_{\text{vad}} = -2$

4.1.2 Number of iterations

The required numbers of iterations are depicted in Figs. 1 and 2. Missing markers indicate that the iteration has not converged. For either value of Ψ_{vad}, the smaller parameter $L = 0.15$ in the $L-$scheme yielded the criterion for convergence to be fulfilled after fewer iterations than $L = 0.25$.

For $\Psi_{\text{vad}} = -2$, the modified Picard scheme required less iterations than the $L-$scheme on coarse meshes, but for $h \leq \frac{1}{40}$, it needed at least as many iterations as the $L-$scheme with $L = 0.15$. Newton's method featured an even smaller number of iterations which was found to be independent of the mesh size in our computation. The number of iterations for the mixed Picard/Newton scheme did not differ significantly from the one for Newton's method, while the mixed $L-$scheme/Newton needed the least iterations on each mesh.

For $\Psi_{\text{vad}} = -3$, the modified Picard scheme had a benefit over the $L-$scheme in view of the number of iterations whenever it converged, although the number of iterations increased considerably as the mesh became finer. The mixed schemes gave the best results with respect to the number of iterations, the application of the mixed Picard/Newton scheme however being limited to coarse meshes.

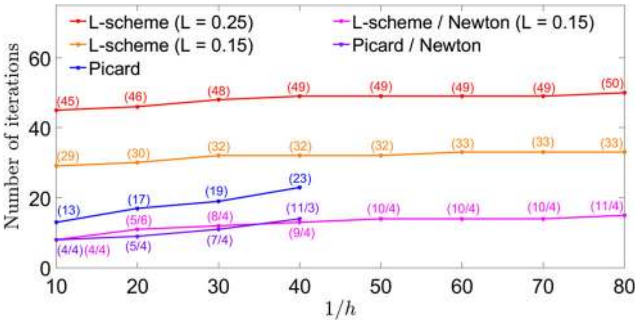

Fig. 2 Numbers of iterations for several mesh sizes, $\Psi_{\text{vad}} = -3$

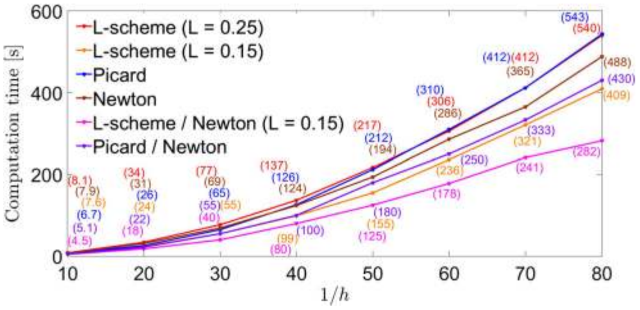

Fig. 3 Computation times for several mesh sizes, $\Psi_{\text{vad}} = -2$

4.1.3 Computation time

Figure 3 shows the computation times for $\Psi_{\text{vad}} = -2$. Although the modified Picard scheme needed less iterations than the $L-$scheme with $L = 0.25$, the differences of computation times were small, since the modified Picard scheme requires the computation of matrices including $\theta'(\Psi)$. For Newton's method, $K'(\Psi)$ has to be calculated in addition. Nevertheless, it converged more rapidly than the modified Picard scheme for $h \leq \frac{1}{40}$. As reported by [20], combination of the modified Picard scheme and Newton's method further improved the performance in terms of computation time. However, both $L-$scheme with $L = 0.15$ and mixed $L-$scheme/Newton exhibited faster convergence than the mixed Picard scheme on dense grids, the mixed $L-$scheme/Newton only taking 65.6% of computation time compared to the mixed Picard/Newton scheme for $h = \frac{1}{80}$.

