ELECTRONIC RESEARCH ARCHIVE
Volume **29**, Number **5**, November **2021**

doi:10.3934/era.2021045
pp. **3405–3427**

AN ADJOINT-BASED A POSTERIORI ANALYSIS OF NUMERICAL APPROXIMATION OF RICHARDS EQUATION

Victor Ginting*

Department of Mathematics and Statistics
University of Wyoming
Laramie, WY 82071, USA

Abstract. This paper formulates a general framework for a space-time finite element method for solving Richards Equation in one spatial dimension, where the spatial variable is discretized using the linear finite volume element and the temporal variable is discretized using a discontinuous Galerkin method. The actual implementation of a particular scheme is realized by imposing certain finite element space in temporal variable to the variational equation and appropriate "variational crime" in the form of numerical integrations for calculating integrations in the formulation. Once this is in place, adjoint-based error estimators for the approximate solution from the scheme is derived. The adjoint problem is obtained from an appropriate linearization of the nonlinear system. Numerical examples are presented to illustrate performance of the methods and the error estimators.

1. **Introduction.** The subject of investigation in this paper is numerical solutions of the Richards Equation [23]. This equation is a governing mathematical principle for modeling the water flow in an unsaturated porous medium that is driven by the gravity and capillarity that disregards air flow. Since ability to construct closed form solutions to this equation is very limited (see for example [25, 18, 22, 24] for some related effort on the subject), a reliance on numerical approximations is a necessity. However, even with the emergence of many advances of computing technology, this equation remains one of the most challenging problems in porous media flow and transport. Recent review on its numerical solutions can be found in [17].

There are several outstanding issues attributed to the challenge. Richards Equation is strongly nonlinear, which appears as the dependence of the soil unsaturated hydraulic conductivity (κ) and the water content (ϑ) to the pressure head (u). Note that the presence of the water content in the equation is in terms of its temporal rate of change. Inclusion of gravity in the Darcy's velocity q, written as $q = -\kappa(u)\partial_z(u - z)$, can potentially create instability in the numerical solutions, in particular, when simulating dry soil conditions. The variable z in the expression of q denotes the vertical spatial coordinate, which is positive in the downward direction. It represents the influence of gravity to the flow. Furthermore, some of the more realistic scenario requires taking into account the soil heterogeneity in the simulations.

2020 *Mathematics Subject Classification.* Primary: 65M60, 65M08; Secondary: 65Z05.

Key words and phrases. Richards equation, space-time finite element, finite volume element, a posteriori error estimation, adjoint methods.

* Corresponding author: Victor Ginting.

The mixed-form (or coupled-form) of Richards Equation in one dimensional soil column is written as follows:

$$\begin{cases} \partial_t \vartheta(u) - \partial_z\big(\kappa(u)\,\partial_z(u-z)\big) = 0, \text{ in } (a,b)\times(0,T), \\ u(z,0) = u_0(z),\ z \in (a,b), \\ \text{Boundary Conditions: } \mathcal{B}_a u = g_a(t),\ \mathcal{B}_b = g_b(t). \end{cases} \tag{1.1}$$

Two typical boundary conditions are

$$\big\{\mathcal{B}_a u = \kappa(u)\partial_z(u-z)\big|_{(a,t)},\ \mathcal{B}_b u = u(b,t)\big\} \text{ or } \big\{\mathcal{B}_a u = u(a,t),\ \mathcal{B}_b u = u(b,t)\big\}.$$

Here we assume that $\vartheta : (-\infty, 0] \to (0,1)$ and $\kappa : (-\infty, 0] \to (\kappa_{\min}, \kappa_s)$ with $\kappa_s > \kappa_{\min} > 0$. The choice $(-\infty, 0]$ as the domain of ϑ and κ is done to reflect the physical relevance that the pressure head u is always nonpositive and Richards Equation typically models unsaturated flow.

The major theme in this paper is two-fold. One aspect centers on the development of a numerical approximation of Richards Equation in space-time finite element methods obtained from an appropriate variational formulation. Space-time finite element methods have been previously used for parabolic equations (see for example [20]) and for reaction-diffusion system (see for example [15]). A recent work on application of control volume finite element in combination with method of lines for solving Richards Equation is recorded in [10]. To the best of the author's knowledge, there has not been any attempt to apply space-time finite element methods technique to Richards Equation. In particular, a finite volume element spatial discretization (with linear finite element) is employed due to its inherent local mass conservation property. This is an important trait commonly desired and in some cases imperative in order to produce reliable numerical simulations of flow and transport in porous media (see for example [5, 8, 16]). The space-time variational formulation in combination with a certain variational crime in the form of numerical integration techniques would in turn yield implementable time marching schemes for approximate solution of Richards Equation.

