Computers and Mathematics with Applications 80 (2020) 405–416

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Auxiliary space preconditioning for mixed finite element discretizations of Richards' equation☆

Juan Batista [b], Xiaozhe Hu [a], Ludmil T. Zikatanov [b,c,*]

[a] Department of Mathematics, Tufts University, Medford, MA 02155, USA
[b] Department of Mathematics, The Pennsylvania State University, University Park, PA 16801, USA
[c] Institute for Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

ARTICLE INFO

Article history:
Available online 24 September 2019

Keywords:
Auxiliary space method
Mixed finite element methods
Richards' equation
Edge average finite element scheme

ABSTRACT

We propose an auxiliary space method for the solution of the indefinite problem arising from mixed method finite element discretizations of scalar elliptic problems. The proposed technique uses conforming elements as an auxiliary space and utilizes special interpolation operators for the transfer of residuals and corrections between the spaces. We show that the corresponding method provides optimal solver for the indefinite problem by only solving symmetric and positive definite auxiliary problems. We apply this preconditioner to the mixed form discretization of Richards' equation linearized with the L-scheme. We provide numerical tests validating the theoretical estimates.

1. Introduction

This paper is on uniform preconditioning of the linear systems resulting from the lowest order mixed finite element discretizations of the Darcy equation. This is motivated as such systems need to be solved repeatedly during simulations involving numerical solution of nonlinear equations, and in particular, Richards' equation. Richards' equation [1] is used in many applications to model the flow of water through an unsaturated porous medium. L. Richards [1] derived it in an attempt to develop a model of groundwater dynamics that incorporates nonlinear capillary effects without having to model the air as an unknown itself; it can be seen as a simplification of two phase flow in which the air phase is assumed to have constant pressure, and the effects of air saturation on the flow of the water phase are reflected by both water content and hydraulic conductivity as functions of water pressure head [2]. Numerical simulations based on this equation are used in agricultural, geochemical, and nuclear-waste-disposal applications, among others [3]. It is a natural nonlinear extension of Darcy's Law, with the nonlinearity being present in the hydraulic conductivity of the solid matrix. In many cases, the linear relation Darcy flow imposes is an over-simplification, as the material properties can depend on the flow in a nonlinear fashion.

Given a bounded region $\Omega \subset \mathbb{R}^d$, $d = 2,\ 3$, filled with a porous material, such as soil, and partially saturated with water, we denote the pressure head of the fluid as our primary unknown $\Psi(\boldsymbol{x}, t)$, the water content at a particular point as $\theta(\boldsymbol{x}, t, \Psi)$ (scaled so that $\theta = 0$ implies relatively dry soil and $\theta = 1$ implies fully saturated soil), the hydraulic

☆ The work of the first author was supported in part by NSF grant DMS-1720114. The work of the second author was supported in part by NSF grant DMS-1620063. The work of the third author was supported in part by NSF grants DMS-1720114 and DMS-1819157.

* Corresponding author at: Department of Mathematics, The Pennsylvania State University, University Park, PA 16801, USA.
E-mail addresses: jxb1131@psu.edu (J. Batista), Xiaozhe.Hu@tufts.edu (X. Hu), ludmil@psu.edu (L.T. Zikatanov).
URLs: http://math.tufts.edu/faculty/xhu/ (X. Hu), http://math.psu.edu/ltz (L.T. Zikatanov).

https://doi.org/10.1016/j.camwa.2019.09.011
0898-1221/

conductivity of the porous medium as $K(\boldsymbol{x}, t, \theta)$, and $f(\boldsymbol{x}, t)$ as a source term. The formulation of the Richards' equation we use to describe the pressure head at any point in Ω with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ is the following:

$$\begin{aligned}
&\partial_t \theta(\Psi) - \operatorname{div}(K(\theta(\Psi))\nabla(\Psi + z)) = f, && (\boldsymbol{x}, t) \in \Omega \times [0, T), \\
&\Psi(\cdot, t) = g_D(\boldsymbol{x}, t), && (\boldsymbol{x}, t) \in \Gamma_D \times [0, T), \\
&K(\theta(\Psi))\nabla(\Psi + z) \cdot \nu = g_N(\boldsymbol{x}, t), && (\boldsymbol{x}, t) \in \Gamma_N \times [0, T), \\
&\theta(\Psi(\boldsymbol{x}, 0)) = \theta_0, && (\boldsymbol{x}, t) \in \Omega \times \{0\}.
\end{aligned} \tag{1}$$

This is a nonlinear elliptic–parabolic equation with potential degeneracies in both the parabolic and elliptic terms. As such, its well posedness is a rather involved matter. We point the reader to the classical results in [4], and [5] for more details on existence and uniqueness of solutions to such degenerate elliptic–parabolic equations. We also note that in some very special cases, closed form analytic solutions can be found [6].

The numerical solution of the Richards' equation involves two main steps: linearization and discretization. Several linearization techniques can be used, and the ones we focused on are combined with the implicit Euler discretization in time to yield a sequence of discrete problems whose solutions approximate the pressure head. Well known linearizations used in practice are the Newton–Raphson and Picard methods. Additionally, there are some specialized linearization techniques such as the modified Picard method [7], and the L-scheme [8]. The former method treats the conductivity as in a Picard iteration and uses a lumped-mass matrix for the discretization of the time derivative of the water content. The latter further simplifies the linearization by replacing the time derivative term with an upper bound. The convergence rate of the L-scheme is slower than Newton–Raphson, but it is much more robust in the choice of the initial guess, and much cheaper to compute, as the Jacobian is replaced with an upper bound for all iterations. For a recent review of these and other linearization schemes for Richards' equation we refer to [8].

As discretization strategies, finite elements have been suggested in several works (see [9–12], and [8]) and we focus on mixed finite elements as our discretization. It is well known that different choices of physical parameters have notable effect on both the accuracy of discretization and the efficiency of the linear solvers involved [13–15]. This is especially true in the case of saturated–unsaturated media: near saturation fronts, the PDE switches type and the coefficients $K(\theta(\Psi))$ and $\theta(\Psi)$ may develop steep gradients or even become singular for certain materials [14,15].

Many different approaches have been taken to accommodate such scenarios, including interpolation techniques for the parameters to make numerical computations more amenable [11,13]. Alternatively, there have been many advances in making the numerical schemes more robust [7,8,10].

It is also clear that the nonlinear iterations and the transient character of the problem require repeated solution of ill-conditioned, large scale linear systems. Their optimal solution demands preconditioners which are robust with respect to physical and discretization parameters and the changing type of the Richards' equation across the computational domain.

In this paper we propose an auxiliary space preconditioner for the lowest order mixed finite-element discretization and show that it is uniform. Auxiliary space preconditioners have been a popular trend in the design of fast solvers for Darcy flow and other mixed methods. We refer to [16,17] for classical results on such methods. In our presentation, we follow the framework outlined in [18] and the techniques proposed in this work have been used to solve different mixed methods, including the Darcy flow [19], flow in fractured porous media [20], time-dependent Maxwell equations [21], and Biot's equations [22].

As auxiliary space, we utilize the conforming (Lagrange) linear finite elements on a simplicial mesh. An application of the preconditioner requires a relaxation step followed by solving systems corresponding to discretizations of the same problem with piece-wise linear continuous elements, which involves solving a system with significantly fewer degrees of freedom.