The computation times for $\Psi_{\text{vad}} = -3$ are presented in Fig. 4. The mixed schemes computed the solution faster than the non-mixed schemes on each mesh, the mixed $L-$scheme/Newton taking roughly half the computation time in comparison to the non-mixed $L-$scheme with $L = 0.15$. In Table 1 we present also computations for several time step sizes and fixed $h = \frac{1}{40}$. One clearly see that with increasing time step size, the mixed scheme performed much better than the Newton or Picard/Newton schemes. For $\tau = 2$ only the $L-$schemes are converging, with mixed L-scheme being the fastest. For the smallest time step, $\tau =$

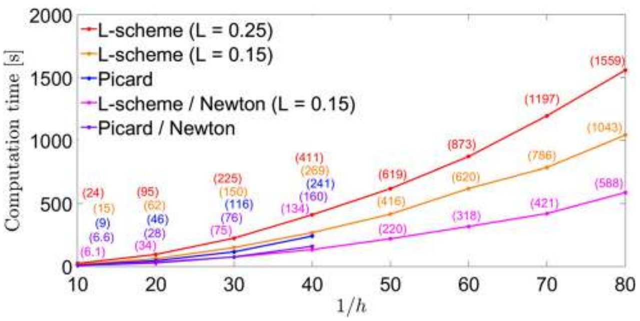

Fig. 4 Computation times for several mesh sizes, $\Psi_{\text{vad}} = -3$

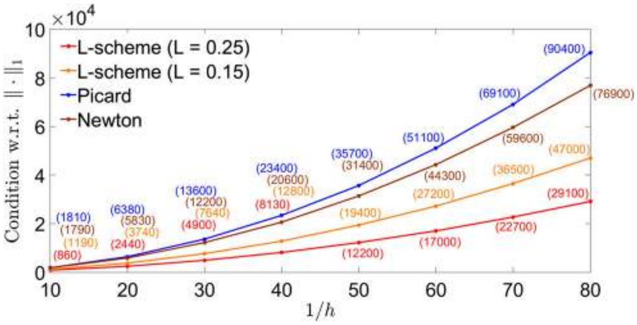

Fig. 5 Condition numbers for several mesh sizes, $\Psi_{\text{vad}} = -2$

0.001 the Picard/Newton scheme was faster than the mixed L-scheme.

4.1.4 Condition numbers

In light of the accuracy of the numerical results, it is interesting to examine the condition numbers of the left-hand side matrices in the system of linear equations for the coefficient vector. Estimations for the condition numbers with respect to $L^1(\Omega)$, denoted by $\|\cdot\|_1$ calculated using the MATLAB function condest() are plotted in Figs. 5 and 6 for the non-mixed methods, averaged over all iterations. They did hardly differ from each other at several iteration steps and condition numbers for the mixed methods corresponded approximately to the ones of the respective non-mixed method in each iteration. For both values of Ψ_{vad}, the $L-$scheme with $L = 0.25$ featured the lowest condition numbers, followed by its counterpart with $L = 0.15$. In case of Newton's method being convergent, it exhibited higher condition numbers than the $L-$scheme. In all computations, the condition numbers in the modified Picard scheme were the highest, furthermore, they increased most rapidly when it came to denser meshes.

All methods required more iterations and computation time when the vadose zone was taken to be dryer initially and the arising matrices were worse-conditioned than for the moister setting.