The other aspect is concerned with an a posteriori error estimation of the resulting numerical approximation. In this regard, some investigations on a posteriori error analysis of numerical methods for Richards Equation are already available. Bause and Knabner [3] use adaptive mixed hybrid finite element discretizations to solve Richards Equation, where the adaptivity is performed under an a posteriori error indicator that is based on either superconvergence or residual of the approximation. Baron et al. [2] employ Discrete Duality Finite Volume (DDFV) scheme along with second-order backward differentiation formula to solve the equation. They derive an a posteriori error bound of the approximation using the equilibrated fluxes method. Bernardi et al. [4] perform a semi discretization of Richards Equation by finite element method and apply Backward Euler scheme to get the full discretization. Then a posteriori error bounds are derived that aim at distinguishing components of contribution of spatial discretization from temporal discretization.

In many practical situations, it may not be necessary to measure global property of the approximate solution. More often, an accuracy is desired only for some specified quantities of interest associated with the numerical approximation, which is usually expressed as a functional of the approximate solution. For this purpose, a suitable a posteriori analysis is based on duality, adjoint operators governing the generalized Green's function, and a variational formulation. This approach is

adopted in this paper. It is also suitable because, as mentioned before, the numerical approximation of Richards Equation is based on certain variational equations. Utilizations of adjoint equations are not new (see for example [21] for an extensive exposition on their applications). On various roles of adjoint methodologies in performing a posteriori error estimations of numerical methods for differential equations, one can consult [19, 15, 11, 13, 12, 14, 1] and references therein.

The rest of this paper is organized as follows. A space-time variational equation governing the numerical approximation of Richards Equation is derived in Section 2. The description includes examples of application of numerical integration to produce the time marching schemes. Section 3 carries out the formulation of an adjoint-based a posteriori error analysis of the quantities of interest calculated using the approximate solution. Since the variational equation of the solution is nonlinear, an appropriate linearization is conducted that would make construction of the corresponding adjoint equation amenable. Some numerical examples to demonstrate performance of the numerical methods and the error estimators are shown in Section 4. Here much of the effort is devoted to illustrate global accuracy of numerical methods and reliability of the error estimators in terms of their capability to decompose the total error into relevant components. A comparison to the actual error in some specified quantities of interest is conducted. Finally, the conclusion and future work is discussed in Section 5.

2. **Finite volume element in space and finite element in time.** In what follows, we assume that

$$\big\{\mathcal{B}_a u = \kappa(u)\partial_z(u-z)\Big|_{(a,t)} = g_a(t),\ \mathcal{B}_b u = u(b,t) = 0\big\}. \tag{2.1}$$

are supplied to (1.1). Denoting

$$H^1_{\mathrm{D}} = \Big\{w : [a,b] \to \mathbb{R} : w \in L^2(a,b),\ w' \in L^2(a,b),\ w(b) = 0\Big\},$$

the solution of (1.1) supplied with (2.1) satisfies

$$\begin{cases} \langle \partial_t \vartheta(u), v\rangle + A(u;u,v) + g_a(t)v(a) = 0,\ \forall v \in H^1_{\mathrm{D}}, \\ \langle u(\cdot,0), v\rangle = \langle u_0, v\rangle,\ \forall v \in H^1_{\mathrm{D}}. \end{cases} \tag{2.2}$$

Here $\langle\cdot,\cdot\rangle$ is the usual scalar product in $L^2(a,b)$, and $A : C[a,b] \times H^1_{\mathrm{D}} \times H^1_{\mathrm{D}} \to \mathbb{R}$ is defined as

$$A(w;u,v) = \langle \kappa(w)\partial_z(u-z), \partial_z v\rangle.$$

The spatial domain (a,b) is partitioned into a collection of M subintervals $\mathcal{T}_h$, such that $\tau_j = (z_{j-1}, z_j) \in \mathcal{T}_h$, with length $h_j = z_j - z_{j-1}$, for $j = 1, \cdots, M$ and $(a,b) = \cup_{j=1}^{M}\tau_j$, where $h = \max\{h_j : 1 \le j \le M\}$. On this $\mathcal{T}_h$, let

$$\mathcal{X}_h = \big\{\ w \in H^1_{\mathrm{D}}\ :\ w \text{ in } \tau_j \text{ is linear } \forall \tau_j \in \mathcal{T}_h\ \big\} = \mathrm{span}\big\{\phi_j\big\}_{j=0}^{M-1},$$

where $\phi_j(z)$ is the usual 'hat' function such that $\phi_j(z_i) = \delta_{ij}$.

