Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1452–1478

ORDER OF CONVERGENCE ESTIMATES FOR AN EULER IMPLICIT, MIXED FINITE ELEMENT DISCRETIZATION OF RICHARDS' EQUATION*

FLORIN RADU†, IULIU SORIN POP‡, AND PETER KNABNER†

Abstract. We analyze a discretization method for a class of degenerate parabolic problems that includes the Richards' equation. This analysis applies to the pressure-based formulation and considers both variably and fully saturated regimes. To overcome the difficulties posed by the lack in regularity, we first apply the Kirchhoff transformation and then integrate the resulting equation in time. We state a conformal and a mixed variational formulation and prove their equivalence. This will be the underlying idea of our technique to get error estimates.

A regularization approach is combined with the Euler implicit scheme to achieve the time discretization. Again, equivalence between the two formulations is demonstrated for the semidiscrete case. The lowest order Raviart–Thomas mixed finite elements are employed for the discretization in space. Error estimates are obtained, showing that the scheme is convergent.

Key words. error estimates, Euler implicit scheme, mixed finite elements, regularization, degenerate parabolic problems, porous media, Richards' equation

AMS subject classifications. 65M12, 65M15, 65M60, 76S05, 35K65, 35K55

DOI. 10.1137/S0036142902405229

1. Introduction.

A commonly accepted mathematical model of water flow in porous media is the Richards' equation, a nonlinear, possibly degenerate, parabolic differential equation. In the pressure formulation, Richards' equation [5] is expressed as

$$\partial_t \Theta(\psi) - \nabla \cdot K(\Theta)\nabla(\psi + z) = 0, \tag{1.1}$$

where ψ is the pressure head, Θ the saturation, K the conductivity, and z the height against the gravitational direction. The equation (1.1) models the flow of a wetting fluid (water) in a porous media in the presence of a nonwetting fluid (air) supposed to be at constant pressure, 0. In the saturated region (where only water is present) we have $\psi \geq 0$, while $\psi < 0$ in the unsaturated domain. Different functional dependencies (retention curves) between ψ, K and Θ are proposed in the literature. These are provided essentially by soil particularities and allow reducing all the unknowns in the above equation to a single one. Here we are interested in both partially saturated and saturated flow, therefore we retain the pressure ψ as primary unknown.

As suggested in [1], applying the Kirchhoff transformation

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\longrightarrow \mathbb{R}, \\ \psi &\longmapsto \int_0^{\psi} K(\Theta(s))\, ds \end{aligned} \tag{1.2}$$

*Received by the editors April 9, 2002; accepted for publication (in revised form) October 29, 2003; published electronically December 16, 2004. This work was supported by the Netherlands Organization for Scientific Research (NWO) through project 809.62.010 of Earth and Life Sciences (ALW).

http://www.siam.org/journals/sinum/42-4/40522.html

†Institute of Applied Mathematics, University Erlangen-Nürnberg, Martensstr. 3, D-91058 Erlangen, Germany (raduf@am.uni-erlangen.de, knabner@am.uni-erlangen.de).

‡Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (I.Pop@tue.nl).

leads to unknowns that are more regular. Since $K(\Theta(s))$ is positive, this transformation can be inverted and equation (1.1) can be rewritten in terms of a new variable, $u := \mathcal{K}(\psi)$. Now defining

$$\begin{aligned} b(u) &:= \Theta \circ \mathcal{K}^{-1}(u), \\ k(b(u)) &:= K \circ \Theta \circ \mathcal{K}^{-1}(u) \end{aligned} \tag{1.3}$$

and letting e_z denote the vertical unit vector, (1.1) becomes

$$\partial_t b(u) - \nabla \cdot (\nabla u + k(b(u))\ e_z) = 0 \quad \text{in} \quad (0,T] \times \Omega. \tag{1.4}$$

By the above transformation, diffusion becomes linear in equation (1.1). However, the problem may still remain degenerate, leading to solutions lacking regularity. Since this equation models important practical problems, several papers are dealing with analysis and numerical methods for it. Euler methods are often employed for the discretization in time. Adaptive time stepping is studied in [25], [14], or [28]. In case of an implicit discretization, iterative methods are considered (see, for example, [16], [8], whose method was already proposed in [12] and used also in [15], and [14]).

For the spatial discretization, mixed finite elements or finite volumes provide a good approximation of the solution [17], [4], [6], [10]. The most comprehensive algorithmic approach has been presented in the thesis [25], where hybrid mixed finite elements and an implicit Euler discretization are used. The set of nonlinear equations is solved by a Newton/multigrid method, while time and space adaptive strategies are constructed on the basis of rigorous error indicators. However, most of the authors are mainly interested in computational aspects and less concerned with rigorous convergence results. With respect to this last aspect we mention [2], where a model nonlinear, degenerate, advection-diffusion equation is considered. Through time integration a mixed variational formulation respecting the known minimal regularity of the solution is obtained. Raviart–Thomas lowest order finite elements are used. A priori error estimates are derived for the time integral of the flux and for the saturation. The estimates are optimal for the semidiscrete (continuous in time), noncomputable scheme. In the degenerate case, an explicit order of convergence for the fully discrete scheme can be deduced only by assuming extra (nonrealistic) regularity for the solution. Using similar techniques, [26] proved also some a priori error estimates for a mixed finite element discretization of Richards' equation. Unfortunately, again an explicit order of convergence for the scheme can be derived only in the nondegenerate case. Another important paper is [21], which deals with a class of multidimensional degenerate parabolic equations, including Richards' equation. A fully discrete scheme based on C^0 piecewise linear finite elements in space and a semi-implicit discretization in time is proposed and analyzed. An explicit order of convergence $(\tau^{1/2} + h)$ is proved. The techniques used here to cope with degenerate parabolic equations, which have been not used so far for discretizations based on mixed finite element method, will permit us to extend the results in [2] to the general, degenerate case. As in [26], error bounds for the time integral of the pressure will be also derived. We note also the recent paper [29], where the techniques from [2] are used for the numerical analysis of an expanded mixed finite element discretization of the Richards' equation. Also employed here are the Raviart–Thomas lowest order finite elements. Convergence rates depending on the Hölder continuity of the capacity term are derived for the entire regime of fully saturated to fully unsaturated flow. Nevertheless, the expanded mixed finite element method is not equivalent with the standard mixed finite element method and their results cannot be simply transferred to our method. Finally, we

mention also [10] (where convergence of an implicit finite volume method is proven by compactness arguments), [13] (for a relaxation scheme that applies to this equation too), and [23] (where error estimates are obtained for the unsaturated regime).

Here we consider an increasing and Lipschitz continuous b. Nevertheless, $b'(u)$ may be 0 for some values of u (not necessary isolated). Our numerical approach employs the lowest order Raviart–Thomas finite elements in space and Euler implicit in time, together with a regularization step. Specifically, with $N > 0$ integer, set $\tau = T/N$ and let $\mathcal{T}_h$ be a decomposition of Ω into closed d-simplices; h stands for the mesh-size. In a formal writing, the numerical scheme under consideration reads as

$$\begin{aligned} b_\epsilon(p_h^n) + \tau \nabla \cdot q_h^n &= b_\epsilon(p_h^{n-1}), \\ q_h^n + \nabla p_h^n + k(b(p_h^n))e_z &= 0 \end{aligned}$$

for $n = \overline{1, N}$; p_h^0 approximates u^0 in the finite dimensional approximation space. The term ∇p_h^n should be understood in a weak sense. Here b_ϵ is a regular approximation of b depending on the small parameter $\epsilon > 0$. By p_h^n we denote a piecewise constant approximation of u and q_h^n is a Raviart–Thomas (RT_0) approximation of the flux $-(\nabla u + k(b(u))e_z)$, based on $\mathcal{T}_h$, both at $t = n\tau$.

As suggested in [2], to overcome the difficulties posed by the lack in regularity, equation (1.4) is first integrated in time. For the resulting problem a mixed variational formulation is stated.

Convergence is shown by obtaining first error estimates for the time discrete scheme, by following the ideas in [21]. Since we work in a slightly more general framework, we include for completeness the proof for the conformal formulation. Next, using the procedure described in [2, 26], error estimates for the fully discrete scheme are obtained. In this setting, the equivalence between the mixed and conformal formulations becomes essential since, in this way, results obtained for one case can be transferred to the other one.

The outline of the paper is as follows. First, we state the main assumptions and notations used throughout the paper, define the problem to be solved, and discuss questions regarding existence and regularity of a solution. In section 2 the equivalence between a conformal and a mixed variational formulations is proved, for the continuous case as well as for the time discrete one. In section 3 we investigate the stability of the numerical scheme, while error estimates are derived in section 4.

1.1. Notations and assumptions. In what follows we let Ω be a domain in $\mathbb{R}^d$ (with $d = 1, 2,$ or 3). Let $J = (0, T]$ be a finite time interval. We are interested in solving (1.4) endowed with initial and boundary conditions,

$$\begin{aligned} \partial_t b(u) - \nabla \cdot (\nabla u + k(b(u))e_z) &= 0 \qquad \text{in} \quad J \times \Omega, \\ u &= u^0 \qquad \text{in} \quad 0 \times \Omega, \\ u &= 0 \qquad \text{on} \quad J \times \Gamma. \end{aligned} \tag{1.5}$$

Throughout this paper we make use of the following assumptions.

(A1) $\Omega \subset \mathbb{R}^d$ is bounded with Lipschitz continuous boundary.

(A2) $b \in C^1$ is nondecreasing and Lipschitz continuous.

(A3) $k(b(z))$ is continuous and bounded in z and satisfies, for all $z_1, z_2 \in \mathbb{R}$, $| k(b(z_2)) - k(b(z_1)) |^2 \leq C_k(b(z_2) - b(z_1))(z_2 - z_1)$.

(A4) $b(u_0)$ is essentially bounded (by 0 and 1) in Ω and $u_0 \in L^2(\Omega)$.