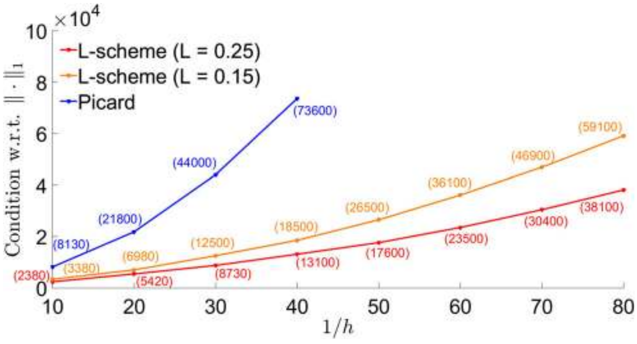

Fig. 6 Condition numbers for several mesh sizes, $\Psi_{\text{vad}} = -3$

4.2 Example 2 (benchmark problem)

In order to compare the linearization methods in the numerical simulation of a recognized benchmark problem, we consider an example used by [13, 17] and [31] amongst others. It describes the recharge of a groundwater reservoir from a drainage trench in two spatial dimensions (Fig. 7). The domain $\Omega \subset \mathbb{R}^2$ represents a vertical section of the subsurface. On the right hand side of Ω, the groundwater table is fixed by a Dirichlet condition for the pressure height for $z \in [0, 1]$. The drainage trench is modelled by a transient Dirichlet condition on the upper boundary for $x \in [0, 1]$. On the remainder of the boundary $\partial\Omega$, no-flow conditions are imposed. Hence, the left boundary can be construed as symmetry axis of the geometry and the lower boundary as transition to an aquitard. Altogether, the geometry is given by

$$
\begin{aligned}
\Omega &= (0, 2) \times (0, 3),\\
\Gamma_{D_1} &= \{(x, z) \in \partial\Omega \mid x \in [0, 1] \wedge z = 3\},\\
\Gamma_{D_2} &= \{(x, z) \in \partial\Omega \mid x = 2 \wedge z \in [0, 1]\},\\
\Gamma_D &= \Gamma_{D_1} \cup \Gamma_{D_2},\\
\Gamma_N &= \partial\Omega \setminus \Gamma_D.
\end{aligned}
$$

The initial and boundary conditions are taken as

$$
\begin{aligned}
\Psi(t, x, z) &= \begin{cases} -2 + 2.2\, t/\Delta t_D, & \text{on } \Gamma_{D_1}, t \leq \Delta t_D,\\ 0.2, & \text{on } \Gamma_{D_1}, t > \Delta t_D,\\ 1 - z, & \text{on } \Gamma_{D_2}, \end{cases}\\
-K(\Psi(t, x, z))(\nabla\Psi(t, x, z) + \mathbf{e_z}) \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_N,\\
\Psi^0(x, z) = 1 - z & \qquad \text{on } \Omega,
\end{aligned}
$$

in which $\mathbf{n}$ denotes the outward pointing normal vector. Initially, a hydrostatic equilibrium is thus assumed. The computations are undertaken for two sets of parameters adopted from [33], characterizing silt loam respectively Beit Netofa clay. For both soil types, the solution is computed over $N = 9$ time levels. The time unit is 1 day and dimensions are given in meters. The van Genuchten parameters employed as well as the parameter Δt_D governing the time evolution of the upper Dirichlet boundary, the time step τ and the simulation end time T are listed in Table 2. We used a regular mesh consisting of 651 nodes. The simulations invoking the $L-$scheme were carried out with $L = \sup_\Psi \theta'(\Psi)$ (referred to as $L-$scheme 1) and with L slightly smaller (referred to as $L-$scheme 2) for both soil types, that is $L = 4.501 \cdot 10^{-2}$ and $L = 3.500 \cdot 10^{-2}$ for the silt loam soil and $L = 7.4546 \cdot 10^{-3}$ and $L = 6.500 \cdot 10^{-3}$ for the clay soil. The mixed methods switched to Newton's method when condition (9) held true for $\delta_a = 0.2$ and $\delta_r = 0$.