2.1. **A brief excursion to finite volume element.** The finite volume approximations rely on a local conservation property associated with the governing equation, in particular with respect to the second order differential operator in (1.1). To fix the idea, consider $\tau^* = (z_l, z_r) \subset (a,b)$ and apply fundamental theorem of calculus to get

$$\int_{\tau^*} -\partial_z\big(\kappa(u)\partial_z(u-z)\big)\,\mathrm{d}z = -\kappa(u)\partial_z(u-z)\ \Big|_{\partial\tau^*} := -\kappa(u)\partial_z(u-z)\ \Big|_{z_l}^{z_r}. \tag{2.3}$$

The above τ^* is called a control volume. We choose M control volumes. Specifically, given z_j, for $j = 1, \cdots, M-1$, we set $\tau_j^* = (z_{j-1/2}, z_{j+1/2})$, where $z_{j-1/2}$ is the mid-point of τ_j and $z_{j+1/2}$ is the mid-point of τ_{j+1}. For z_0, we set $\tau_0^* = (z_0, z_{1/2})$. Collection of these control volumes is denoted by $\mathcal{T}_h^*$. On this $\mathcal{T}_h^*$, let

$$\mathcal{Y}_h = \left\{ \ \eta \in L^2(a,b) \ : \ \eta \text{ in } \tau^* \text{ is constant } \forall \tau_j^* \in \mathcal{T}_h^* \ \right\} = \operatorname{span}\left\{\phi_j^*\right\}_{j=0}^{M-1},$$

where $\phi_j^*(z)$ is a piecewise constant function such that it is equal to 1 in τ_j^* and zero elsewhere. Set $J_h : H^1_{\mathrm{D}} \to \mathcal{Y}_h$ such that $[J_h v](z_j) = v(z_j)$, i.e., given $v \in H^1_{\mathrm{D}}$, $J_h v$ is a piecewise constant function over $\mathcal{T}_h^*$. A standard interpolation estimate suggests

$$\|J_h v - v\| \leq \frac{h}{2} \|\partial_z v\|, \text{ for } v \in H^1_{\mathrm{D}}, \tag{2.4}$$

where $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. By the Cauchy-Schwarz inequality, this implies

$$\langle \chi, J_h v - v \rangle \leq \frac{h}{2} \|\chi\| \|\partial_z v\|, \text{ for } \chi \in L^2(a,b), \ v \in H^1_{\mathrm{D}}. \tag{2.5}$$

To express (2.3) in a variational setting, define $A_h : C[a,b] \times H^1_{\mathrm{D}} \times H^1_{\mathrm{D}} \to \mathbb{R}$ as

$$A_h(w; u, v) = \sum_{\tau_j^* \in \mathcal{T}_h^*} -\kappa(w) \partial_z (u - z) \Big|_{\partial \tau_j^* \setminus a} [J_h v](z_j).$$

Exclusion of a in the above equation is due to the Neumann boundary condition at that point. Next, we quantify the discrepancy of $A_h(\cdot; \cdot, \cdot)$ from $A(\cdot; \cdot, \cdot)$.

Proposition 2.1. *Let $w \in C[a,b]$, $u, v \in H^1_{\mathrm{D}}$. Then*

$$A(w; u, v) = A_h(w; u, v) + \varepsilon_A(w; u, v), \tag{2.6}$$

where

$$\varepsilon_A(w; u, v) = \sum_{\tau_j \in \mathcal{T}_h} \int_{\tau_j} \partial_z \big(\kappa(w) \partial_z (u - z) \big) \, (J_h v - v) \, \mathrm{d}z. \tag{2.7}$$

Furthermore, when $u, w \in \mathcal{X}_h$ and $\kappa'(r)$ is bounded for every $r \in (-\infty, 0]$, then

$$|\varepsilon_A(w; u, v)| \leq \frac{C_\kappa}{2} h \|\partial_z w \partial_z (u - z)\| \|\partial_z v\|, \tag{2.8}$$

where $C_\kappa = \sup_{u \in (-\infty, 0]} |\kappa'(u)|$.