The rest of the paper is organized as follows. In Section 2, we present the mixed form of the Richards' equation and its discretization. We then linearize it and develop the Schur iteration that we use to precondition the full mixed system on each iteration. In Section 3, we introduce our auxiliary space preconditioner for the approximate Schur complement solve in the Schur iteration, and prove its uniformity. Section 4 closes by showing some test results that numerically verify our preconditioner works as intended for various choices of θ, K, and initial and boundary data.

2. Discretization and linearization

In [10,12,23,24], the mixed finite-element method has been used to solve the Richards' equation (1). The mixed form of (1) is:

$$\begin{aligned}
&\partial_t \theta(\Psi) - \operatorname{div} q = f, && (\boldsymbol{x}, t) \in \Omega \times [0, T), \\
&K^{-1}(\theta(\Psi))q - \nabla(\Psi + z) = 0, && (\boldsymbol{x}, t) \in \Omega \times [0, T), \\
&\Psi(\cdot, t) = g_D(\boldsymbol{x}, t), && (\boldsymbol{x}, t) \in \Gamma_D \times [0, T), \\
&q \cdot \nu = g_N(\boldsymbol{x}, t), && (\boldsymbol{x}, t) \in \Gamma_N \times [0, T), \\
&\theta(\Psi(\boldsymbol{x}, 0)) = \theta_0, && (\boldsymbol{x}, t) \in \Omega \times \{0\}.
\end{aligned} \tag{2}$$

The implicit Euler method with time step τ gives the following sequence of problems

$$\theta(\Psi^n) - \tau \operatorname{div} q^n = \tau f^n + \theta(\Psi^{n-1}),$$
$$K^{-1}(\theta(\Psi^n))q^n - \nabla(\Psi^n + z) = 0,$$

where $(\cdot)^n$ is the value of the argument at timestep $t_n := n\tau$.

By introducing $H(\operatorname{div}, \Omega) := \{r \in L^2(\Omega) | \operatorname{div} r \in L^2(\Omega)\}$, we consider the following variational formulation of this problem: Find $(\Psi^n, q^n) \in L^2(\Omega) \times H(\operatorname{div}, \Omega)$ such that for every $(v, r) \in L^2(\Omega) \times H(\operatorname{div}, \Omega)$,

$$\begin{aligned}(\theta(\Psi^n), v) - \tau(\operatorname{div} q^n, v) &= \tau(f^n, v) + (\theta(\Psi^{n-1}), v),\\ (K^{-1}(\Psi^n)q^n, r) + (\Psi^n + z, \operatorname{div} r) &= (g_D, r)_{\partial\Omega},\end{aligned} \tag{3}$$

where $(\cdot, \cdot)$ denotes the usual L_2 inner product, and $(\cdot, \cdot)_{\partial\Omega}$ is the usual boundary inner product. For the sake of simplicity, we focus on solving (3) with boundary conditions $\Gamma_D = \partial\Omega$, $g_D = 0$ in space, and assuming for simplicity that the map $\theta(\Psi)$ is invertible at time $t = 0$, so that some initial pressure head Ψ_0 can be found in d dimensions ($d = 2$ or 3).

Our choice of discretization for the variational problem is the lowest order mixed finite elements, namely, the standard Raviart–Thomas-Nédélec elements [25–27]. The primary purpose of choosing this discretization is the desirable property of conservation of mass, both globally and locally [25]. However, the drawback is the introduction of many more degrees of freedom in the form of the local fluxes, q^n.

We assume for simplicity that Ω is a Lipschitz polyhedron (polygon in 2D) formed by a union of shape-regular simplices $T \in \mathcal{T}_h$: $\Omega = \cup_{T\in\mathcal{T}_h} T$. We then choose subspaces $S_h \subset L^2(\Omega)$ and $Q_h \subset H(\operatorname{div}, \Omega)$, spanned by piecewise constant basis functions for Ψ^n and first order Raviart–Thomas basis functions for q^n, respectively. Thus, our fully discrete problem for each timestep is finding a solution $(\Psi_h^n, q_h^n) \in S_h \times Q_h$ such that for every $(\varphi, \eta) \in S_h \times Q_h$,

$$\begin{aligned}(\theta(\Psi_h^n), \varphi) - \tau(\operatorname{div} q_h^n, \varphi) &= \tau(f^n, \varphi) + (\theta(\Psi_h^{n-1}), \varphi),\\ (K^{-1}(\Psi_h^n)q^n, \eta) + (\Psi^n + z, \operatorname{div} \eta) &= (g_D, r)_{\partial\Omega}.\end{aligned} \tag{4}$$

To linearize (4), we linearize the primal form of (1) using the L-scheme mentioned earlier. Each linear problem then is written in mixed form. The corresponding mixed form L-scheme is as follows: Find $(\Psi_\epsilon, q_\epsilon) \in S_h \times Q_h$ so that for every $(\varphi, \eta) \in S_h \times Q_h$,

$$(L_\theta \Psi_\epsilon, \varphi) - \tau(\operatorname{div} q_\epsilon, \varphi) = \tau(f, \varphi) + (\theta^{n-1}, \varphi) - ((\theta^{n,j-1})\Psi_h^{n,j-1}, \varphi) + \tau(\operatorname{div} q_h^{n,j-1}, \varphi), \tag{5}$$
$$(K^{-1}(\Psi_h^{n,j-1})q_\epsilon, \eta) + (\Psi_\epsilon, \operatorname{div} \eta) = -(K^{-1}(\Psi_h^{n,j-1})q_h^{n,j-1}, \eta) - (\Psi_h^{n,j-1}, \operatorname{div} \eta). \tag{6}$$

We then update,

$$\Psi_h^{n,j} = \Psi_h^{n,j-1} + \Psi_\epsilon, \quad q_h^{n,j} = q_h^{n,j-1} + q_\epsilon.$$

Here, $(\cdot)^{n,j}$ is the jth iterate of the iterative solution method being implemented to solve at timestep n. In [8] it is shown that the choice of $L_\theta = \sup_\psi |\theta'(\Psi)|$ gives a convergent and robust nonlinear iteration. It is also clear from the above mixed formulation that on every time step $n\tau$ we need to solve several indefinite problems corresponding to the discretization of the mixed form of the L-scheme.

2.1. Preconditioner for the mixed system

For each time step, the resulting fully discrete linear system that needs to be solved on every non-linear iteration has the following form:

$$\begin{bmatrix} A_{qq} & B_{\text{div}}^T \\ B_{\text{div}} & -D_\theta \end{bmatrix} \begin{pmatrix} q_\epsilon \\ \Psi_\epsilon \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ \tilde{g} \end{pmatrix}, \tag{7}$$

where $A_{qq} \leftarrow \tau(K^{-1}(\Psi_h^{n,j-1})q, \eta)$, $B_{\text{div}} \leftarrow \tau(\operatorname{div} q, \varphi)$, and $D_\theta \leftarrow (L_\theta \Psi, \varphi)$.

To solve this system, we use an inexact Uzawa-type iterative solver which we refer to as the Schur iteration, which utilizes an approximation $\widetilde{S}_R$ of the pressure Schur complement of the system, $S_R := -(D_\theta + B_{\text{div}} A_{qq}^{-1} B_{\text{div}}^T)$. Such iteration for the indefinite problem is given in Algorithm 2.1 . We note that this method only requires solutions of systems with $\widetilde{S}_R$ and A_{qq}.