REMARK 1.1. *By* (A3)*, the convection term is bounded. This restriction is not unrealistic since, for Richards' equation, k stands for the conductivity of the medium.*

This assumption makes our analysis easier, but can be avoided. Moreover, the growth condition on $k(b(\cdot))$ (see also [11], [27], [30], *or* [23]*) relaxes the more often assumed Lipschitz continuity of k (see, e.g.,* [21], [2]*). It gives uniqueness for the weak solution, as shown in* [1]*. In addition, source terms can also be considered here, provided that they satisfy a similar growth condition as $k(b(u))$.*

REMARK 1.2. *In the transformed version, Richards' equation fits in our framework. However, since b is Lipschitz, a vanishing permeability in* (1.1) *is not allowed, meaning that our analysis is valid in the variably saturated to fully saturated flow regimes, but not in the completely air saturated one.*

REMARK 1.3. *For the sake of simplicity, we deal with homogeneous Dirichlet boundary conditions. More general situations can be included in a straightforward manner, with similar results. Here nonlinearities depend only on the unknown u, not on x and t. For more general situations, techniques developed in* [2] *can be employed.*

Because of its degenerate character, we do not expect smooth solutions for problem (1.5). For defining a solution in a weak sense we let $(\cdot,\cdot)$ stand for the inner product on $L^2(\Omega)$ or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, $\|\cdot\|$ for the norm in $L^2(\Omega)$, and $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ for the norms in $H^1(\Omega)$ and $H^{-1}(\Omega)$, respectively. We use analogous notations for the inner product and the corresponding norm on $L^2(J;\mathcal{H})$, with $\mathcal{H}$ being either $L^2(\Omega)$, $H^1(\Omega)$, or $H^{-1}(\Omega)$. In addition, we often write u or $u(t)$ instead of $u(t,x)$ and use C to denote a generic positive constant, not depending on the discretization or regularization parameters.

A weak solution for problem (1.5) is defined as follows.

DEFINITION 1.4. *A function u is called a weak solution for equation* (1.5) *iff $b(u) \in H^1(J;H^{-1}(\Omega))$, $u \in L^2(J;H_0^1(\Omega))$, $u(0) = u_0$ (in H^{-1} sense), and for all $\varphi \in L^2(J;H_0^1(\Omega))$ it holds that*

$$\int_0^T (\partial_t b(u(t)),\varphi(t)) \;+\; (\nabla u(t) + k(b(u(t)))e_z, \nabla\varphi(t))dt = 0. \tag{1.6}$$

Existence, uniqueness, and essential bounds for a weak solution of the above problem are studied in several papers (see, for example, [1], [22], [27] and the references therein). In [1] the following regularity result is obtained:

$$b(u) \in L^\infty(J;L^1(\Omega)), \tag{1.7}$$

$$q := -(\nabla u + k(b(u))e_z) \in L^2(J;(L^2(\Omega))^d). \tag{1.8}$$

Here $b(u)$ models the water content, hence it is natural to assume that, after scaling, it lies between 0 and 1 for almost every $(t,x) \in J \times \Omega$. For the same reason, in (A4) similar restrictions are imposed to the initial data. Such essential estimates can be shown, for example, if b and k do not depend explicitly on x, or if $k(b(u))$ is constant for $u = 0$ and $u = 1$. Moreover, $u \in L^2(J;H_0^1(\Omega))$ yields $b(u) \in L^2(J;H_0^1(\Omega))$ due to the Lipschitz continuity of b. Since $b(u) \in H^1(J;H^{-1}(\Omega))$ we have $b(u) \in C([0,T];L^2(\Omega))$ (see [20, Chapter I]), allowing a simplified mixed variational formulation. Following [2] or [29] we integrate (1.5) in time and obtain, for every $t \in J$,

$$b(u(t)) + \nabla \cdot \int_0^t q\ (s)\, ds = b(u^0) \tag{1.9}$$

in L^2 sense. It follows (see [2] or [25]) that the flux $\vec{q}$ defined in (1.8) satisfies

$$\int_0^t q\, d\tau \in H^1(J;(L^2(\Omega))^d) \cap L^2(J;(H^1(\Omega))^d) =: X. \tag{1.10}$$

2. Equivalent formulations. In this section we give the mixed variational formulations and study the equivalence with the conformal ones in both continuous and time discrete cases.

2.1. The continuous case. Integrated in time, problem (1.5) becomes the following.

Problem 1. Find $u \in L^2(J, H_0^1(\Omega))$ such that $b(u) \in L^\infty(J \times \Omega)$, and for all $t \in J$ and $\phi \in H_0^1(\Omega)$ it holds that

$$(b(u(t)) - b(u^0), \phi) + \int_0^t (\nabla u(s) + k(b(u(s)))e_z, \nabla\phi)ds = 0. \tag{2.1}$$

As mentioned in the previous section, this stronger formulation makes sense since $b(u) \in C(J; L^2(\Omega))$.

A mixed formulation for problem (1.5) reads as follows.

Problem 2. Find $(p, \tilde{q}) \in L^2(J \times \Omega)) \times X$ such that $b(p) \in L^\infty(J \times \Omega)$ and for all $t \in J$ the equations

$$(b(p(t)) - b(p^0), w) + (\nabla \cdot \tilde{q}(t), w) = 0, \tag{2.2}$$

$$(\tilde{q}(t), v) - \int_0^t (p(s), \nabla v)ds + \int_0^t (k(b(p(s)))e_z, v)ds = 0 \tag{2.3}$$

hold for all $w \in L^2(\Omega)$ and $v \in H(\mathrm{div}, \Omega)$, with $p^0 = u^0 \in L^2(\Omega)$.

The two problems are equivalent, as shown in Proposition 2.2. In the proof we use the following lemma [7, p. 91].

LEMMA 2.1. *Let $v \in H(\mathrm{div}, \Omega)$ and $\vec{n}$ denote the outer normal to Γ. Then $v \cdot \vec{n}$ is defined in $H^{-1/2}(\Gamma)$ (in the sense of traces) and Green's formula applies for all $p \in H^1(\Omega)$*

$$\int_\Omega \nabla \cdot v \; p \; dx + \int_\Omega v \cdot \; \nabla p \; dx = \int_\Gamma \vec{n} \cdot v \; p \; ds. \tag{2.4}$$

PROPOSITION 2.2. *$u \in L^2(J, H_0^1(\Omega))$ solves Problem 1 iff $(p, \tilde{q}) \in L^2(J \times \Omega)) \times X$ defined as*

$$(p, \tilde{q}) = \left(u, -\int_0^t (\nabla u(s) + k(b(u(s)))e_z)ds\right) \tag{2.5}$$

solves Problem 2. Moreover, in this case we have $p \in L^2(J, H_0^1(\Omega))$.

Proof. We use some ideas from [18].

"$\Rightarrow$" Let $u \in L^2(J, H_0^1(\Omega))$ be a solution of Problem 1 and $(p, \tilde{q})$ defined in (2.5). By (1.10) we have $(p, \tilde{q}) \in L^2(J, H_0^1(\Omega)) \times X$. Fixing now $t > 0$, for any $v \in H(\mathrm{div}, \Omega)$, using Green's formula we get

$$\begin{aligned}(\tilde{q}(t), v) &= -\int_0^t (\nabla u(s) + k(b(u(s)))e_z, v)ds \\ &= \int_0^t (p(s), \nabla v) - (k(b(p(s)))e_z, v)ds,\end{aligned}$$

so (2.3) is proven.

Next, taking any $\phi \in C_0^\infty(\Omega)$ in (2.1) yields

$$(b(u(t)) - b(u^0), \phi) = -\left(\int_0^t (\nabla u(s) + k(b(u(s)))e_z)ds, \nabla\phi\right)$$
$$= (\tilde{q}(t), \nabla\phi) = -(\nabla \cdot \tilde{q}(t), \phi).$$

However, for any $t > 0$, both $b(u(t)) - b(u^0)$ and $\nabla \cdot \tilde{q}(t)$ lie in $L^2(\Omega)$, so the above relations still hold for $\phi \in L^2(\Omega)$, implying (2.2).

"$\Leftarrow$" Let $(p, \tilde{q}) \in L^2(J \times \Omega) \times X$ solving Problem 2 and set $u = p \in L^2(J \times \Omega)$. Taking $v \in (C_0^\infty(\Omega))^d \subset H(\mathrm{div}, \Omega)$ arbitrary, by differentiating (2.3) we get for almost all $t > 0$

$$(\partial_t \tilde{q}(t), v) + (k(b(p(t)))e_z, v) = (p(t), \nabla \cdot v) = -(\nabla p(t), v), \tag{2.6}$$

so $\nabla p = -\partial_t \tilde{q} - k(b(p))e_z$ in a distributional sense. Since both $\partial_t \tilde{q}$ and $k(b(p))e_z$ are in $L^2(J \times \Omega)$, the same holds for ∇p, so $u = p \in L^2(J, H^1(\Omega))$.

Taking now $v \in H(\mathrm{div}, \Omega)$ in (2.3) gives, for every $t \in J$,

$$-\int_0^t (\nabla p, v) \overset{(2.6)}{=} (\tilde{q}(t), v) + \int_0^t (k(b(p))e_z, v) \overset{(2.3)}{=} \int_0^t (p, \nabla \cdot v).$$

In this way, using (2.4) we get

$$\int_0^t \int_\Gamma pv \cdot \vec{n} ds = \int_0^t (\nabla p, v) + \int_0^t (p, \nabla \cdot v) = 0.$$

Here v was chosen arbitrary, so the trace of p on Γ is zero. Thus $p \in L^2(J, H_0^1(\Omega))$ and the same holds for u.

Moreover, taking any $\phi \in H_0^1(\Omega)$ yields, for all $t > 0$,

$$\big(b(u(t)) - b(u^0), \phi\big) \overset{(2.2)}{=} -(\nabla \cdot \tilde{q}(t), \phi) = (\tilde{q}(t), \nabla\phi)$$
$$\overset{(2.3)}{=} -\int_0^t (\nabla u(s) + k(b(u(s)))e_z, \nabla\phi)ds,$$

so u solves (2.1). ∎

2.2. The semidiscrete case. As mentioned in the introduction, for overcoming difficulties due to degeneracy, we first perturb the original equation to obtain a regular parabolic one. Such a technique has been successfully applied in the analysis of degenerate problems and also allows developing effective numerical schemes (see, e.g., [21]).