Table 1 Computation times for several time step sizes, $\Psi_{\text{vad}} = -3$

Scheme	CPU time [s]	Iterations	Scheme	CPU time [s]	Iterations
$\tau = 2$			$\tau = 0.1$		
$L-$scheme 0.25	368	48	$L-$scheme 0.25	319	41
$L-$scheme 0.15	246	32	$L-$scheme 0.15	219	28
Picard	no convergence		Picard	207	20
Newton	no convergence		Newton	no convergence	
$L-$scheme/Newton	129	13 (9/4)	$L-$scheme/Newton	107	10 (6/4)
Picard/Newton	no convergence		Picard/Newton	123	10 (6/4)
$\tau = 1$			$\tau = 0.01$		
$L-$scheme 0.25	411	49	$L-$scheme 0.25	241	31
$L-$scheme 0.15	269	32	$L-$scheme 0.15	163	20
Picard	241	23	Picard	145	14
Newton	no convergence		Newton	no convergence	
$L-$scheme/Newton	134	14 (11/3)	$L-$scheme/Newton	92	8 (4/4)
Picard/Newton	160	13 (9/4)	Picard/Newton	97	8 (5/3)
$\tau = 0.5$			$\tau = 0.001$		
$L-$scheme 0.25	372	47	$L-$scheme 0.25	1120	145
$L-$scheme 0.15	242	31	$L-$scheme 0.15	743	95
Picard	228	22	Picard	83	8
Newton	no convergence		Newton	105	7
$L-$scheme/Newton	138	13 (8/5)	$L-$scheme/Newton	106	8 (2/6)
Picard/Newton	138	12 (9/3)	Picard/Newton	98	8 (5/3)

All the considered linearization methods converged for both soil types. The pressure profiles computed with mixed $L-$scheme 2/Newton at time T are presented in Fig. 8 and are as expected for this benchmark problem. Table 3 shows the total numbers of iterations, the computation times and the average of the estimated condition numbers of the left-hand side matrices with respect to $\| \cdot \|_1$, in case of mixed methods split up in the two involved schemes. In what follows, the foregoing numerical indicators, i.e. the number of iterations, the computational time and the condition numbers are to be discussed in detail.

4.2.1 Numbers of iterations

As to the non-mixed methods, it is not surprising that more complex methods yielded smaller numbers of iterations, i.e. Newton's method converged after the fewest iterations, followed by the modified Picard scheme. $L-$scheme 2 had the

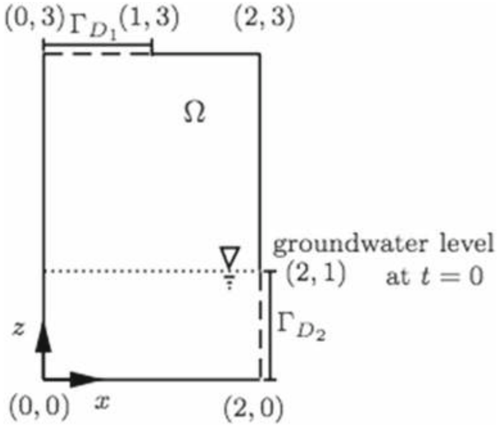

Fig. 7 Geometry for Example 2

Table 2 Simulation parameters for Example 2

	Silt loam	Beit Netofa clay
Van Genuchten parameters:		
θ_S	0.396	0.446
θ_R	0.131	0.0
α	0.423	0.152
n	2.06	1.17
K_S	$4.96 \cdot 10^{-2}$	$8.2 \cdot 10^{-4}$
Time parameters:		
Δt_D	1/16	1
Δt	1/48	1/3
T	3/16	3

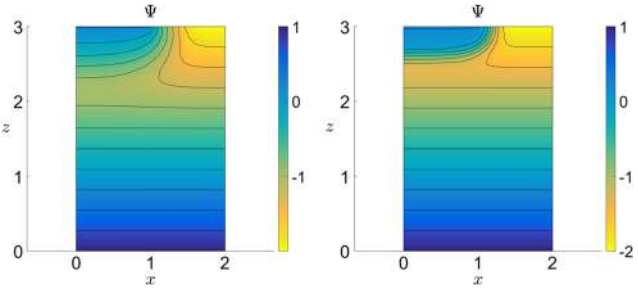

Fig. 8 Pressure profiles after 4.5 [h] for silt loam (left) and 3 [d] for Beit Netofa clay (right)

edge over $L-$scheme 1, but still needed some more iterations than the modified Picard scheme for both soil types. The numbers of iterations of the mixed methods exhibit a salient result: the advantage of the modified Picard scheme over the $L-$scheme with regard to the number of iterations vanished when coupling the schemes to Newton's method and the mixed $L-$scheme 2/Newton required less iterations than the mixed Picard/Newton scheme. This suggests that the $L-$scheme stands out due to a rapid approach towards the solution in the first iteration steps. Among all methods, Newton's method provided convergence after the least number of iterations for both van Genuchten parametrizations.