Algorithm 2.1 (*Schur Iteration*)**.** Given initial guess $\left(q_\epsilon^0, \Psi_\epsilon^0\right)$, we use the following recurrence relation to define $\left(q_\epsilon^{k+1}, \Psi_\epsilon^{k+1}\right)$ in terms of the kth iterates:

1. Solve $A_{qq} u = \widetilde{f} - B_{\text{div}}^T \Psi_\epsilon^k$;
2. Solve,

$$\widetilde{S}_R v = (\widetilde{g} + D_\theta \Psi_\epsilon^k - B_{\text{div}} u); \tag{8}$$

3. Update Ψ_ϵ as $\Psi_\epsilon^{k+1} = \Psi_\epsilon^k + \omega_R v$ and then solve again with A_{qq}, namely, solve

$$A_{qq} w = -B_{\text{div}}^T v;$$

4. Update q_ϵ as $q_\epsilon^{k+1} = u + w$.

In order for our iterative scheme to be uniformly convergent with respect to mesh size and other physical parameters, as a minimum, we need $\widetilde{S}_R$ to be spectrally equivalent to S_R. One way of doing this, on shape regular meshes, is to replace A_{qq} with its diagonal, and define

$$\widetilde{S}_R := -\left(D_\theta + B_{\text{div}} \text{diag}(A_{qq})^{-1} B_{\text{div}}^T\right). \tag{9}$$

We now show that Algorithm 2.1 converges for certain choice of ω_R; this proof is similar in approach to many standard approaches in the literature on iterative methods (see, e.g. Young [28]).

Lemma 2.2. *For sufficiently small ω_R, the iterates Ψ_ϵ^k and q_ϵ^k obtained by Algorithm 2.1 converge to the solution of (7).*

Proof. We first consider $\Psi_\epsilon^{k+1} - \Psi_\epsilon^k$. Note that if $\begin{pmatrix} q_\epsilon \\ \Psi_\epsilon \end{pmatrix}$ is a solution to (7), then $A_{qq} q_\epsilon + B_{\text{div}}^T \Psi_\epsilon = \tilde{f}$ and $B_{\text{div}} q_\epsilon - D_\theta \Psi_\epsilon = \tilde{g}$. We have,

$$\begin{aligned}
\Psi_\epsilon^{k+1} - \Psi_\epsilon &= \Psi_\epsilon^k + v - \Psi_\epsilon \\
&= \Psi_\epsilon^k + \omega_R \widetilde{S}_R^{-1} \left[\underbrace{B_{\text{div}} q_\epsilon - D_\theta \Psi_\epsilon}_{\tilde{g}} + D_\theta \Psi_\epsilon^k - B_{\text{div}} A_{qq}^{-1} (\tilde{f} - B_{\text{div}}^T \Psi_\epsilon^k) \right] - \Psi_\epsilon \\
&= \Psi_\epsilon^k - \Psi_\epsilon + \omega_R \widetilde{S}_R^{-1} \left[B_{\text{div}} q_\epsilon + D_\theta (\Psi_\epsilon^k - \Psi_\epsilon) - B_{\text{div}} A_{qq}^{-1} (\underbrace{A_{qq} q_\epsilon + B_{\text{div}}^T \Psi_\epsilon}_{\tilde{f}} - B_{\text{div}}^T \Psi_\epsilon^k) \right] \\
&= \Psi_\epsilon^k - \Psi_\epsilon + \omega_R \widetilde{S}_R^{-1} \left[D_\theta (\Psi_\epsilon^k - \Psi_\epsilon) + B_{\text{div}} A_{qq}^{-1} B_{\text{div}}^T (\Psi_\epsilon^k - \Psi_\epsilon) \right] \\
&= (I - \omega_R \widetilde{S}_R^{-1} S_R)(\Psi_\epsilon^k - \Psi_\epsilon).
\end{aligned}$$

Thus, in order for $\Psi_\epsilon^k \to \Psi_\epsilon$ as $k \to \infty$, it is necessary and sufficient $\rho(I - \omega_R \widetilde{S}_R^{-1} S_R) < 1$. Obviously, this inequality is satisfied if ω_R is sufficiently small, namely,

$$0 < \omega_R < \frac{2}{\rho(\widetilde{S}_R^{-1} S_R)}. \tag{10}$$

On the other hand, for $q_\epsilon^{k+1} - q_\epsilon$ we have

$$\begin{aligned}
q_\epsilon^{k+1} &= u + w = A_{qq}^{-1} (\tilde{f} - B_{\text{div}}^T \Psi_\epsilon^k) - A_{qq}^{-1} B_{\text{div}}^T v \\
&= A_{qq}^{-1} (A_{qq} q_\epsilon + B_{\text{div}}^T \Psi_\epsilon - B_{\text{div}}^T \Psi_\epsilon^k) - A_{qq}^{-1} B_{\text{div}}^T v \\
&= q_\epsilon + A_{qq}^{-1} B_{\text{div}}^T (\Psi_\epsilon - \Psi_\epsilon^k - v) = q_\epsilon - A_{qq}^{-1} B_{\text{div}}^T (\Psi_\epsilon^{k+1} - \Psi_\epsilon).
\end{aligned}$$

As a result from this relation we get

$$q_\epsilon - q_\epsilon^{k+1} = A_{qq}^{-1} B_{\text{div}}^T (\Psi_\epsilon^{k+1} - \Psi_\epsilon).$$

Thus, since $\Psi_\epsilon^k \to \Psi_\epsilon$ and $A_{qq}^{-1} B_{\text{div}}^T$ is a bounded operator with our choice of finite-element spaces, $q_\epsilon^k \to q_\epsilon$ as well. □

Clearly, (10) holds, provided that ω_R is sufficiently small and A_{qq} is spectrally equivalent to a diagonal matrix, such as $\text{diag}(A_{qq})$. Results along these lines are found in [29] and [30]. These works provide two simple techniques for estimating the parameter ω_R elementwise. Furthermore, we would like to point out that while using the diagonal of A_{qq} in (9) is sufficient to show spectral equivalence between $\widetilde{S}_R$ and S_R, there are also techniques that provide more accurate approximation. We refer to [30] for constructing diagonal (mass lumping) approximations to A_{qq} which also gives a finite element solution with same accuracy as the mixed FE discretization without mass lumping.

To conclude this section, let us point out that in our formulation $\widetilde{S}_R$ and S_R are, in fact, negative definite. However below often we discuss preconditioning of positive definite $(-\widetilde{S}_R)$ and $(-S_R)$. To avoid proliferation of sign changes from now on we refer to $(-S_R)$ and $(-\widetilde{S}_R)$ as Schur complement and approximate Schur complement, respectively.

3. Auxiliary space preconditioner

To motivate the auxiliary space preconditioner for (8), we observe that the bilinear form generated by $\widetilde{S}_R$ is a weighted "graph" Laplacian for Ψ corresponding to taking differences of the values of Ψ on neighboring elements across their

common face. The weights are given by integrating K at each element. Similar formulation for continuous piece-wise linear Lagrange elements, even for discretized convection–diffusion equations, are derived in [31] (see also [32] and discussion there for M-matrix relatives of discretized elliptic operators). The idea we advocate here is to use the Lagrange elements as auxiliary space for the piece-wise constants ($P0$) space using the tools developed in [16,17], and [18] to analyze such an auxiliary space preconditioner. Overall, our aim is to combine the nodal discretization (continuous FE) with a simple smoother and properly constructed interpolation map to precondition the $P0$ system. This preconditioner offers two distinct advantages. First, we show that it is uniform with respect to the mesh size and the time step τ. Second, the number of vertices of any lattice $\mathcal{T}_h$ with no hanging nodes is smaller than the number of simplices forming the triangulation. This can be seen by considering the case of a uniform lattice of size h of the unit square $[0,1]^d$ with $N = h^{-1}$ being an integer. The number of simplices in the mesh is $d!N^d$, while the number of vertices is $(N+1)^d$. Thus, the number of degrees of freedom of our auxiliary space is on the order of $\frac{1}{d!}$ the number of degrees of freedom of the $\widetilde{S}_R$ system that must be solved on each step of our Schur iteration 2.1.