In problem (1.5) degeneracy appears due to the vanishing of b'. Therefore we approximate this nonlinearity by b_ϵ, with $\epsilon > 0$ a small perturbation parameter. A possible choice reads as

$$b_\epsilon(u) = b(u) + \epsilon u. \tag{2.7}$$

Obviously, b_ϵ is Lipschitz continuous (with the same Lipschitz constant as b, if ϵ is small enough), strictly increasing and its derivative is bounded from below by ϵ. The regularized problem becomes

$$\begin{aligned} \partial_t b_\epsilon(u) - \nabla \cdot (\nabla u + k(b(u))e_z) &= 0 \quad && \text{in} \quad (0;T] \times \Omega, \\ u &= u^0 \quad && \text{in} \quad \Omega, \\ u &= 0 \quad && \text{on} \quad J \times \Gamma. \end{aligned} \tag{2.8}$$

We let $N > 1$ be an integer giving a time step $\tau = T/N$, with $t_n = n\tau$. The regularized semidiscrete conformal problem reads

Problem 3. Let $n = \overline{1, N}$ and u^{n-1} be given. Find $u^n \in H_0^1(\Omega)$ such that, for all $\phi \in H_0^1(\Omega)$,

$$(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), \phi) + \tau(\nabla u^n + k(b(u^n))e_z, \nabla\phi) = 0. \tag{2.9}$$

However, our final aim is a mixed discretization. The time discrete regularized mixed problem becomes the following.

Problem 4. Let $n = \overline{1, N}$ and p^{n-1} given. Find $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ such that

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), w) + \tau(\nabla \cdot q^n, w) = 0, \tag{2.10}$$

$$(q^n, v) - (p^n, \nabla v) + (k(b(p^n))e_z, v) = 0, \tag{2.11}$$

for all $w \in L^2(\Omega)$, respectively, $v \in H(\mathrm{div}, \Omega)$, with $p^0 = u^0 \in L^2(\Omega)$.

As in the continuous case, the two problems above are equivalent.

PROPOSITION 2.3. *Let $n = \overline{1, N}$ be fixed and assume $u^{n-1} = p^{n-1}$. Then $u^n \in H_0^1(\Omega)$ solves Problem* 3 *iff $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ defined as*

$$(p^n, q^n) = (u^n, -(\nabla u^n + k(b(u^n))e_z)) \tag{2.12}$$

solve Problem 4. *Moreover, we have $p^n \in H_0^1(\Omega)$.*

Proof. "⇒" Let $u^n \in H_0^1(\Omega)$ be a solution of Problem 3 and (p^n, q^n) be defined in (2.12). For all $v \in H(\mathrm{div}, \Omega)$ we have

$$(q^n, v) = -(\nabla u^n + k(b(u^n))e_z, v) = (p^n, \nabla v) - (k(b(p^n))e_z, v),$$

so (p^n, q^n) verify (2.11).

Next, for all $\phi \in C_0^\infty(\Omega)$ (which is dense in $H_0^1(\Omega)$) we get

$$\begin{aligned}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), \phi) &\overset{(2.9)}{=} -\tau(\nabla u^n + k(b(u^n))e_z, \nabla\phi) = \tau(q^n, \nabla\phi)\\ &= -\tau(\nabla \cdot q^n, \phi).\end{aligned}$$

But $b_\epsilon(p^n) - b_\epsilon(p^{n-1}) \in L^2(\Omega)$, so $\nabla \cdot q^n \in L^2(\Omega)$, implying $q^n \in H(\mathrm{div}, \Omega)$ and that (2.10) holds by density arguments.

"⇐" Let $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ be a solution of Problem 4 and $u^n = p^n \in L^2(\Omega)$. For any $v \in (C_0^\infty(\Omega))^d \subset H(\mathrm{div}, \Omega)$ we have

$$\begin{aligned}(q^n, v) &\overset{(2.11)}{=} (p^n, \nabla v) - (k(b(p^n))e_z, v)\\ &= -(\nabla p^n, v) - (k(b(p^n))e_z, v),\end{aligned}$$

implying

$$\nabla p^n + k(b(p^n))e_z = -q^n$$

in distributional sense. Since both q^n and $k(b(p^n))$ are $L^2(\Omega)$ functions it follows that $p^n \in H^1(\Omega)$. As for the continuous case, using Green's formula (2.4), we get actually $u^n = p^n \in H_0^1(\Omega)$.

Finally, (2.9) results by taking any $\phi \in H_0^1(\Omega)$ in (2.10),

$$\begin{aligned}(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), \phi) &= -\tau(\nabla \cdot q^n, \phi) = \tau(q^n, \nabla\phi)\\ &= -\tau((\nabla p^n + k(b(p^n))e_z), \nabla\phi).\end{aligned}$$

∎

As resulting from the equivalencies proven above, stability and error estimates for the time discrete mixed formulation can be obtained by analyzing the Euler implicit scheme applied to Problem 3. This is the underlying idea in the forthcoming section.

3. Stability estimates. In this section we investigate the stability of our numerical approach. We make use of the lemmas below.

LEMMA 3.1. *For any vectors* $a_k, b_k \in \mathbb{R}^q$ $(k = \overline{1,N}, q \geq 1)$ *we have*

$$2\sum_{n=1}^{N} a_n \sum_{k=1}^{n} a_k = \left(\sum_{n=1}^{N} a_n\right)^2 + \sum_{n=1}^{N} (a_n)^2, \tag{3.1}$$

$$2\sum_{n=1}^{N} (a_n - a_{n-1}, a_n) = |a_N|^2 - |a_0|^2 + \sum_{n=1}^{N} |a_n - a_{n-1}|^2, \tag{3.2}$$

$$\sum_{n=1}^{N} (a_n - a_{n-1}, b_n) = a_N b_N - a_0 b_0 - \sum_{n=1}^{N} (b_n - b_{n-1}, a_{n-1}). \tag{3.3}$$

LEMMA 3.2. *Under the assumption* (A1)*, for any real sequence* x^j*,* $j = \overline{1,n}$ *we have*

$$\sum_{j=1}^{n} (b_\epsilon(x^j) - b_\epsilon(x^{j-1}))\ x^j \geq -C|x^0|^2 + \frac{\epsilon}{2}|x^n|^2. \tag{3.4}$$

Proof. Since $b'_\epsilon \geq \epsilon$, one has, for any reals x and y,

$$((b_\epsilon(x) - b_\epsilon(y))x \geq \int_y^x s b'_\epsilon ds \quad \text{and} \quad \int_0^x s b'_\epsilon(s) ds \geq \frac{\epsilon}{2} x^2.$$

Furthermore,

$$\begin{aligned}\sum_{j=1}^{n} (b_\epsilon(x^j) - b_\epsilon(x^{j-1}))x^j &\geq \sum_{j=1}^{n} \int_{x^{j-1}}^{x^j} s b'_\epsilon(s) ds \\ &= \int_0^{x^n} s b'_\epsilon(s) ds - \int_0^{x^0} s b'_\epsilon(s) ds \geq -C|x^0|^2 + \frac{\epsilon}{2}|x^n|^2,\end{aligned}$$

where the constant C is half of the Lipschitz constant of b. ∎

3.1. Stability in the time discrete conformal case.

PROPOSITION 3.3. *Assume* (A1)–(A4)*. If* u^n *solves Problem* 3 *(*$n = \overline{1,N}$*), we have*

$$\tau \sum_{n=1}^{N} \|u^n\|_1^2 \leq C. \tag{3.5}$$

Proof. Taking $\phi = u^n$ in (2.9) and summing up for $n = \overline{1,N}$ give

$$\sum_{n=1}^{N} \left(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), u^n\right) + \sum_{n=1}^{N} \tau \|\nabla u^n\|^2 + \sum_{n=1}^{N} \tau \left(k(b(u^n))e_z, \nabla u^n\right) = 0. \tag{3.6}$$

Now we estimate the terms on the left in the above. By (3.4), since $u^0 \in L^2(\Omega)$,

$$\sum_{n=1}^{N} \left(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), u^n\right) \geq -C.$$

The second term needs no further treatment. Finally, since k is bounded, applying the Cauchy–Schwarz inequality, we get

$$\tau \sum_{n=1}^{N} |(k(b(u^n))e_z, \nabla u^n)| \leq \frac{\tau}{2} \sum_{n=1}^{N} \|k(b(u^n))e_z\|^2 + \frac{\tau}{2} \sum_{n=1}^{N} \|\nabla u^n\|^2$$

$$\leq C + \frac{\tau}{2} \sum_{n=1}^{N} \|\nabla u^n\|^2 .$$

Inserting the last inequalities into (3.6) and using the inequality of Poincaré gives (3.5). ∎

3.2. Stability for the time discrete mixed formulation. By the equivalence of Problems 3 and 4, Proposition 3.3 provides stability for the time discrete solutions p^n and q^n.

PROPOSITION 3.4. *Assuming* (A1)–(A4)*, if, for any* $n = \overline{1, N}$*,* (p^n, q^n) *solve Problem* 4*, we have*

$$\tau \sum_{n=1}^{N} \|p^n\|_1^2 + \tau \sum_{n=1}^{N} \|q^n\|^2 \leq C. \tag{3.7}$$

Proof. The estimate for p^n is a direct consequence of (3.5). Next, taking $w = p^n$ in (2.10) and $v = \tau q^n$ in (2.11) yields

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n) + \tau(\nabla \cdot q^n, p^n) = 0,$$
$$(q^n, \tau q^n) - (p^n, \tau \nabla \cdot q^n) + (k(b(p^n))e_z, \tau q^n) = 0.$$

Adding these two equations and summing up for $n = 1$ to N give

$$\sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n) + \tau \sum_{n=1}^{N} \|q^n\|^2 + \tau \sum_{n=1}^{N} (k(b(p^n))e_z, q^n) = 0,$$

and the rest of the proof follows exactly as in the previous proposition. ∎

Other stability estimates can be obtained defining an initial flux $q^0 \in [L^2(\Omega)]^d$. In doing so we take $\rho \in C_0^\infty(B_d(0,1))$ ($B_d(0,1)$ being the unit ball in $\mathbb{R}^d$) so that $\int_{B_d(0,1)} \rho(x)dx = 1$ and consider the mollifier sequence $\{\rho_\mu(x) = \frac{1}{\mu^d}\rho(\frac{x}{\mu})\}_{1>>\mu>0}$. Defining q^0 as

$$q^0 = -\nabla(\rho_\mu * p^0) - k(b(p^0))e_z, \tag{3.8}$$

with μ to be chosen further and $*$ denoting the convolution operator, for any $v \in H(\mathrm{div}, \Omega)$ we have

$$(q^0, v) - (\rho_\mu * p^0, \nabla \cdot v) + (k(b(p^0))e_z, v) = 0. \tag{3.9}$$

A mollifying of p^0 in the above is necessary for having $q^0 \in [L^2(\Omega)]^d$. However, since $p^0 \in L^2(\Omega)$, $\|p^0 - \rho_\mu * p^0\|$ goes to 0 as $\mu \searrow 0$, so $\|q^0\|$ is uniformly bounded with respect to μ. Now the following estimates can be obtained.