Table 3 Comparison of the linearization methods for Example 2

	Silt loam	Beit Netofa clay
Total number of iterations:		
$L-$scheme 1	74	74
$L-$scheme 2	65	72
Picard	58	69
Newton	31	48
$L-$scheme 1 / Newton	46 (26/20)	54 (28/26)
$L-$scheme 2 / Newton	40 (22/18)	54 (28/26)
Picard / Newton	43 (25/18)	55 (29/26)
Total computation time [s]:		
$L-$scheme 1	231	237
$L-$scheme 2	210	225
Picard	234	285
Newton	184	289
$L-$scheme 1 / Newton	200	247
$L-$scheme 2 / Newton	180	243
Picard / Newton	213	278
Averaged condition number [10^3]:		
$L-$scheme 1	6.84	51.2
$L-$scheme 2	7.86	56.0
Picard	90.1	321
Newton	90.1	321
$L-$scheme 1 / Newton	6.84/90.1	51.2/321
$L-$scheme 2 / Newton	7.86/90.1	56.0/321
Picard / Newton	9.01/90.1	321/321

4.2.2 Computation times

When it comes to the comparison of computation times, it is striking that the performances of the methods substantially varied between the simulations for silt loam and Beit Netofa clay. While Newton's method featured the shortest computation time among the non-mixed methods in case of silt loam owing to the low number of required iterations, computation in case of the clayey soil took long using Newton's method as compared to the $L-$scheme. In the silt loam simulation, computation times of the $L-$scheme were clearly greater than the ones of Newton's method, but switching to Newton's method vastly improved the computation time so that the $L-$scheme 2/Newton turned out to be the fastest method. In contrast, the computation times for the clay soil demonstrate that in some cases, switching to Newton's method may even be disadvantageous. Although the mixed $L-$scheme/Newton converged in fewer iteration steps than the non-mixed ones, changing to Newton's method provoked a deterioration of the computation time. This might indicate that the $L-$scheme be less susceptible to parametrizations of the hydraulic relationships lacking of regularity than the modified Picard scheme and Newton's method since the hydraulic conductivity for the parametrization of the Beit Netofa clay is not Lipschitz continuous. The modified Picard scheme was found to be the slowest method for the silt loam soil, the computation time for Beit Netofa clay was barely less than the one related to Newton's method.

4.2.3 Condition numbers

In view of the condition numbers of the left-hand side matrices, the $L-$scheme excelled for both soil types: The condition numbers with either value of L were remarkably lower than the ones arising when Newton's method or the modified Picard scheme were employed, to be more specific by a factor of minimum 11 for the silty soil and still by a factor of minimum 5 for the clayey soil. Apparently, incorporation of the derivative of the water content entailed a considerable deterioration of the condition. The virtual equality of the condition numbers for the modified Picard scheme and Newton's method was probably due to the proximity of the solution to a hydrostatic equilibrium which caused the only term distinguishing Newton's method from the modified Picard scheme in Eq. 5 to be small because of $\nabla \Psi_h^n \approx -\mathbf{e_z}$.

5 Conclusions

In this paper we considered linearization methods for the Richards equation. The methods were comparatively studied w.r.t. convergence, computational time and condition number of the resulting linear systems. The analysis was done in connection with Galerkin finite elements, but the schemes can be applied to any other discretization method as well, and similar results are expected. We focused on the Newton method, the modified Picard method, the Picard/Newton and the $L-$scheme. We proposed also a new mixed scheme, the $L-$scheme/Newton which seems to perform best. We conducted a theoretical analysis for the $L-$scheme for Richards' equation, showing that it is robust and linearly convergent. We also discussed the convergence of the modified Picard and Newton methods.