Next, to define the auxiliary space preconditioner, we follow the notation in [18] and introduce the following components.

- The fictitious space $\bar{V} = S_h \times V_h$, with $V_h \subset H_0^1(\Omega)$ being the space of piece-wise linear and continuous functions with zero trace on the boundary. For typical elements $p \in S_h$ and $v \in V_h$ we have

$$p = \sum_{T\in\mathcal{T}_h} p_T \chi_T, \quad v = \sum_{j=1}^{N} v_j \phi_j, \tag{11}$$

  with $\{\chi_T\}_{T\in\mathcal{T}_h}$ being the set of characteristic functions for each simplex in the triangulation $\mathcal{T}_h$, which form a basis for S_h, and $\{\phi_j\}_{j=1}^N$ being the standard piecewise-linear Lagrange tent functions defined to be 1 on one of the N vertices forming the triangulation, and 0 at every other vertex, which form a basis for V_h.
- The map between the auxiliary space and S_h, $\Pi = \begin{pmatrix} I & \Pi_J \end{pmatrix}$, with $\Pi_J : V_h \to S_h$. The action of Π_J amounts to taking the average per element T of the values of v on its vertices $j \in T$, namely, given $v \in V_h$, $v = \sum_{i=1}^N v_i\phi_i$, we define

$$[\Pi_J(v)]_T := p_T = \frac{1}{d+1}\sum_{j\in T} v_j, \quad \Pi_J(v) = \sum_T p_T\chi_T, \quad \text{for all} \quad v \in V_h.$$

- Our smoother is the Jacobi smoother, which just uses the diagonal of $\widetilde{S}_R$ (which may be scaled if needed);

$$D_{\widetilde{S}} = \operatorname{diag}(\widetilde{S}_R) : S_h \to S_h'.$$

The preconditioner B is then

$$B = \Pi \begin{pmatrix} D_{\widetilde{S}} & 0 \\ 0 & A_E \end{pmatrix}^{-1} \Pi^* = D_{\widetilde{S}}^{-1} + \Pi_J A_E^{-1} \Pi_J^*, \tag{12}$$

where A_E denotes the linearization and discretization of (1) on the auxiliary space V_h.

Given a right hand side r, an algorithm for the action of B is as follows.

Algorithm 3.1 (*Auxiliary Space Preconditioner B: $z \leftarrow Br$*).

1. Transfer the right hand side r to the auxiliary space V_h: $r_{V_h} \leftarrow \Pi r$,
2. Solve the auxiliary problem on V_h: $e_{V_h} \leftarrow A_E^{-1} r_{V_h}$,
3. Transfer the correction e_{V_h} back to S_h: $z \leftarrow \Pi^* e_{V_h}$,
4. Smooth the correction with Jacobi iteration: $z \leftarrow z + D_{\widetilde{S}}^{-1}(r - \widetilde{S}_R z)$.

To discretize the problem on the auxiliary space, we wish to use a discretization on V_h that formulates the diffusion operator on vertices as differences along edges of elements, as that facilitates the analysis we use to prove the uniformity of our preconditioner. Our choice is to discretize using edge averaged finite elements, or EAFE [31] (see also [33] for such a discretization in 2D). While standard Lagrange elements can also be used, our choice is motivated by the fact that such discretizations are robust with respect to jumps in the coefficient K, as they use harmonic averages of the coefficient K along edges. The second reason is that such discretizations can be used for higher order, e.g., Newton–Raphson linearizations of the Richards' equation, which result in discrete convection–diffusion problems and EAFE provides accurate discretization in this case too. Moreover, for symmetric linearizations like the L-scheme, as shown in [31] and [33] this discretization guarantees monotone discretizations under reasonable conditions on the mesh. For details on such geometric conditions we refer the reader also to [34], and [35].

In what follows we need to use spectrally equivalent forms of A_E and $\widetilde{S}_R$ which are as follows.

$$(A_E v, w) = \sum_{e\in\mathcal{T}_h} \omega_e \delta_e v \delta_e w \quad \text{and} \quad (\widetilde{S}_R p, s) = \sum_{f\in\mathcal{T}_h} d_f (p_{T^+} - p_{T^-})(s_{T^+} - s_{T^-}). \tag{13}$$

where for an edge e connecting vertices i and j (where we assume that $i > j$ without loss of generality), we define $\delta_e f = f(i) - f(j)$, and for a face $f \in \mathcal{T}_h$, an assigned ordering of $\mathcal{T}_h$, and a function $f \in S_h$, we take f_+ and f_- to be the value of f at the higher (resp. lower) numbered simplex sharing the face. Direct computations yield

$$d_f = \tau \left[\int_\Omega K^{-1} |\phi_f|^2 \, \mathrm{d}x \right]^{-1}.$$

A is well known, for the symmetric problem the weight

$$\omega_e = \tau \left[\frac{1}{|\tau_e|} \int_e K^{-1} \, \mathrm{d}s \right]^{-1} \widetilde{\omega}_e,$$

with $|\tau_e|$ being the length of edge e, and $\widetilde{\omega}_e$, defined in [31], equation (2.5) uses geometric properties of the individual simplices, namely the angle between faces across edges and lengths of edges opposite these angles. The scaling of both of these coefficients is h^{d-2} and, as is immediately seen, the ratio of these terms is independent of h and τ, only dependent on mesh geometry and average values of K over edges versus over elements.

Taking $s = p$ and $v = w$, we arrive at

$$(A_E v, v) = \sum_{e \in \mathcal{T}_h} \omega_e (\delta_e v)^2 \quad \text{and} \quad (\widetilde{S}_R p, p) = \sum_{f \in \mathcal{T}_h} d_f (p_{T^+} - p_{T^-})^2. \tag{14}$$

Next, we show that (12) is a uniform preconditioner for the problem (8) under certain assumptions. To this end, we introduce the following notation,

- Ω_i is the subdomain consisting of simplices sharing vertex i, $\Omega_i = \cup_{T \ni i} T$,
- $\mathcal{F}_i$ is the set of faces containing vertex i, $\mathcal{F}_i = \{f \ni i\}$,
- $\mathcal{N}_i$ is the number of simplices sharing vertex i,
- $\mathcal{N}_i^f$ is the number of faces in $\mathcal{F}_i$,
- For an edge e with vertices i and j, $\mathcal{N}_{i\cup j}$ is the number of simplices in $\Omega_i \cup \Omega_j$, and $\mathcal{N}_{i\cap j}$ is the number of simplices in $\Omega_i \cap \Omega_j$.

To prove that this auxiliary space preconditioner is uniform, it is sufficient to prove the following three properties of the transfer operators Π_J, the auxiliary problem A_E, and the smoother $D_{\widetilde{S}}$ hold independent of h.