PROPOSITION 3.5. *Assuming* (A1)–(A4)*, if, for all* $n = \overline{1, N}$*,* (p^n, q^n) *solve Problem* 4*, for any* $k > 0$ *we have*

$$\sum_{n=1}^{k} (b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\|q^k\|^2 + \tau \sum_{n=1}^{k} \|q^n - q^{n-1}\|^2 \leq C\tau. \tag{3.10}$$

Proof. First we take $w = p^n - p^{n-1} \in L^2(\Omega)$ in (2.10) and subtract equation (2.11) at time step $n-1$ from the one at time step n. Testing with $v = \tau q^n$ in the resulting equality yields

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau(\nabla \cdot q^n, p^n - p^{n-1}) = 0,$$
$$\tau(q^n - q^{n-1}, q^n) - \tau(p^n - p^{n-1}, \nabla \cdot q^n) + \tau((k(b(p^n)) - k(b(p^{n-1})))e_z, q^n) = 0.$$

For $n = 1$ the second equation above reads as

$$\tau(q^1 - q^0, q^1) - \tau(p^1 - p^0, \nabla \cdot q^1) + \tau((k(b(p^1)) - k(b(p^0)))e_z, q^1) = \tau(p^0 - \rho_\mu * p^0, \nabla \cdot q^1).$$

Adding the above pairs of equalities and summing the result up for $n = \overline{1,k}$ yields

$$\begin{aligned} &\sum_{n=1}^{k}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\sum_{n=1}^{k}(q^n - q^{n-1}, q^n) \\ &\quad + \tau\sum_{n=1}^{k}((k(b(p^n)) - k(b(p^{n-1})))e_z, q^n) = \tau(p^0 - \rho_\mu * p^0, \nabla \cdot q^1). \end{aligned} \tag{3.11}$$

Denoting the terms above by $T_1, \dots, T_4$, we first notice that T_1 is positive by the monotonicity of b_ϵ. Next, by (3.2),

$$\begin{aligned} T_2 &= \tau\sum_{n=1}^{k}(q^n - q^{n-1}, q^n) \\ &= \frac{\tau}{2}\|q^k\|^2 - \frac{\tau}{2}\|q^0\|^2 + \frac{\tau}{2}\sum_{n=1}^{k}\|q^n - q^{n-1}\|^2. \end{aligned}$$

Recalling (A3) and the Cauchy–Schwarz inequality, for T_3 we get

$$\begin{aligned} |T_3| &\le \frac{\delta_1}{2}\sum_{n=1}^{k}\|(k(b(p^n)) - k(b(p^{n-1})))e_z\|^2 + \frac{\tau^2}{2\delta_1}\sum_{n=1}^{k}\|q^n\|^2 \\ &\le \frac{\delta_1 C_k}{2}\sum_{n=1}^{k}(b(p^n) - b(p^{n-1}), p^n - p^{n-1}) + \frac{\tau^2}{2\delta_1}\sum_{n=1}^{k}\|q^n\|^2. \end{aligned}$$

Estimating T_4 follows as before,

$$|T_4| \le \tau\|p^0 - \rho_\mu * p^0\|\|\nabla \cdot q^1\| \le \delta_2\|p^0 - \rho_\mu * p^0\|^2 + \frac{\tau^2}{4\delta_2}\|\nabla \cdot q^1\|^2.$$

To estimate $\|\nabla \cdot q^1\|$ we use (2.10) for $n = 1$, test with $w = \nabla \cdot q^1 \in L^2(\Omega)$ and obtain

$$\tau\|\nabla \cdot q^1\|^2 \le \|b_\epsilon(p^1) - b_\epsilon(p^0)\|\|\nabla \cdot q^1\| \le \frac{C}{2\tau}\left(b_\epsilon(p^1) - b_\epsilon(p^0), p^1 - p^0\right) + \frac{\tau}{2}\|\nabla \cdot q^1\|^2$$

by the Lipschitz continuity of b_ϵ. In this way we get

$$\tau\|\nabla \cdot q^1\|^2 \le \tfrac{C}{\tau}(b_\epsilon(p^1) - b_\epsilon(p^0), p^1 - p^0).$$

Using these estimates in (3.11) and choosing the δ's properly give

$$\begin{aligned} &\sum_{n=1}^{k}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\|q^k\|^2 + \tau\sum_{n=1}^{k}\|q^n - q^{n-1}\|^2 \\ &\quad \le C_1\tau + C_2\|p^0 - \rho_\mu * p^0\|^2 + C_3\tau^2\sum_{n=1}^{k}\|q^n\|^2. \end{aligned}$$

We still have to choose μ in (3.8). Since $\|p^0 - \rho_\mu * p^0\|$ converges to 0, taking μ sufficiently small, the right term in the above becomes

$$C_4\tau + C_3\tau^2 \sum_{n=1}^{k} \|q^n\|^2.$$

Now (3.10) follows by the discrete Gronwall lemma. ∎

REMARK 3.6. *If $p^0 \in H^1(\Omega)$, q^0 can be defined without using a mollifier,*

$$q^0 = -\nabla p^0 - k(b(p^0))e_z. \tag{3.12}$$

Then $T_4 = 0$ in (3.11)*, without changing* (3.10).

A direct consequence of the stability estimates above follows.

PROPOSITION 3.7. *In the setting of Proposition* 3.5 *we have*

$$\sum_{n=1}^{N} \tau\|\nabla \cdot q^n\|^2 \le C. \tag{3.13}$$

Proof. Taking $w = \nabla \cdot q^j$ in equation (2.10) and applying the Cauchy–Schwarz inequality one gets

$$\tau\|\nabla \cdot q^j\|^2 \le \frac{1}{2\tau}\|b_\epsilon(p^j) - b_\epsilon(p^{j-1})\|^2 + \frac{\tau}{2}\|\nabla \cdot q^j\|^2,$$

so

$$\tau\|\nabla \cdot q^j\|^2 \le \frac{1}{\tau}\|b_\epsilon(p^j) - b_\epsilon(p^{j-1})\|^2.$$

Summing up the above for $j = \overline{1,N}$, using the Lipschitz continuity of b_ϵ and (3.10) leads to (3.13). ∎

4. Error estimates. In this section we obtain a priori error estimates for both time discrete scheme, as well as for the fully discrete one.

4.1. Error estimates for the semidiscrete approximation. To obtain error estimates for the time discrete scheme we employ techniques developed in [21] and make use of the Green operator $G : H^{-1}(\Omega) \to H_0^1(\Omega)$ defined as

$$(\nabla(G\psi), \nabla\phi) = (\psi, \phi) \quad \text{for all} \quad \phi \in H_0^1(\Omega). \tag{4.1}$$

Obviously, G is linear and self-adjoint. Moreover, by the Cauchy–Schwarz inequality, using (3.3) yields the following lemma.

LEMMA 4.1. *For all $f, f_k \in H^{-1}(\Omega)$ $(k = \overline{1,N})$ and $g \in H^1(\Omega)$ we have*

$$(f, g) \le \|f\|_{-1}\|\nabla g\|,$$

$$\|\nabla Gf\|^2 = (f, Gf) = \|f\|_{-1}^2,$$

$$2\sum_{k=1}^{N}(f_k - f_{k-1}, Gf_k) = \|f_k\|_{-1,\Omega}^2 - \|f_0\|_{-1,\Omega}^2 + \sum_{k=1}^{N}\|f_k - f_{k-1}\|_{-1,\Omega}^2.$$

Further, we use the notations

$$\begin{aligned} \overline{u}^n &= \frac{1}{\tau}\int_{t_{n-1}}^{t_n} u(t)dt, \\ u_\Delta(t) &= u^n \quad \text{for} \quad t \in (t_{n-1}, t_n], \\ e_b(u) &= b(u) - b_\epsilon(u_\Delta), \end{aligned} \tag{4.2}$$

where $n = \overline{1, N}$ and $\overline{u}^0 = u^0$.

It is worth pointing out here that, by Propositions 2.2 and 2.3, estimates obtained for the conformal discretization can be transferred to the mixed case.

PROPOSITION 4.2. *Assuming* (A1)–(A4)*, if* u *is the weak solution of Problem* 1 *and* u^n *solves, for each* $n = \overline{1, N}$*, Problem* 3*, then*

$$\begin{aligned} \max_{n=\overline{1,N}} \left\| \overline{e_b(u)}^n \right\|_{-1}^2 + \|e_b(u)\|_{L^2(J\times\Omega)}^2 + \int_0^T (b_\epsilon(u(t)) - b_\epsilon(u_\Delta), u(t) - u_\Delta)dt \\ \le C\,(\tau + \epsilon). \end{aligned} \tag{4.3}$$

Proof. Subtracting (2.1) at $t = t_{j-1}$ from the one at $t = t_j$ and then subtracting (2.9) with $n = j$ from the result give

$$\begin{aligned} &(b(u(t_j)) - b(u(t_{j-1})) - b_\epsilon(u^j) + b_\epsilon(u^{j-1}), \phi) \\ &\quad + \tau(\nabla(\overline{u}^j - u^j), \nabla\phi) + \tau((\overline{k(b(u))}^j - k(b(u^j)))e_z, \nabla\phi) = 0. \end{aligned}$$

Taking $\phi = G\overline{e_b(u)}^j \in H_0^1(\Omega)$ into above and summing up for $j = \overline{1, n}$ (with $n \le N$) yield

$$\begin{aligned} &\sum_{j=1}^n (b(u(t_j)) - b(u(t_{j-1})) - b_\epsilon(u^j) + b_\epsilon(u^{j-1}), G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n \tau(\nabla\overline{u}^j - \nabla u^j, \nabla G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n \tau((\overline{k(b(u))}^j - k(b(u^j)))e_z, \nabla G\overline{e_b(u)}^j) = 0. \end{aligned} \tag{4.4}$$

We estimate now each of terms in (4.4), denoted by T_1, T_2, and T_3:

$$\begin{aligned} T_1 &= \sum_{j=1}^n (b(u(t_j)) - \overline{b(u)}^j - b(u(t_{j-1})) + \overline{b(u)}^{j-1}, G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n (\overline{b(u)}^j - b_\epsilon(u^j) - \overline{b(u)}^{j-1} + b_\epsilon(u^{j-1}), G\overline{e_b(u)}^j) \\ &=: T_{11} + T_{12}. \end{aligned}$$