The $L-$scheme is very easy to be implemented, does not involve the computation of any derivatives and the resulting linear systems are much better conditioned as the modified Picard or Newton methods. Although it is only linearly convergent, it seems to be not much slower than the Newton (or Picard/Newton) method, and in some cases even faster. The $L-$scheme is the only robust one, a result which can be shown theoretically and it is supported by the numerical findings. Only a relatively mild constraint on the time step length is required. Furthermore, when the hydraulic conductivity K is a constant, there is no restriction in the time step size. In this case the only condition necessary for the global convergence of the $L-$method is $L \geq \frac{L_\theta}{2}$.

We proposed a new mixed scheme, the $L-$scheme/Newton which is more robust than Newton but still quadratically convergent. This new mixed method performed best from all the considered methods with respect to computational time. Even in cases when Newton converges, the $L-$scheme/Newton seems to be worth, being faster in the examples considered.

The present study is based on two illustrative numerical examples, with realistic parameters. The examples are two dimensional. One of the examples is a known benchmark problem. The numerical findings are sustaining the theoretical analysis.

Acknowledgments Thank the DFG-NRC NUPUS for support and our colleagues J.M. Nordbotten, I.S. Pop and K. Kumar for very helpful discussions. Part of this work has been finished during the visit of F.A. Radu at TU Eindhoven, supported by the Meltzer foundation, University of Bergen and the Dutch Research Council (NOW) Visitors Grant 040.11.499.

References

1. Alt, H.W., Luckhaus, S.: Quasilinear elliptic-parabolic differential equations. Math. Z. **183**, 311–341 (1983)
2. Arbogast, T.: An error analysis for Galerkin approximations to an equation of mixed elliptic-parabolic type, Technical Report TR90-33, Department of Computational and Applied Mathematics, Rice University, Houston, TX (1990)
3. Arbogast, T., Obeyesekere, M.M., Wheeler, M.F.: Numerical methods for the simulation of flow in root-soil systems. SIAM J. Numer. Anal. **30**, 1677–1702 (1993)
4. Arbogast, T., Wheeler, M.F., Zhang, N.Y.: A non-linear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. SIAM J. Numer. Anal. **33**, 1669–1687 (1996)
5. Bause, M., Knabner, P.: Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods. Adv. Water Resour. **27**, 565–581 (2004)
6. Bear, J., Bachmat, Y.: Introduction to modelling of transport phenomena in porous media. Kluwer Academic (1991)
7. Bergamashi, N., Putti, M.: Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. Int. J. Numer. Methods Eng. **45**, 1025–1046 (1999)
8. Celia, M., Bouloutas, E., Zarba, R.: A general mass-conservative numerical solution for the unsaturated flow equation. Water Resour. Res. **26**, 1483–1496 (1990)
9. Ebmeyer, C.: Error estimates for a class of degenerate parabolic equations. SIAM J. Numer. Anal. **35**, 1095–1112 (1998)
10. Eymard, R., Gutnic, M., Hilhorst, D.: The finite volume method for Richards equation. Comput. Geosci. **3**, 259–294 (1999)
11. Eymard, R., Hilhorst, D., Vohralík, M.: A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. Numer. Math. **105**, 73–131 (2006)
12. Forsyth, P.A., Wu, Y.S., Pruess, K.: Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. Adv. Water Resour. **18**, 25–38 (1995)
13. Haverkamp, R., Vauclin, M., Touma, J., Wierenga, P.J., Vachaud, G.: A comparison of numerical simulation models for one-dimensional infiltration. Soil Sci. Soc. Am. J. **41**, 285–294 (1977)
14. Huang, K., Mohanty, B.P., van Genuchten, M.T.: A new convergence criterion for the modified Picard iteration method to solve the variably saturated flow equation. J. Hydrol. **178**, 69–91 (1996)
15. Kačur, J.: Solution to strongly non-linear parabolic problems by a linear approximation scheme. IMAJNA **19**, 119–145 (1995)
16. Klausen, R.A., Radu, F.A., Eigestad, G.T.: Convergence of MPFA on triangulations and for Richards' equation. Int. J. Numer. Meth. Fluids **58**, 1327–1351 (2008)
17. Knabner, P.: Finite element simulation of saturated-unsaturated flow through porous media. LSSC **7**, 83–93 (1987)
18. Knabner, P., Angermann, L.: Numerical methods for elliptic and parabolic partial differential equations. Springer (2003)
19. Kräutle, S.: The semismooth Newton method for multicomponent reactive transport with minerals. Adv. Water Resour. **34**, 137–151 (2011)
20. Lehmann, F., Ackerer, P.: Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. Transp. Porous Media **31**, 275–292 (1998)
21. Lott, P.A., Walker, H.F., Woodward, C.S., Yang, U.M.: An accelerated Picard method for non-linear systems related to variably saturated flow. Adv. Water Resour. **38**, 92–101 (2012)
22. Nochetto, R.H., Verdi, C.: Approximation of degenerate parabolic problems using numerical integration. SIAM J. Numer. Anal. **25**, 784–814 (1988)