Lemma 3.2 (*Estimate for the Transfer Operator*)**.** *There exists $c_J > 0$ such that*

$$(\widetilde{S}_R \Pi_J v, \Pi_J v)^2 \le c_J^2 (A_E v, v), \quad \forall v \in V_h, \tag{15}$$

with c_J independent of mesh size.

Proof. Let $v \in V_h$. To show (15), we use the definition of Π_J and the relation (14),

$$\begin{aligned}(\widetilde{S}_R \Pi_J v, \Pi_J v)^2 &= \sum_{f \in \mathcal{T}_h} d_f \left(\frac{1}{d+1} \sum_{i \in T^+} v_i - \frac{1}{d+1} \sum_{j \in T^-} v_j \right)^2 \\ &= \sum_{f \in \mathcal{T}_h} \frac{d_f}{(d+1)^2} (v_{f,+} - v_{f,-})^2,\end{aligned}$$

where $v_{f,+}$ and $v_{f,-}$ are the values of v at the vertices in T^+ and T^- opposite face f, respectively. These vertices are not connected by any edge in $\mathcal{T}_h$, but can be connected via two edges $e^+ \in T^+$ and $e^- \in T-$. Using this fact we can relate the two bilinear forms:

$$\begin{aligned}&\sum_{f \in \mathcal{T}_h} \frac{d_f}{(d+1)^2} (v_{f,+} - v_{f,-})^2 = \sum_{f \in \mathcal{T}_h} \frac{d_f}{(d+1)^2} (\delta_{e^+} v + \delta_{e^-} v)^2 \\ &\le \sum_{f \in \mathcal{T}_h} \frac{2 d_f}{(d+1)^2} \left[(\delta_{e^+} v)^2 + (\delta_{e^-} v)^2 \right] \le \frac{2 D \kappa_e}{(d+1)^2} \sum_{e \in \mathcal{T}_h} \omega_e (\delta_e v)^2 \\ &= c_J^2 (A_E v, v),\end{aligned}$$

with D being a scaling constant changing from the weights d_f to ω_e that is independent of discretization parameters h and τ, and κ_e being an upper bound on the number of times an edge can be used on the sum over faces, $\kappa_e = \max_{e \in \mathcal{T}_h} 2\mathcal{N}_{i\cap j}$. Thus, this c_J is indeed independent of mesh size. □

Remark 3.3. Note that the estimate for κ_e is conservative. In three dimensions, one can almost always use different edges to connect the points e^+ and e^-, so that in practice, c_J is smaller.

Lemma 3.4 (*Continuity of the Smoother*)**.**

$$\exists\, c_{\widetilde{S}_R} > 0: \quad (\widetilde{S}_R v, v) \le c^2_{\widetilde{S}_R} (D_{\widetilde{S}} v,\ v), \quad \forall v \in S_h.$$

with $c_{\widetilde{S}_R}$ independent of mesh size.

Proof. Given $\widetilde{S}_R$ is symmetric positive definite (SPD), we use the Cauchy–Schwarz inequality using the standard Euclidean basis $\{e_i\}$ to obtain

$$|\widetilde{S}^{ij}_R| = |(\widetilde{S}_R e_i, e_j)| = |(e_i, e_j)_{\widetilde{S}_R}| \le \|e_i\|_{\widetilde{S}_R} \|e_j\|_{\widetilde{S}_R} = \sqrt{\widetilde{S}^{ii}_R \widetilde{S}^{ij}_R}.$$

Next, using this inequality we obtain

$$\left| [D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2}]_{ij} \right| = \frac{|\widetilde{S}^{ij}_R|}{\sqrt{\widetilde{S}^{ii}_R \widetilde{S}^{ij}_R}} \le 1. \tag{16}$$

Therefore, we have

$$\begin{aligned} c_{\widetilde{S}_R} &= \max_{v \in S_h} \frac{(\widetilde{S}_R v, v)}{(D_{\widetilde{S}} v, v)} = \max_{w = D_{\widetilde{S}}^{1/2} v \in S_h} \frac{(D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2} w, w)}{(w, w)} \\ &= \rho\left(D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2} \right) \le \| D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2} \|_\infty. \end{aligned}$$

From the inequality (16) it follows that $\| D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2} \|_\infty$ is bounded by the number of nonzeros per row in $\widetilde{S}_R$, which can be bounded by the number of faces an element has, i.e., $(d+1)$. □

Lemma 3.5 (*Stable Decomposition*)**.** *For every $p \in S_h$, there exist $p_0 \in S_h$ and $w_J \in V_h$ such that $p = p_0 + \Pi_J w_J$ and*

$$(D_{\widetilde{S}} p_0,\ p_0) + (A_E w_J,\ w_J) \le c_0^2 \|p\|^2_{\widetilde{S}_R}, \tag{17}$$

where $c_0 > 0$ should be small and independent of mesh size.

Proof. To bound the first term on the left hand side of (17), we set the value of $w_J \in V_h$ at a vertex i to equal the average of the values of $p \in S_h$ on the simplices T which surround the vertex i. More precisely,

$$V_h \ni w_J = \sum_i [w_J]_i \phi_i, \quad [w_J]_i = \frac{1}{\mathcal{N}_i} \sum_{T \in \Omega_i} p_T.$$

By the definition (14) and rearranging the decomposition of p, for each $T \in \mathcal{T}_h$

$$\begin{aligned} [p_0]_T = [p - \Pi_J w_J]_T &= p_T - \frac{1}{d+1} \sum_{i \in T} [w_J]_i \\ &= p_T - \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} p_{T'} \\ &= \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} (p_T - p_{T'}). \end{aligned}$$

We now make the following estimate for this fixed T, using two applications of Cauchy–Schwarz:

$$\begin{aligned} &\left(p_T - \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} p_{T'} \right)^2 \\ &= \left(\frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} (p_T - p_{T'}) \right)^2 \\ &\le \frac{1}{(d+1)^2 n_T^2} \left(\sum_{i \in T} \sum_{T' \in \Omega_i} (p_T - p_{T'}) \right)^2 \\ &\le \frac{1}{(d+1) n_T} \sum_{i \in T} \sum_{T' \in \Omega_i} (p_T - p_{T'})^2. \end{aligned}$$

with $n_T = \min_{i\in T} \mathcal{N}_i$. For each simplex $T' \in \Omega_i$, we expand the difference $(p_T - p_{T'})$ as a telescoping sum of differences across faces, $\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})$. To define this set of faces $\mathcal{F}_i(T, T')$, we must first define a chain of pairwise–adjacent simplices $\{T_j\}_{j=1}^{J_i} \subset \Omega_i$ of minimal length, with $T_1 = T$ and $T_{J_i} = T'$. Then for this chain,

$$\mathcal{F}_i(T, T') = \{f \in \mathcal{F}_i : f_j = T_{j+1} \cap T_j, \quad j = 1, \dots, J_i - 1\}.$$

Denoting $\mathcal{N}^f_{\mathcal{F}_i(T,T')}$ as the number of faces in $\mathcal{F}_i(T, T')$, we get the following:

$$\begin{aligned}
&\left(p_T - \frac{1}{d+1}\sum_{i\in T}\frac{1}{\mathcal{N}_i}\sum_{T'\in\Omega_i} p_{T'}\right)^2 \\
&\le \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}(p_T - p_{T'})^2 = \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}\left(\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})\right)^2 \\
&\le \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2. \qquad (18)
\end{aligned}$$

Since $\{T_j\}_{j=1}^{J_i} \subset \Omega_i$, the length J_i of each simplex chain length connecting the simplices T and T' is bounded from above by $\mathcal{N}_i/2$; otherwise a shorter chain of simplices connecting T and T' must exist.