Further, by (3.3) and recalling that $b(u(0)) = \overline{b(u)}^0$ we have

$$\begin{aligned} T_{11} &= \sum_{j=1}^n (b(u(t_j)) - \overline{b(u)}^j - b(u(t_{j-1})) + \overline{b(u)}^{j-1}, G\overline{e_b(u)}^j) \\ &= (b(u(t_n)) - \overline{b(u)}^n, G\overline{e_b(u)}^n) \end{aligned}$$

$$
\begin{aligned}
&-\sum_{j=1}^{n}(b(u(t_{j-1}))-\overline{b(u)}^{j-1},G\overline{e_b(u)}^{j}-G\overline{e_b(u)}^{j-1})\\
&=: T_{111}-T_{112}.
\end{aligned}
$$

For T_{111} we make use of Lemma 4.1 and obtain

$$
\begin{aligned}
|T_{111}| &\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}|(b(u(t_n))-b(u(t)),G\overline{e_b(u)}^{n})|dt\\
&\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}\int_{t}^{t_n}|(\partial_s b(u(s)),G\overline{e_b(u)}^{n})|dsdt\\
&\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}\sqrt{\tau}\,\|\partial_t b(u)\|_{L^2(t_{n-1},t_n;H^{-1})}\,\|\overline{e_b(u)}^{n}\|_{-1}dt \qquad (4.5)\\
&\le \sqrt{\tau}\,\|\partial_s b(u)\|_{L^2(t_{n-1},t_n;H^{-1})}\,\|\overline{e_b(u)}^{n}\|_{-1}\\
&\le \tau\,\|\partial_s b(u)\|^2_{L^2(t_{n-1},t_n;H^{-1})}+\frac{1}{4}\|\overline{e_b(u)}^{n}\|^2_{-1}.
\end{aligned}
$$

Proceeding as before, T_{112} can be estimated as

$$
|T_{112}|\le \tau\,\|\partial_t b(u)\|^2_{L^2(0,t_n;H^{-1})}+\frac{1}{4}\sum_{j=1}^{n}\|\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1}\|^2_{-1}. \qquad (4.6)
$$

Using Lemma 4.1 again, since $\overline{e_b(u)}^{0}=0$, T_{12} gives

$$
\begin{aligned}
T_{12} &= \sum_{j=1}^{n}(\overline{b(u)}^{j}-b_\epsilon(u^j)-\overline{b(u)}^{j-1}+b_\epsilon(u^{j-1}),G\overline{e_b(u)}^{j}))\\
&= \frac{1}{2}(\overline{e_b(u)}^{n},G\overline{e_b(u)}^{n}) \qquad (4.7)\\
&\quad+\frac{1}{2}\sum_{j=1}^{n}(\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1},G\overline{e_b(u)}^{j}-G\overline{e_b(u)}^{j-1})\\
&= \frac{1}{2}\|\overline{e_b(u)}^{n}\|^2_{-1}+\frac{1}{2}\sum_{j=1}^{n}\|\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1}\|^2_{-1}.
\end{aligned}
$$

For T_2 we have

$$
\begin{aligned}
T_2 &= \sum_{j=1}^{n}\tau(\nabla\overline{u}^{j}-\nabla u^j,\nabla G\overline{e_b(u)}^{j})\\
&= \tau\sum_{j=1}^{n}\left(\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(u(t)-u^j)dt,\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(b(u(s))-b_\epsilon(u^j))ds\right)\\
&= \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\big(u(t)-u^j,b(u(t))-b_\epsilon(u^j)\big)\,dt\\
&\quad+\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(u(t)-u^j,\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(b(u(s))-b(u(t)))ds\right)dt\\
&=: T_{21}+T_{22}.
\end{aligned}
$$

T_{21} can be decomposed as follows:

$$T_{21} = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b(u(t)) - b_\epsilon(u(t))\big)\, dt$$
$$+ \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b_\epsilon(u(t)) - b_\epsilon(u^j)\big)\, dt \;=: T_{211} + T_{212}.$$

The definition of b_ϵ in (2.7) gives

$$|T_{211}| = \left| \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, \epsilon\, u(t)\big)\, dt \right|$$
$$\leq \epsilon \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|\, \|u(t)\|\, dt \tag{4.8}$$
$$\leq \frac{\epsilon}{4} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|^2\, dt + \epsilon\, \|u\|^2_{L^2(0,t_n;L^2(\Omega))}.$$

Since b_ϵ is monotone, T_{212} is positive; moreover, it holds

$$T_{212} \geq \frac{1}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b_\epsilon(u(t)) - b_\epsilon(u^j)\big)\, dt$$
$$+ \frac{\epsilon}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|^2\, dt. \tag{4.9}$$

Proceeding as for T_{111} and recalling the a priori estimates in Proposition 3.3, since $b(u) \in H^1(J; H^{-1})$ and $u \in L^2(J; H^1)$ we obtain

$$T_{22} = \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b(u(s)) - b(u(t))\big)\, ds dt$$
$$= \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \left(\int_t^s (u(t) - u^j, \partial_r b(u)) dr \right) ds dt$$
$$\leq \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \int_t^s \|\nabla(u(t) - u^j)\| \|\partial_r b(u)\|_{-1} dr ds dt \tag{4.10}$$
$$\leq \frac{\tau}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|\nabla(u(t) - u^j)\big\|^2 + \frac{\tau}{2} \|\partial_r b(u)\|^2_{L^2(0,t_n;H^{-1})}$$
$$\leq C\, \tau.$$

For the T_3 we proceed as follows:

$$|T_3| \leq \frac{\tau}{4\delta} \sum_{j=1}^{n} \|\overline{k(b(u))}^j - k(b(u^j))\|^2 + \delta\tau \sum_{j=1}^{n} \|\overline{e_b(u)}^j\|^2_{-1}$$
$$= T_{31} + \delta\tau \sum_{j=1}^{n} \|\overline{e_b(u)}^j\|^2_{-1}.$$

Applying (A3) and taking $\delta = C_k$ gives

$$
\begin{aligned}
|T_{31}| &= \frac{\tau}{4\delta}\frac{1}{\tau^2}\sum_{j=1}^{n}\int_\Omega\left(\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))dt\right)^2 dx\\
&\le \frac{1}{4\delta\tau}\sum_{j=1}^{n}\int_\Omega \tau\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))^2 dtdx\\
&\le \frac{1}{4}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(b(u)-b(u^j),u-u^j\right)dt \qquad (4.11)\\
&\le \frac{1}{4}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(b_\epsilon(u)-b_\epsilon(u^j),u-u^j\right)dt.
\end{aligned}
$$

Since $b(u)\in H^1(J;H^{-1})$ and $u\in L^2(J;H^1(\Omega))$, inserting (4.5)–(4.11) into (4.4) yields

$$
\begin{aligned}
&\|\overline{e_b(u)}^n\|_{-1}^2+\sum_{j=1}^{n}\|\overline{e_b(u)}^j-\overline{e_b(u)}^{j-1}\|_{-1}^2+\epsilon\int_{t_{j-1}}^{t_j}\left\|u(t)-u^j\right\|^2dt\\
&\quad+\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(u(t)-u^j,b_\epsilon(u(t))-b_\epsilon(u^j)\right)dt\\
&\quad\le C(\tau+\epsilon)+4C_k\tau\sum_{j=1}^{n}\|\overline{e_b(u)}^j\|_{-1}^2,
\end{aligned}
$$

and (4.3) is a direct consequence of the discrete Gronwall lemma. □

Using the above result an error estimate for the L^2 norm of the time integrated gradient can be obtained. Such an estimate is essential for our analysis because it provides also an error estimate for the time integral of the flux in the mixed formulation.

PROPOSITION 4.3. *Under the assumptions in Proposition* 4.2 *we have*

$$
\left\|\int_0^T(u(t)-u_\Delta(t))dt\right\|_1^2\le C\,(\tau+\epsilon). \qquad (4.12)
$$

Proof. Following the ideas in [21], we first add (2.9) for $n=1$ to N, subtract the result from (2.1) at $t=t_N=T$ and end up with

$$
\begin{aligned}
&\left(b(u(T))-b_\epsilon(u^N),\phi\right)-\left(b(u(t_0))-b_\epsilon(u^0),\phi\right)\\
&\quad+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}\nabla(u(t)-u^j)dt,\nabla\phi\right)\\
&\quad+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z dt,\nabla\phi\right)=0
\end{aligned}
$$

for all $\phi\in H_0^1(\Omega)$. Now taking $\phi=\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)$ into the total above gives

$$
\begin{aligned}
&\left(b(u(T))-b_\epsilon(u^N),\tau\sum_{j=1}^{N}\left(\overline{u}^j-u^j\right)\right)\\
&-\left(\epsilon u^0,\tau\sum_{j=1}^{N}\left(\overline{u}^j-u^j\right)\right)+\left\|\tau\sum_{j=1}^{N}\nabla\left(\overline{u}^j-u^j\right)\right\|^2\\
&+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z,\nabla\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)=0.
\end{aligned}
\tag{4.13}
$$

Denoting the terms in (4.13) by T_1, T_2, T_3, and T_4, we proceed by estimating each of them separately. T_1 yields

$$
T_1=\left(b(u(T))-\overline{b(u)}^N+\overline{b(u)}^N-b_\epsilon(u^N),\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)=:T_{11}+T_{12}.
$$

As in (4.5), since $\partial_t b(u)\in L^2(J;H^{-1})$, T_{11} gives

$$
\begin{aligned}
|T_{11}|&\le\frac{1}{\tau}\int_{t_{N-1}}^{t_N}\int_{t}^{t_N}\left|\left(\partial_s b(u),\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)\right|dsdt\\
&\le\frac{C_{11}}{2\delta_{11}}\tau+\frac{\delta_{11}}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\end{aligned}
\tag{4.14}
$$

Applying the Cauchy–Schwarz inequality, for T_{12} we obtain

$$
|T_{12}|\le\frac{1}{2\delta_{12}}\left\|\overline{e_b(u)}^N\right\|_{-1}^2+\frac{\delta_{12}}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.15}
$$

Analogously, T_2 gives

$$
|T_2|\le\frac{1}{2\delta_2}\epsilon\left\|u^0\right\|^2+\frac{\delta_2}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.16}
$$

For T_3 we recall the inequality of Poincaré:

$$
T_3=\left\|\sum_{j=1}^{N}\tau\nabla\left(\overline{u}^j-u^j\right)\right\|^2\ge C\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.17}
$$

Analogously, T_4 can be estimated as

$$
|T_4|\le\frac{1}{2\delta_4}\left\|\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z dt\right\|^2+\frac{\delta_4}{2}\left\|\nabla\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|^2.
$$

For the first term above—denoted by T_{41}—we get, by (A3),

$$
\begin{aligned}
T_{41}&\le N\sum_{j=1}^{N}\tau\int_{t_{j-1}}^{t_j}\|k(b(u))-k(b(u^j))\|^2dt\\
&\le TC_k\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}\left(b(u(t))-b(u^j),u(t)-u^j\right).
\end{aligned}
\tag{4.18}
$$

Inserting (4.14)–(4.18) into (4.13), choosing the $\delta's$ properly and recalling the estimates in Proposition 4.2 we obtain (4.12). ∎

Propositions 4.2 and 4.3 can be summarized in the following.