23. Park, E.J.: Mixed finite elements for non-linear second-order elliptic problems. SIAM J. Numer. Anal. **32**, 865–885 (1995)
24. Pop, I.S.: Error estimates for a time discretization method for the Richards' equation. Comput. Geosci. **6**, 141–160 (2002)
25. Pop, I.S., Radu, F.A., Knabner, P.: Mixed finite elements for the Richards' equations: linearization procedure. J. Comput. Appl. Math. **168**, 365–373 (1999)
26. Radu, F.A., Pop, I.S., Knabner, P.: Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. SIAM J. Numer. Anal. **42**, 1452–1478 (2004)
27. Radu, F.A., Pop, I.S., Knabner, P.: On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation, Numerical Mathematics and Advanced Applications, pp. 1194–1200. Springer (2006)
28. Radu, F.A., Pop, I.S., Knabner, P.: Error estimates for a mixed finite element discretization of some degenerate parabolic equations. Numer. Math. **109**, 285–311 (2008)
29. Radu, F.A., Wang, W.: Error estimates for a mixed finite element discretization of some degenerate parabolic equations. Nonlinear Anal. Real World Appl. **15**, 266–275 (2014)
30. Radu, F.A., Nordbotten, J.M., Pop, I.S., Kumar, K.: A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media. J. Comput. Appl. Math. **289**, 134–141 (2015)
31. Schneid, E.: Hybrid-Gemischte Finite-Elemente-Diskretisierung der Richards-Gleichung. PhD Thesis (in german) University of Erlangen-Nürnberg, Germany (2004)
32. Slodička, M.: A robust and efficient linearization scheme for doubly non-linear and degenerate parabolic problems arising in flow in porous media. SIAM J. Sci. Comput. **23**, 1593–1614 (2002)
33. van Genuchten, M.Th.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J., 44 (1980)
34. Woodward, C., Dawson, C.: Analysis of expanded mixed finite element methods for a non-linear parabolic equation modeling flow into variably saturated porous media. SIAM J. Numer. Anal. **37**, 701–724 (2000)
35. Yotov, I.: A mixed finite element discretization on non–matching multiblock grids for a degenerate parabolic equation arizing in porous media flow. East–West J. Numer. Math. **5**, 211–230 (1997)
36. Yong, W.A., Pop, I.S.: A numerical approach to porous medium equations, Preprint 95-50 (SFB 359), IWR. University of Heidelberg (1996)