Given this observation, we can reorder the two innermost sums in the last inequality above:

$$\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2 \le \sum_{f\in\mathcal{F}_i}(p_{T+} - p_{T-})^2\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')},$$

which gives us the bound

$$\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2 \le \frac{F(F+1)}{2}\sum_{f\in\mathcal{F}_i}(p_{T+} - p_{T-})^2, \qquad (19)$$

with $F = \max_{i\in\mathcal{T}_h}\max_{T,\,T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}$. As each $\mathcal{N}^f_{\mathcal{F}_i(T,T')} = J_i - 1$, taking $J = \max_{i\in\mathcal{T}_h} J_i$ it follows that $F \le \max_{i\in\mathcal{T}_h}\frac{\mathcal{N}_i}{2} - 1$, which gives us a uniform estimate for F.

Combining (18) and (19) with the observation that any face $f \in \mathcal{F}_i$, $i \in T$ is shared by at most d vertices in T, and that each face f is globally shared by at most 2 simplices in the mesh, we get the final set of inequalities,

$$\begin{aligned}
(D_{\widetilde{S}}p_0,\ p_0)^2 &= \sum_{T\in\mathcal{T}_h} d_T(p_0)_T^2 = \sum_{T\in\mathcal{T}_h} d_T(p - \Pi_J w_J)^2 \\
&\le \frac{dF(F+1)D_*}{(d+1)n}\sum_{f\in\mathcal{T}_h} d_f(p_{T+} - p_{T-})^2 = c_D^2\|p\|^2_{\widetilde{S}_R},
\end{aligned}$$

with $n = \min_{T\in\mathcal{T}_h} n_T$ and $D_* = \max_{T,f\in\mathcal{T}}\frac{d_T}{d_f}$, thus giving us a constant independent of mesh size due to the weights d_T and d_f being of the same order in h.

Now we need to bound the second term on the left hand side of (17). Taking the same definition of w_J as above, our goal is to show $(A_E w_J, w_J) \le c_A^2\|p\|^2_{\widetilde{S}}$. Since $(A_E w_J, w_J) = \sum_{e\in\mathcal{T}_h} w_e\left(\frac{1}{\mathcal{N}_i}\sum_{T\in\Omega_i} p_T - \frac{1}{\mathcal{N}_j}\sum_{T'\in\Omega_j} p_{T'}\right)^2$, we fix an edge $e \in \mathcal{T}_h$ to estimate each term of the sum first.

Note that for any constant C, $\left(\frac{1}{\mathcal{N}_i}\sum_{T\in\Omega_i} C - \frac{1}{\mathcal{N}_j}\sum_{T'\in\Omega_j} C\right)^2 = (C - C)^2 = 0$. Then we have,

$$\begin{aligned}
&\left(\frac{1}{\mathcal{N}_i}\sum_{T\in\Omega_i} p_T - \frac{1}{\mathcal{N}_j}\sum_{T'\in\Omega_j} p_{T'}\right)^2 = \left(\sum_{T\in\Omega_i\cup\Omega_j}(p_T - C)\left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} - \frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j}\right)\right)^2 \\
&\le \sum_{T\in\Omega_i\cup\Omega_j}(p_T - C)^2\sum_{T\in\Omega_i\cup\Omega_j}\left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} - \frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j}\right)^2 \\
&\le \left(\sum_{T\in\Omega_i\cup\Omega_j}\left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i}\right)^2 + \sum_{T\in\Omega_i\cup\Omega_j}\left(\frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j}\right)^2\right)\inf_{C\in\mathbb{R}}\|\boldsymbol{p}_{\Omega_i\cup\Omega_j} - C\mathbf{1}\|^2_{\ell^2},
\end{aligned}$$

where the first inequality is the Cauchy–Schwarz inequality and the second is due to both the Cauchy–Schwarz inequality and the non-negativity of the terms in the second summand. Here $\chi_{\Omega_k}(T)$ is the characteristic function on set Ω_k, the vector $\boldsymbol{p}_{\Omega_i\cup\Omega_j} = (p_{T_1}, \ldots, p_{T_{\mathcal{N}_{i\cup j}}})^T$ denotes the values of p on each simplex in $\Omega_i \cup \Omega_j$, and $\mathbf{1}$ is the vector of the same size as $\boldsymbol{p}_{\Omega_i\cup\Omega_j}$ with ones on each entry. It is well known that the C that will minimize the ℓ^2-norm in this scenario is the average of p over each simplex in the union, $\bar{p} = \frac{1}{\mathcal{N}_{i\cup j}} \sum_{T\in\Omega_i\cup\Omega_j} p_T$. Using this fact and that both $\frac{\chi_{\Omega_k}(T)}{\mathcal{N}_k} \leq 1$ and $\sum_{T\in\Omega_i\cup\Omega_j} \frac{\chi_{\Omega_k}(T)}{\mathcal{N}_k} = 1$ for $k = i, j$, we can continue our estimates,

$$\begin{aligned}
&\left(\sum_{T\in\Omega_i\cup\Omega_j} \left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i}\right)^2 + \sum_{T\in\Omega_i\cup\Omega_j} \left(\frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j}\right)^2 \right) \inf_{C\in\mathbb{R}} \|\boldsymbol{p}_{\Omega_i\cup\Omega_j} - C\mathbf{1}\|^2_{\ell^2} \\
&\leq 2\|\boldsymbol{p}_{\Omega_i\cup\Omega_j} - \bar{p}\mathbf{1}\|^2_{\ell^2} \leq 2\mathcal{N}_{i\cup j}\mathrm{Diam}(\Omega_i\cup\Omega_j)(Lp,p)_{\Omega_i\cup\Omega_j} \\
&= \gamma^2_{P,e} \sum_{f\in\Omega_i\cup\Omega_j} w_{f,L}(p_{T+} - p_{T-})^2.
\end{aligned}$$

The second inequality is due to the Poincaré inequality, with $\|\nabla p\|^2_{\ell^2}$ expressed as the bilinear form $(Lp, p)_{\Omega_i\cup\Omega_j}$, which is the local action of the graph Laplacian. The weights $w_{f,L}$ are the weights required to form the local Laplacian bilinear form across faces.

Finally, we incorporate the sum over all edges,

$$\begin{aligned}
(A_E w_J, w_J) &= \sum_{e\in\mathcal{T}_h} w_e(\delta_e w_J)^2 \\
&\leq \gamma_P^2 \sum_{e\in\mathcal{T}_h} \sum_{f\in\Omega_i\cup\Omega_j} w_{f,L}(p_{T+} - p_{T-})^2 \\
&\leq D\gamma_P^2 \sum_{f\in\mathcal{T}_h} d_f(p_{T+} - p_{T-})^2 = c_A^2\|p\|^2_{\widetilde{S}},
\end{aligned}$$

where $\gamma_P = \max_{e\in\mathcal{T}_h} \gamma_{P,e}$, $D = d\,\alpha\,\widetilde{D}$, with $\alpha = \max_{i\in\mathcal{T}_h}\{\#\text{ vertices} \in \Omega_i\}$ and $\widetilde{D}$ is a scaling factor used to change from the weights $w_{f,L}$ to d_f, again independent of mesh size due to both weights being of the same order in h.