THEOREM 4.4. *If u is the solution of Problem 1 and u^n solves Problem 3 ($n = \overline{1,N}$), we have*

$$\max_{n=\overline{1,N}} \|\overline{e_b(u)}^n\|_{-1}^2 + \|e_b(u)\|_{L^2(J\times\Omega)}^2 + \left\|\int_0^T (u(t)-u_\Delta(t))dt\right\|_1^2 + \int_0^T (b_\epsilon(u(t))-b_\epsilon(u_\Delta(t)), u(t)-u_\Delta(t))dt \le C(\tau+\epsilon). \tag{4.19}$$

REMARK 4.5. *The estimates above do not change if we replace the last term on the left by $\int_0^T (b(u(t))-b(u_\Delta(t)), u(t)-u_\Delta(t))dt$.*

Since Problems 3 and 4 are equivalent we immediately obtain the following theorem.

THEOREM 4.6. *In the setting of Theorem 4.4, if (p^n, q^n) solve Problem 4 ($n = \overline{1,N}$), we get*

$$\begin{aligned}&\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (b_\epsilon(u(t))-b_\epsilon(p^n), u(t)-p^n)dt \\ &+ \left\|\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (u(t)-p^n)dt\right\|_1^2 + \left\|\tilde{q}(T) - \tau\sum_{n=1}^N q^n\right\|^2 \\ &\le C(\tau+\epsilon).\end{aligned} \tag{4.20}$$

REMARK 4.7. *As in Remark 4.5, we can replace the scalar product in (4.20) by $\int_0^T (b(u(t))-b(u_\Delta(t)), u(t)-u_\Delta(t))dt$. This immediately implies an error estimate for the saturation,*

$$\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \|b(u(t))-b(p^n)\|^2 dt \le C(\tau+\epsilon).$$

4.2. Error estimates for the fully discrete mixed discretization. The next step in our analysis is proving error estimates for the fully discrete approximation. We first estimate the error for the flux variable and then proceed with estimates for the p unknowns.

In doing so we denote by W and V the spaces $L^2(\Omega)$ and $H(\mathrm{div},\Omega)$. Let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size (see [9]). Here we assume $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$, hence Ω is polygonal. Thus we neglect the errors caused by an approximation of a nonpolygonal domain, avoiding an excess of technicalities (a complete analysis in this sense can be found in [21]).

The discrete subspaces $W_h \times V_h \subset W \times V$ are defined as

$$\begin{aligned}W_h &:= \{p \in W |\ p \text{ is constant on each element } T \in \mathcal{T}_h\}, \\ V_h &:= \{\overrightarrow{q} \in V |\ \overrightarrow{q}_{|T} = \overrightarrow{a} + b\overrightarrow{x} \ \text{ for all } T \in \mathcal{T}_h\}.\end{aligned} \tag{4.21}$$

So W_h denotes the space of piecewise constant functions, while V_h is the RT_0 space (see [7]). Further we make use of the usual L^2 projector

$$P_h : L^2(\Omega) \to W_h, \quad ((P_h w - w), w_h) = 0 \qquad \forall w_h \in W_h. \tag{4.22}$$

Taking a $\widetilde{V}$ slightly better than V (for example, $V \cap (L^s(\Omega))^d$ with an $s > 2$), a projector Π_h can be defined as (see [7, p. 131])

$$\Pi_h : \widetilde{V} \to V_h, \quad (\nabla \cdot (\Pi_h v - v), w_h) = 0 \tag{4.23}$$

for all $w_h \in W_h$. With $r \in (0, 1]$, for the operators defined above we have

$$\begin{aligned} \|w - P_h w\| &\le C h^r \|w\|_r, \\ \|v - \Pi_h v\| &\le C h^r \|v\|_r \end{aligned} \tag{4.24}$$

for any $w \in H^r(\Omega)$ and $v \in (H^r(\Omega))^d$.

The following technical lemma is proven in [25].

LEMMA 4.8. *Assuming* (A1)*, taking $f_h \in W_h$, a $v_h \in V_h$ exists so that*

$$\begin{aligned} \nabla \cdot v_h &= f_h, \\ \|v_h\| &\le C \|\nabla \cdot v_h\|, \end{aligned}$$

$C > 0$ being a generic constant not depending on h, f_h, or v_h.

Before proceeding with the fully discrete approximation scheme, we rewrite Problem 4 (continuous in space) as

Problem 5. Let $n = \overline{1, N}$. Find $(p^n, q^n) \in W \times V$ such that

$$(b_\epsilon(p^n), w) - (b_\epsilon(p^0), w) + \tau \left(\sum_{j=1}^{n} \nabla \cdot q^j, w \right) = 0, \tag{4.25}$$

$$(q^n, v) - (p^n, \nabla \cdot v) + (k(b(p^n))e_z, v) = 0 \tag{4.26}$$

for all $w \in W$ and $v \in V$, with $p^0 = u^0$.

The fully discrete mixed finite element approximation reads the following.

Problem 6. Let $n = \overline{1, N}$. Find $(p_h^n, q_h^n) \in W_h \times V_h$ such that

$$(b_\epsilon(p_h^n), w_h) + \tau \left(\sum_{j=1}^{n} \nabla \cdot q_h^j, w_h \right) = (b_\epsilon(p_h^0), w_h), \tag{4.27}$$

$$(q_h^n, v_h) - (p_h^n, \nabla \cdot v_h) + (k(b(p_h^n))e_z, v_h) = 0 \tag{4.28}$$

for all $w_h \in W_h$ and $v_h \in V_h$.

Initially we take $p_h^0 = b_\epsilon^{-1}(P_h b_\epsilon(u^0))$. Since $P_h b_\epsilon(u^0)$ is constant on any $T \in \mathcal{T}_h$, the same holds for $b_\epsilon^{-1}(P_h b_\epsilon(u^0))$, so $p_h^0 \in W_h$. Moreover, with this choice, for all $w_h \in W_h$, we obtain

$$(b_\epsilon(p_h^0), w_h) = (b_\epsilon(u^0), w_h) = (b_\epsilon(p^0), w_h).$$

We start with some stability estimates for the fully discrete case.

PROPOSITION 4.9. *Assuming* (A1)–(A4)*, if (p_h^n, q_h^n) solve Problem* 6 *($n = \overline{1, N}$), we have*

$$\begin{aligned} \|p_h^n\|^2 + \|q_h^n\|^2 &\le C, \\ \sum_{k=1}^{n} (b_\epsilon(p_h^k) - b_\epsilon(p_h^{k-1}), p_h^k - p_h^{k-1}) &\le C\tau. \end{aligned} \tag{4.29}$$

Proof. Applying the arguments used in Propositions 3.4 and 3.7 we immediately obtain the estimates (4.29), excepting the one for $\|p_h^n\|$. To complete the proof we apply Lemma 4.8. Consequently, there exists a $v_h \in V_h$ such that $\nabla \cdot v_h = p_h^n$ and $\|v_h\| \leq C\|p_h^n\|$. Using this as a test function in (4.28) gives

$$\|p_h^n\|^2 = (q_h^n, v_h) + (k(b(p_h^n))e_z, v_h) \leq C(\|q_h^n\| + \|k(b(p_h^n))\|)\|p_h^n\|,$$

and the estimate follows from the estimates for $\|q_h^n\|$ and the boundedness of k. ∎

Applying now techniques developed in [2] we estimate the errors induced by the spatial discretization.

PROPOSITION 4.10. *Let $n = \overline{1,N}$. If $(p^n, q^n) \in W \times V$, $(p_h^n, q_h^n) \in W_h \times V_h$ solve Problem 5, respectively 6, assuming* (A1)–(A4) *yields*

$$\sum_{n=1}^{N} \left\{ (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n) + \tau \|\Pi_h q^n - q_h^n\|^2 \right\} + \tau \left\| \sum_{n=1}^{N} (\Pi_h q^n - q_h^n) \right\|^2$$

$$\leq C \sum_{n=1}^{N} \{ \|q^n - \Pi_h q^n\|^2 + \|P_h p^n - p^n\|^2 \}. \tag{4.30}$$

Proof. Subtracting (4.27) from (4.25) and (4.28) from (4.26) gives

$$(b_\epsilon(p^n) - b_\epsilon(p_h^n), w_h) + \tau \left(\sum_{j=1}^{n} \nabla \cdot (q^j - q_h^j), w_h \right) = 0,$$

$$(q^n - q_h^n, v_h) - (p^n - p_h^n, \nabla \cdot v_h) + ((k(b(p^n)) - k(b(p_h^n)))e_z, v_h) = 0.$$

Taking $w_h = P_h p^n - p_h^n \in W_h$ and $v_h = \tau \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \in V_h$ into the above leads to

$$(b_\epsilon(p^n) - b_\epsilon(p_h^n), P_h p^n - p_h^n) + \tau \left(\sum_{j=1}^{n} \nabla \cdot (\Pi_h q^j - q_h^j), P_h p^n - p_h^n \right) = 0,$$

$$\tau \left(q^n - q_h^n, \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right) - \tau \left(P_h p^n - p_h^n, \nabla \cdot \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right)$$

$$+ \tau \left((k(b(p^n)) - k(b(p_h^n)))e_z, \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right) = 0.$$

Adding these equalities and summing the result up from 1 to N yields

$$\sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), P_h p^n - p_h^n) + \sum_{n=1}^{N} \left(q^n - q_h^n, \sum_{j=1}^{n} \tau (\Pi_h q^j - q_h^j) \right)$$

$$+ \sum_{n=1}^{N} \left((k(b(p^n)) - k(b(p_h^n)))e_z, \sum_{j=1}^{n} \tau (\Pi_h q^j - q_h^j) \right) = 0. \tag{4.31}$$