Taking the max of c_D and c_A as c_0 in (17) gives us the required uniform bound, which completes the proof. □

As shown in [18] and [17], the last three lemmas guarantee that our preconditioner is uniform, as they combine to provide the hypotheses required to apply the fictitious space lemma, introduced in [16]:

Theorem 3.6 (*Fictitious Space Lemma*)**.** *Assume that Π is surjective and*

$$\exists c_0 > 0: \quad \forall v \in V: \quad \exists \bar{v} \in \bar{V}: \quad v = \Pi\bar{v} \quad \text{and} \quad \|\bar{v}\|_{\bar{A}} \leq c_0\|v\|_A, \tag{20}$$

$$\exists c_1 > 0: \quad \|\Pi\bar{v}\|_A \leq c_1\|\bar{v}\|_{\bar{A}} \quad \forall \bar{v} \in \bar{V}. \tag{21}$$

Then

$$c_0^{-2}\|v\|^2_A \leq (BAv,\ v)_A \leq c_1^2\|v\|^2_A \quad \forall v \in V,$$

with c_0 and c_1 independent of mesh size.

Corollary 3.7. *The auxiliary space preconditioner defined in (12) provides a uniform preconditioner for $\widetilde{S}_R$.*

Proof. Taking $V = V_h$, $\bar{V} = S_h \times V_h$ and $\Pi = \begin{pmatrix} I & \Pi_J \end{pmatrix}$ as defined before. Let $(v, v)_A = (\widetilde{S}_R v, v)$, $v \in S_h$, $(\bar{v}, \bar{v})_{\bar{A}} = (D_{\widetilde{S}_R} v_0, v_0) + (A_E v_h, v_h)$, $\bar{v} = (v_0, v_h) \in \bar{V}$, then Lemmas 3.2 and 3.4 prove that our scheme fulfills the second assumption (21) of Theorem 3.6, and Lemma 3.5 shows that the first assumption (20) is satisfied. Then direct application of Theorem 3.6 shows that B is a uniform preconditioner for $\widetilde{S}_R$. □

4. Numerical tests

In our numerical tests we solved the linear system (7) for the first L-scheme iteration as outlined above with an outer GMRES iterations preconditioned with the Schur iteration Algorithm 2.1 with relative residual tolerance 5×10^{-8}. The inner solve of $\widetilde{S}_R$ system (8) is done using Preconditioned Conjugate Gradient (PCG) with the auxiliary space method described in the previous section as preconditioner. The inner iterations were stopped when the relative residual was smaller than 10^{-9}. To solve the auxiliary problem A_E to machine precision, we used unsmoothed aggregation algebraic multigrid (UA-AMG) method.

Table 1
Number of outer GMRES/average inner PCG iterations (rounded to nearest integer) for solving the linearization of the mixed form of RE, using the analytic K and θ as described in Test 1. Here the mesh size is $h = 2^{-p}$ and timestep $\tau = 1$.

p	2	3	4	5	6
Outer/Inner	5/13	5/15	5/15	5/15	5/16

For both examples outlined below, we set $\Omega = [0, 2]^3$ with uniform tesselation $\mathcal{T}_h$ of characteristic element size $h = 2^{-p}$, and defined an analytic solution $\Psi_e(x, t)$, with source term being determined analytically by plugging in the solution into Richards' equation. We used Dirichlet data for $\Psi(x, t)$ for all $t > 0$, $\Psi(x, t)|_{\partial\Omega} = \Psi_e(x, t)$, and initial condition $\Psi_0 = \Psi_e(x, 0)$.

4.1. Example 1: Continuously varying K

For our first example, we chose $\Psi_e(x, t) = -10t|x|^2$,

$$\theta(\Psi) = \exp(\Psi), \quad K(\theta) = (K_{\max} - K_{\min})\theta + K_{\min},$$

with $K_{\min} = 1 \times 10^{-6}$ and $K_{\max} = 1$. Thus, K varies continuously by several orders of magnitude from the bottom of the cube to the top.

Table 1 lists the number of outer GMRES and average inner PCG iterations for the preconditioned linear solve of the first L-scheme step for this problem. To give some perspective on the size of the respective mesh sizes used to test, for $h = 2^{-6}$, the size of the full system is over 4.5 million DOF, with over 1.5 million pressure unknowns; the size of the auxiliary system used to precondition (8) in the Schur iteration is around 262,000 unknowns, which is a reduction in degrees of freedom of roughly 1/6. It should be noted that since the auxiliary system is as poorly conditioned as (8), a preconditioner can (and should in practice) be used to solve the auxiliary problem efficiently, as the performance of the auxiliary space preconditioner depends on solving the auxiliary problem accurately.

4.2. Example 2: Van Genuchten–Mualem (VGM) model

For the next example we use a setup given in [8] and we consider the so-called Van Genuchten–Mualem (VGM) model for K and θ:

$$\theta(\Psi) = \begin{cases} \theta_R + (\theta_S - \theta_R)\left[1 + |\alpha\Psi|^n\right]^{\frac{1}{n}-1}, & \Psi < 0, \\ \theta_S, & \Psi \geq 0. \end{cases} \tag{22}$$

$$K(\theta) = \begin{cases} K_S\theta_N(\Psi)^{\frac{1}{2}}\left[1 - \left(1 - \theta_N(\Psi)^{\frac{n}{n-1}}\right)^{\frac{n-1}{n}}\right]^2, & \Psi < 0, \\ K_S, & \Psi \geq 0. \end{cases} \tag{23}$$

Let θ_N be the dimensionless water content defined as

$$\theta_N(\Psi) = \frac{\theta(\Psi) - \theta_R}{\theta_S - \theta_R}.$$

The parameters $n > 1$ and α are typically determined using the so called log-slope technique based on test data from field experiments of water infiltration using different compositions of materials [36]. Elementary analysis shows that θ is Lipschitz continuous for $n > 1$ at $\theta = 1$, and that K is also Lipschitz continuous for $n > 2$; this corresponds to media with sufficiently uniform pore size distribution.

For the second test related to the VGM model, we considered the same choice of test function and boundary/initial conditions as the first, but use the VGM K and θ as defined above. The test runs for values of α, n, and K_S for three different media: Beit Netofa clay ($\alpha = 0.152$, $n = 1.17$, $K_S = 8.2 \times 10^{-4}$), silt loam ($\alpha = 0.423$, $n = 2.06$, $K_S = 5 \times 10^{-2}$), and clay loam ($\alpha = 1.9$, $n = 1.31$, $K_S = 6.2 \times 10^{-2}$). The Beit Netofa clay and the silt loam examples are from in [8], and the clay loam example is from [13].

Note that K reduces roughly 4 orders of magnitude from the bottom of the cube to the top, roughly 8 orders for the Beit Netofa clay, and roughly 10 orders for the clay loam. Due to the high contrast, we increased the number of V-cycles to 60 for the auxiliary space solve.

As Table 2 shows, even for large contrast K, our preconditioner maintains its robustness, though the authors would like to remind that getting linearization schemes to converge for such high contrast K is a different challenge that has been well documented in the literature (e.g. [6,11,13,15]).

Table 2
Average inner PCG iterations for Stilde solve for each of three different media after preconditioning with aux space preconditioner. Here the mesh size is $h = 2^{-p}$ and timestep $\tau = 1$.

p	2	3	4	5	6
Beit Netofa clay	8	12	15	15	15
Silt loam	11	14	15	15	15
Clay loam	9	13	16	17	17

Acknowledgment

The authors would like to thank Anna Mazzucato for many productive discussions on the theoretical results concerning the well-posedness of the Richards' equation.

References

[1] L. Richards, Capillary conduction of liquids through porous medium, Physics 1 (1931) 318–333.
[2] Y. Wu, J. Kool, J. McCord, An evaluation of alternative numerical formulations for two-phase air water flow in unsaturated soils, in:Proceedings of the American Geophysical Union Spring meeting, Montreal, 1992.
[3] P. van der Heidje, Compilation of Saturated and Unsaturated Zone Modeling Software, Tech. Rep. EPA/R-96/009, 1996.
[4] H.W. Alt, S. Luckhaus, Quasilinear elliptic–parabolic differential equations, Math. Z. 183 (3) (1983) 311–341, http://dx.doi.org/10.1007/BF01176474.
[5] F. Otto, L1-contraction and uniqueness for quasilinear elliptic–parabolic equations, J. Differential Equations 131 (0155) (1996) 20–38.
[6] C.T. Miller, et al., Multiphase flow and transport modeling in heterogeneous porous media: Challenges and approaches, Adv. Water Resour. 21 (1998a) 77–120.
[7] M.A. Celia, E.T. Bouloutas, R.L. Zarba, A general mass-conservative numerical solution for the unsaturated flow equation, Water Resour. Res. 26 (7) (1990) 1483–1496, http://dx.doi.org/10.1029/WR026i007p01483.
[8] F. List, F.A. Radu, A study on iterative methods for solving Richards' equation, Comput. Geosci. 20 (2) (2016) 341–353, http://dx.doi.org/10.1007/s10596-016-9566-3.
[9] T. Arbogast, M.F. Wheeler, N.-Y. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, SIAM J. Numer. Anal. 33 (4) (1996) 1669–1687, http://dx.doi.org/10.1137/S0036142994266728.
[10] L. Bergamaschi, M. Putti, Mixed finite elements and Newton-type linearizations for the solution of Richards' equation, Internat. J. Numer. Methods Engrg. 45 (8) (1999) 1025–1046, http://dx.doi.org/10.1002/(SICI)1097-0207(19990720)45:8<1025::AID-NME615>3.3.CO;2-7.
[11] P. Forsyth, Y. Wu, K. Pruess, Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media, Adv. Water Resour. 18 (1) (1995) 25–38, http://dx.doi.org/10.1016/0309-1708(95)00020-J, URL http://www.sciencedirect.com/science/article/pii/030917089500020J.
[12] I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure, J. Comput. Appl. Math. 168 (1–2) (2004) 365–373, http://dx.doi.org/10.1016/j.cam.2003.04.008.
[13] C.T. Miller, G.A. Williams, C. Kelley, M.D. Tocci, Robust solution of Richards' equation for nonuniform porous media, Water Resour. Res. 34 (1998b) 2599–2610.
[14] P. Forsyth, M. Kropinski, Monotonicity considerations for saturated-unsaturated subsurface flow, SIAM J. Sci. Comput. 18 (1997) 1328–1354.
[15] M.W. Farthing, F.L. Ogden, Numerical solution of Richards' equation: A review of advances and challenges, Soil Sci. Soc. Am. J. 81 (2017) 1257–1269.
[16] S.V. Nepomnyaschikh, Decomposition and fictitious domains methods for elliptic boundary value problems, in: Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Norfolk, VA, 1991, SIAM, Philadelphia, PA, 1992, pp. 62–72.
[17] J. Xu, The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids, Computing 56 (1996) 215–235.
[18] R. Hiptmair, J. Xu, Nodal auxiliary space preconditioning in H(curl) and H(div) spaces, SIAM J. Numer. Anal. 45 (2007) 2483–2509.
[19] R.S. Tuminaro, J. Xu, Y. Zhu, Auxiliary space preconditioners for mixed finite element methods, in: M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (Eds.), Domain Decomposition Methods in Science and Engineering XVIII, in: Lecture Notes in Computational Science and Engineering, Springer, Berlin Heidelberg, 2009, pp. 99–109.
[20] A. Budiša, X. Hu, Block preconditioners for mixed-dimensional discretization of flow in fractured porous media, 2019, arXiv preprint arXiv:1905.13513.
[21] J. Adler, X. Hu, L. Zikatanov, Robust solvers for Maxwell's equations with dissipative boundary conditions, SIAM J. Sci. Comput. 39 (5) (2017) S3–S23, http://dx.doi.org/10.1137/16M1073339.
[22] J.H. Adler, F.J. Gaspar, X. Hu, C. Rodrigo, L.T. Zikatanov, Robust Block Preconditioners for Biot's Model, in: Lecture Notes in Computational Science and Engineering, vol. 125, Springer, 2018, pp. 3–16, arXiv:1705.08842.
[23] M.W. Farthing, C. Kees, C. Miller, Mixed finite element methods and higher order temporal approximations for variably saturated groundwater flow, Adv. Water Resour. 26 (2003) 373–394.
[24] C.S. Woodward, C.N. Dawson, Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media, SIAM J. Numer. Anal. 37 (3) (2000) 701–724, http://dx.doi.org/10.1137/S0036142996311040 (electronic).
[25] P.-A. Raviart, J.M. Thomas, A mixed finite element method for 2nd order elliptic problems, in: Mathematical Aspects of Finite Element Methods, Proc. Conf., Consiglio Naz. delle Ricerche (CNR), Rome, 1975, in: Lecture Notes in Math., vol. 606, Springer, Berlin, 1977, pp. 292–315.
[26] J.-C. Nédélec, Mixed finite elements in $\mathbf{R}^3$, Numer. Math. 35 (3) (1980) 315–341, http://dx.doi.org/10.1007/BF01396415.
[27] J.-C. Nédélec, A new family of mixed finite elements in $\mathbf{R}^3$, Numer. Math. 50 (1) (1986) 57–81, http://dx.doi.org/10.1007/BF01389668.
[28] D.M. Young, Iterative Solution of Large Linear Systems, Academic Press, New York-London, 1971.
[29] R. Hiptmair, Finite elements in computational electromagnetism, Acta Numer. 11 (2002) 237–339.
[30] F. Brezzi, M. Fortin, L. Marini, Error analysis of piecewise constant pressure approximations of darcy's law, Comput. Methods Appl. Mech. Eng. 195 (2006) 1547–1559.
[31] J. Xu, L. Zikatanov, A monotone finite element scheme for convection–diffusion equations, Math. Comp. 68 (228) (1999) 1429–1446, http://dx.doi.org/10.1090/S0025-5718-99-01148-5.
[32] J. Xu, L. Zikatanov, Algebraic multigrid methods, Acta Numer. 26 (2017) 591–721, http://dx.doi.org/10.1017/S0962492917000083.

[33] P. Markowich, M. Zlamal, Inverse-average-type finite element discretizations of self-adjoint second order elliptic problems, Math. Comp. 51 (1989) 431–449.
[34] T. Barth, Aspects of Unstructured Grids and Finite-Volume Solvers for the Euler and Navier-Stokes Equations, Tech. Rep. 787, 1992, AGARD, special course on unstructured grids methods for advection dominated flows.
[35] M. Bern, D. Eppstein, Mesh generation and optimal triangulation, in: Computing in Euclidean Geometry, World Scientific, 1992, pp. 23–90.
[36] M.T. Van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Sci. Soc. Am. J. 44 (5) (1980) 892–898.