We estimate now each of the terms above, denoted by T_1, T_2, and T_3:

$$T_1 = \sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n)$$

$$+\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),P_hp^n-p^n)=:T_{11}+T_{12}. \tag{4.32}$$

By (A2) and the definition (2.7) of b_ϵ, a $C>0$ independent of τ and h exists such that

$$\begin{aligned}T_{11}&\geq\frac{1}{2}\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)\\&\quad+C\left(\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\epsilon\sum_{n=1}^{N}\|p^n-p_h^n\|^2\right).\end{aligned} \tag{4.33}$$

Applying the inequality of Cauchy T_{12} yields

$$|T_{12}|\leq\frac{\mu}{2}\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\frac{1}{2\mu}\sum_{n=1}^{N}\|P_hp^n-p^n\|^2. \tag{4.34}$$

Rewriting T_2 as

$$\begin{aligned}T_2&=\sum_{n=1}^{N}\left(q^n-\Pi_hq^n,\sum_{j=1}^{n}\tau(\Pi_hq^j-q_h^j)\right)\\&\quad+\sum_{n=1}^{N}\left(\Pi_hq^n-q_h^n,\sum_{j=1}^{n}\tau(\Pi_hq^j-q_h^j)\right)=:T_{21}+T_{22},\end{aligned} \tag{4.35}$$

we estimate T_{21} and T_{22}. For T_{21} we get

$$|T_{21}|\leq\frac{1}{2}\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2+\frac{\tau^2}{2}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2, \tag{4.36}$$

while for T_{22} we use (3.1) to obtain

$$T_{22}=\frac{\tau}{2}\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+\frac{\tau}{2}\sum_{n=1}^{N}\|\Pi_hq^n-q_h^n\|^2. \tag{4.37}$$

Using (A3), T_3 gets

$$\begin{aligned}|T_3|&\leq\frac{\delta}{2}\sum_{n=1}^{N}\|k(b(p^n))-k(b(p_h^n))\|^2+\frac{\tau^2}{2\delta}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2\\&\leq\frac{C_k\delta}{2}\sum_{n=1}^{N}(b(p^n)-b(p_h^n),p^n-p_h^n)+\frac{\tau^2}{2\delta}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2.\end{aligned} \tag{4.38}$$

Inserting (4.32)–(4.38) into (4.31) and choosing μ and δ properly gives

$$\begin{aligned}&\sum_{n=1}^{N}\{(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\|\Pi_hq^n-q_h^n\|^2\}+\tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2\\&\quad\leq C\sum_{n=1}^{N}\left\{\|q^n-\Pi_hq^n\|^2+\|P_hp^n-p^n\|^2+\tau^2\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2\right\}.\end{aligned}$$

Finally, (4.30) follows applying the discrete Gronwall lemma. ∎

REMARK 4.11. *By the equivalence proven in Proposition* 2.3, $p^n \in H^1(\Omega)$ *for all* n. *Now using* (4.24) *and* (3.7) *we get*

$$\sum_{n=1}^{N} \|P_h p^n - p^n\|^2 \le Ch^2 \sum_{n=1}^{N} \|p^n\|_1^2 \le C\frac{h^2}{\tau},$$

and the estimates (4.30) *can be modified accordingly.*

Similar estimates can be obtained for the p-unknowns.

PROPOSITION 4.12. *Under the assumptions of Proposition* 4.10 *we have*

$$\tau \left\| \sum_{n=1}^{N} (P_h p^n - p_h^n) \right\|^2 \le C \Bigg\{ \sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n) + \tau \left\| \sum_{n=1}^{N} (\Pi_h q^n - q_h^n) \right\|^2 + \sum_{n=1}^{N} \|q^n - \Pi_h q^n\|^2 \Bigg\}. \tag{4.39}$$

Proof. Subtracting (4.28) from (4.26), recalling the definition of P_h, and summing up for $n = 1$ to N yield

$$\left(\sum_{n=1}^{N} (q^n - q_h^n), v_h \right) - \left(\sum_{n=1}^{N} (P_h p^n - p_h^n), \nabla \cdot v_h \right) + \left(\sum_{n=1}^{N} \tau (k(b(p^n)) - k(b(p_h^n)))e_z, v_h \right) = 0 \tag{4.40}$$

for any $v_h \in V_h$. Using now Lemma 4.8, a $v_h \in V_h$ exists such that

$$\nabla \cdot v_h = \sum_{n=1}^{N} \tau (P_h p^n - p_h^n) \tag{4.41}$$

and $\|v_h\| < C\|\tau \sum_{n=1}^{N} (P_h p^n - p_h^n)\|$. In this case (4.40) gives

$$\tau \left\| \sum_{n=1}^{N} (P_h p^n - p_h^n) \right\|^2 = \left(\sum_{n=1}^{N} (q^n - q_h^n), v_h \right) + \left(\sum_{n=1}^{N} (k(b(p^n)) - k(b(p_h^n)))e_z, v_h \right). \tag{4.42}$$

Denoting by T_1 and T_2 the terms on the right into above, applying the inequality of Cauchy and recalling the estimates on $\|v_h\|$ leads to

$$|T_1| \le \frac{\tau}{2\delta_1} \left\| \sum_{n=1}^{N} (q_h^n - q^n) \right\|^2 + \frac{\delta_1}{2\tau} \|v_h\|^2$$

$$\le \frac{\tau}{2\delta_1} \left\| \sum_{n=1}^{N} (q_h^n - q^n) \right\|^2 + \frac{C\tau\delta_1}{2} \left\| \sum_{n=1}^{N} (p_h^n - P_h p_h^n) \right\|^2. \tag{4.43}$$

Similarly, by (A3) we obtain

$$|T_2| \le \frac{\tau}{2\delta_2}\left\|\sum_{n=1}^{N}(k(b(p_h^n))-k(b(p^n)))\right\|^2 + \frac{\delta_2}{2\tau}\|v_h\|^2$$

$$\le \frac{C}{2\delta_2}\sum_{n=1}^{N}(b(p_h^n)-b(p^n),p_h^n-p^n) + \frac{C\tau\delta_2}{2}\left\|\sum_{n=1}^{N}(p_h^n-P_hp_h^n)\right\|^2. \tag{4.44}$$

Choosing δ_1 and δ_2 properly, (4.42)–(4.44) gives

$$\tau\left\|\sum_{n=1}^{N}(P_hp^n-p_h^n)\right\|^2$$

$$\le C\left\{\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2\right\}.$$

The last term above can be rewritten as

$$\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2 \le \tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+\tau\left\|\sum_{n=1}^{N}(q^n-\Pi_hq^n)\right\|^2$$

$$\le \tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+T\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2,$$

which completes the proof. □

The following is a direct consequence of Propositions 4.10 and 4.12.

THEOREM 4.13. *Assuming* (A1)–(A4), *if* $(p^n,q^n)\in W\times V$, $(p_h^n,q_h^n)\in W_h\times V_h$ *solve, for* $n=\overline{1,N}$, *Problems* 5 *and* 6, *we obtain*

$$\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\sum_{n=1}^{N}\|\Pi_hq^n-q_h^n\|^2$$

$$+\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2+\tau\left\|\sum_{n=1}^{N}(p^n-p_h^n)\right\|^2 \tag{4.45}$$

$$\le C\left(\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2+\sum_{n=1}^{N}\|P_hp^n-p^n\|^2\right).$$

Combining the estimates in Theorems 4.6 and 4.13 and recalling Remark 4.11 we get, for the fully discrete scheme, the following.

THEOREM 4.14. *Assuming* (A1)–(A4), *we get*

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p_h^n)dt\right\|^2+\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q_h^n\right\|^2$$

$$\le C\left(\tau+\epsilon+h^2+\tau\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right). \tag{4.46}$$

Proof. Let T_1 and T_2 denote the terms on the left in (4.46). For T_1, by the properties of norms we have

$$T_1 \le 2\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p^n)dt\right\|^2 + 2\tau^2\left\|\sum_{n=1}^{N}(p^n-p_h^n)\right\|^2.$$

Estimates (4.20) in Theorem 4.4 give

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p^n)dt\right\|^2 \le C(\tau+\epsilon),$$

which, together with (4.45), imply

$$T_1 \le C(\tau+\epsilon). \tag{4.47}$$

Analogously, for T_2 we obtain

$$\begin{aligned} T_2 &\le 2\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q^n\right\|^2 + 2\tau^2\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2 \\ &\le C_1(\tau+\epsilon) + C_2\left(\tau+\epsilon+h^2+\tau\sum_{n=1}^{N}\|q^n-\Pi_h q^n\|^2\right) \end{aligned} \tag{4.48}$$

by the arguments above. Now (4.46) follows from (4.47) and (4.48). ∎

COROLLARY 4.15. *Under the assumptions of Theorem* 4.14*, for the scalar product we have*

$$\begin{aligned} &\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt \\ &\quad\le C\left(\tau^{\frac12}+\epsilon^{\frac12}+h^2/\tau^{\frac12}+\tau^{\frac12}\sum_{n=1}^{N}\|q^n-\Pi_h q^n\|^2\right). \end{aligned} \tag{4.49}$$

Proof. We decompose the scalar product as follows:

$$\begin{aligned} &\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt \\ &\quad=\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p^n),u(t)-p_h^n)dt \\ &\qquad+\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(p^n)-b_\epsilon(p_h^n),u(t)-p_h^n)dt =: T_1+T_2. \end{aligned}$$

Applying Cauchy's inequality T_1 yields

$$\begin{aligned} |T_1| &\le \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|\|u(t)-p_h^n\|dt \\ &\le \frac{1}{4\delta_1}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|^2dt + \delta_1\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt. \end{aligned}$$

Since b_ϵ is Lipschitz, using (4.20), the first sum gives

$$\begin{aligned}
&\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|^2dt\\
&\quad\le C\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p^n),u(t)-p^n)dt\\
&\quad\le C(\tau+\epsilon).
\end{aligned}$$

Having $u\in L^2(J\times\Omega)$, by (4.29) we get

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt\le C.$$

In this way, choosing $\delta_1=(\tau+\epsilon)^{\frac12}$ yields

$$|T_1|\le C(\tau+\epsilon)^{\frac12}\le C(\tau^{\frac12}+\epsilon^{\frac12}). \tag{4.50}$$

For T_2 we obtain

$$\begin{aligned}
|T_2|&\le\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|\|u(t)-p_h^n\|dt\\
&\le\frac{\tau^{\frac12}}{4}\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\tau^{\frac12}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt.
\end{aligned}$$

As before, the second sum above is uniformly bounded, while for the first one we write

$$\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2\le C\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n).$$

Using (4.45) gives

$$\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2\le C\left(h^2/\tau+\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right),$$

so T_2 is bounded by

$$|T_2|\le C\left(\tau^{\frac12}+h^2/\tau^{\frac12}+\tau^{\frac12}\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right). \tag{4.51}$$

The result follows now from (4.50) and (4.51). ∎

Assuming additionally

(A5) $q^n\in H^1(\Omega)^d$ for all $n=\overline{1,N}$,

using (4.24) and the estimates in Theorem 4.14 and Corollary 4.15 we obtain the following theorem.

THEOREM 4.16. *Assuming* (A1)–(A5) *we have*

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p_h^n)dt\right\|^2+\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q_h^n\right\|^2\le C(\tau+\epsilon+h^2),$$

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt\le C(\tau^{\frac{1}{2}}+\epsilon^{\frac{1}{2}}+h^2/\tau^{\frac{1}{2}}). \tag{4.52}$$

REMARK 4.17. *Obviously* (A5) *is fulfilled in one spatial dimension, since then* $H(\mathrm{div},\Omega)$ *and* $H^1(\Omega)$ *coincide. Assumption* (A5) *also holds in the multidimensional case, provided* $\partial\Omega$ *is smooth enough and* k *is differentiable. Using* (A2) *and* (A3)*, since* $k(b(\cdot))\in C^1(0,1)$ *we have*

$$|\partial_u k(b(u))|\le\lim_{\delta\to 0}\left|\frac{k(b(u+\delta))-k(b(u))}{\delta}\right|\le\sqrt{C_k\frac{b(u+\delta)-b(u)}{\delta}}\le C.$$

Following [19, *Chapter* 4, *Theorems* 5.1 *and* 5.2]*, for any* $n=\overline{1,N}$*,* u^n *solving Problem* 3 *is in* $H^2(\Omega)$ *and the corresponding norm is bounded uniformly in* n *by a constant that, nevertheless, may depend on* τ*. Therefore* $q^n\in H^1(\Omega)$ *for all* $n\ge 1$ *and* $\|q^n\|_1\le C(\tau)$.

To confirm our theoretical results we present a numerical test. We consider a problem allowing for a travel wave solution, as proposed in [12] which refers to the Richards' equation in its form after the Kirchhoff transformation (1.4), without gravitation term and with

$$b(u)=\begin{cases}\frac{\pi^2}{2}-\frac{u^2}{2} & \text{for} \quad u\le 0,\\ \frac{\pi^2}{2} & \text{for} \quad u>0.\end{cases}$$

For this problem an exact solution is known:

$$u_{\mathrm{ex}}(t,x,y)=\begin{cases}\frac{-2(e^s-1)}{e^s+1} & \text{for} \quad s\ge 0,\\ -s & \text{for} \quad s<0,\end{cases}$$

where $s=x-y-t$. The equation has been solved in the unit square Ω, with Dirichlet boundary condition given by $u=u_{\mathrm{ex}}$ on $\partial\Omega$ and initial value u_{ex} at $t=0$. For mixed finite element discretizations the emerging system of equations is difficult to solve due to being the solution of a saddle point problem. A common implementation trick is to enlarge the system by adding Lagrange multipliers on edges (hybridization of the method). Briefly, within one timestep the resulting algorithm reads as follows: first the flux variable is eliminated on each element; then the continuity equation is locally solved for pressure by a variably damped Newton's method. The global system is set for the Lagrange multipliers and solved using the Newton method and a multigrid solver for the linear subproblems in the Newton iterations (for details see [25]). An alternative linearization approach is discussed in [24]. The implementation is based on the package UG (version 3.8, see also [3]), and the computations have been done on a SUN workstation.

To verify the theoretical estimates, we have started performing computations on a uniform triangular mesh with $h=0.25$, and a time step $\tau=0.04$. Then τ and h^2 are successively halved, up to $\tau=0.000625$ and $h=0.03125$. The final time was set

TABLE 4.1
Numerical results.

N	τ	h	Error	$\tau + h^2$	Convergence Order
1	0.04	0.25	6.344201e-06	1.025000e-01	—
2	0.02	0.176	3.620119e-06	5.125000e-02	0.81 0.81
3	0.01	0.125	2.057356e-06	2.562500e-02	0.82 0.81
4	0.005	0.088	9.574634e-07	1.281250e-02	1.10 0.91
5	0.0025	0.0625	5.362175e-07	6.406250e-03	0.84 0.89
6	0.00125	0.044	2.431734e-07	3.203250e-03	1.14 0.94
7	0.000625	0.03125	1.355397e-07	1.601562e-03	0.84 0.92

to be 1.0 for all the computations. Knowing the exact solution, the square of the total error (as written in (4.52)) is given by

$$E_{tot}^2 = \left\| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (u_{\mathrm{ex}}(t) - p_h^n)\, dt \right\|^2 + \left\| \tilde{q}_{\mathrm{ex}}(T) - \tau \sum_{n=1}^{N} q_h^n \right\|^2,$$

where $\tilde{q}_{\mathrm{ex}}(T) = \int_0^T \nabla\, u_{\mathrm{ex}}(t)\, dt$ is the exact flux. The order of convergence (for the squared error) is estimated by dividing the errors above, computed for two sets of parameters (refined according to the procedure mentioned above). Dividing the natural logarithm of the result by the natural logarithm of the refinement ratio yields an approximation of the convergence order. Results are displayed in Table 4.1. As predicted by Theorem 4.14, the order of E_{tot}^2 is between 0.8 and 1.1. Thus we can conclude that the numerical results are in concordance with our theoretical analysis, in particular confirming the convergence of the scheme.

Acknowledgments. We would like to thank Prof. C. J. van Duijn and Dr. E. F. Kaasschieter for useful discussions and suggestions.

REFERENCES

[1] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] T. ARBOGAST, M. F. WHEELER, AND N. Y. ZHANG, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[3] P. BASTIAN, K. BIRKEN, K. JOHANSSEN, S. LANG, N. NEUSS, H. RENTZ-REICHERT, AND C. WIENERS, *UG–a flexible toolbox for solving partial differential equations*, Comput. Visualiz. Sci., 1 (1997), pp. 27–40.

[4] R. G. BACA, J. N. CHUNG, AND D. J. MULLA, *Mixed transform finite element method for solving the non-linear equation for flow in variably saturated porous media*, Internat. J. Numer. Methods Fluids, 24 (1997), pp. 441–455.

[5] J. BEAR AND Y. BACHMAT, *Introduction to Modelling of Transport Phenomena in Porous Media*, Kluwer Academic, Dordrecht, The Netherlands, 1991.

[6] L. BERGANASCHI AND M. PUTTI, *Mixed finite elements and Newton-type linearizations for the solution of Richards' equation*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 1025–1046.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] M. A. CELIA, E. T. BOULOUTAS, AND R. L. ZARBA, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res., 26 (1990), pp. 1483–1496.

[9] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] R. EYMARD, M. GUTNIC, AND D. HILLHORST, *The finite volume method for Richards equation*, Comput. Geosci., 3 (1999), pp. 259–294.

[11] K. Fadimba and R. Sharpley, *A priori estimates and regularization for a class of porous medium equations*, Nonlinear World, 2 (1995), pp. 13–41.
[12] U. Hornung and W. Messing, *Poröse Medien—Methoden und Simulation*, Verlag Beiträge zur Hydrologie, Kirchzarten, 1984.
[13] W. Jäger and J. Kačur, *Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes*, RAIRO Model. Math. Numer. Anal., 29 (1995), pp. 605–627.
[14] D. Kavetski, P. Binning, and S. W. Sloan, *Adaptive time stepping and error control in a mass conservative numerical solution of the mixed form of Richards equation*, Adv. Water Res. 24 (2001), pp. 595–605.
[15] P. Knabner, *Finite element simulation of saturated-unsaturated flow through porous media*, in Large Scale Scientific Computing, Progress in Scientific Computing 7, P. Deuflhard and B. Engquist, eds., Birkhäuser, Boston, 1987, pp. 83–93.
[16] P. Knabner and E. Schneid, *Adaptive hybrid mixed finite element discretization of instationary variably saturated flow in porous media*, in High Performance Scientific and Engineering Computing, M. Breuer, F. Durst, and C. Zenger, eds., Springer-Verlag, Berlin, 2002, pp. 37–44.
[17] P. Knabner and E. Schneid, *Numerical solution of unsteady saturated/unsaturated flow through porous media*, in Numerical Modelling in Continuum Mechanics, Part II, M. Feistauer, R. Rannacher, and K. Kozel, eds., Matfyzpress, Prague, 1997, pp. 337–343.
[18] P. Knabner and G. Summ, *Efficient realization of the mixed finite element discretization for nonlinear problems*, Math. Comp., submitted.
[19] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, London, 1968.
[20] J. L. Lions and E. Magenes, *Non Homogenous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, 1972.
[21] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[22] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.
[23] I. S. Pop, *Error estimates for a time discretization method for the Richards' equation*, Comput. Geosci., 6 (2002), pp. 141–160.
[24] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: Linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373.
[25] E. Schneid, *Hybrid-Gemischte Finite-Elemente-Diskretisierung der Richards-Gleichung* (in German), Ph.D. thesis, University of Erlangen–Nürnberg, 2000; also available at http://www.am.uni-erlangen.de/am1/publications/dipl_phd_thesis/dipl_phd_thesis.html.
[26] E. Schneid, P. Knabner, and F. Radu, *A priori error estimates for a mixed finite element discretization of the Richards' equation*, Numer. Math., 98 (2004), pp. 353–370.
[27] M. Watanabe, *An approach by difference to the porous medium equation with convection*, Hiroshima Math. J., 25 (1995), pp. 623–645.
[28] G. A. Williams and C. T. Miller, *An evaluation of temporally adaptive transformation approaches for solving Richards' equation*, Adv. Water Res. 22 (1999), pp. 831–840.
[29] C. Woodward and C. 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 (2000), pp. 701–724.
[30] I. Yotov, *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 (1997), pp. 211–230.