Proof. For any $\tau_j \in \mathcal{T}_h$, integration by parts gives

$$\int_{\tau_j} -\partial_z \big(\kappa(w) \partial_z (u - z) \big) \, v \, \mathrm{d}z = \int_{\tau_j} \kappa(w) \partial_z (u - z) \, \partial_z v \, \mathrm{d}z - \kappa(w) \partial_z (u - z) \, v \Big|_{\partial \tau_j},$$

which when applied to $A(\cdot; \cdot, \cdot)$ gives

$$\begin{aligned} A(w; u, v) &= \sum_{\tau_j \in \mathcal{T}_h} \int_{\tau_j} \kappa(w) \partial_z (u - z) \, \partial_z v \, \mathrm{d}z \\ &= \sum_{\tau_j \in \mathcal{T}_h} \Big(\int_{\tau_j} -\partial_z \big(\kappa(w) \partial_z (u - z) \big) \, v \, \mathrm{d}z + \kappa(w) \partial_z (u - z) \, v \Big|_{\partial \tau_j \setminus a} \Big). \end{aligned} \tag{2.9}$$

For $j = 0$, set $K_j = \tau_j^*$. For $j = 1, \cdots, M-1$, fix a $\tau_j^* \in \mathcal{T}_h^*$ and suppose $\tau_j, \tau_{j+1} \in \mathcal{T}_h$ are such that $K_j = \tau_j \cap \tau_j^* = (x_{j-1/2}, x_j)$, $K_{j+1} = \tau_{j+1} \cap \tau_j^* = (x_j, x_{j+1/2})$. By fundamental theorem of calculus,

$$\int_{K_e} -\partial_z \big(\kappa(w) \partial_z (u - z) \big) \, J_h v \, \mathrm{d}z = -\kappa(w) \partial_z (u - z) \Big|_{\partial K_e} [J_h v](z_j), \tag{2.10}$$

for $e = j, j+1$. Recognizing that

$$\partial_z u\Big|_{\partial\tau_j^*} = \sum_{e=j,j+1} \partial_z u\Big|_{\partial K_e} + \partial_z u\Big|_{x_j^-}^{x_j^+},$$

we may apply (2.10) in $A_h(\cdot;\cdot,\cdot)$ to get

$$A_h(w;u,v) = \sum_{\tau_j\in\mathcal{T}_h} \Big(\int_{\tau_j} -\partial_z\big(\kappa(w)\partial_z(u-z)\big)\, J_h v \,\mathrm{d}z + \kappa(w)\partial_z(u-z)\, J_h v\Big|_{\partial\tau_j\setminus a}\Big). \quad (2.11)$$

Subtraction of (2.9) from (2.11) and recalling that $J_h v(z_j) = v(z_j)$ gives (2.6).

Furthermore, when $w, u \in \mathcal{X}_h$, product rule of differentiation for $z \in \tau_j$ gives

$$\partial_z\big(\kappa(w)\partial_z(u-z)\big) = \kappa'(w)\partial_z w\partial_z(u-z) + 0,$$

so using this identity and the Cauchy-Schwarz inequality,

$$\begin{aligned}
|\varepsilon_A(w;u,v)| &= \Big| \sum_{\tau_j\in\mathcal{T}_h} \int_{\tau_j} \kappa'(w)\partial_z w\partial_z(u-z)(J_h v - v)\,\mathrm{d}z\Big| \\
&\le C_\kappa \sum_{\tau_j\in\mathcal{T}_h} \|\partial_z w\partial_z(u-z)\|_{L^2(\tau_j)}\, \|J_h v - v\|_{L^2(\tau_j)} \\
&\le C_\kappa \|\partial_z w\partial_z(u-z)\|\, \|J_h v - v\| \\
&\le \frac{C_\kappa}{2} h \|\partial_z w\partial_z(u-z)\| \|\partial_z v\|.
\end{aligned}$$

This completes the proof. □

Remark 2.1. The foregoing exposition gives an indication that $\langle w, J_h v - v\rangle \to 0$ and $A_h(w;u,v) \to A(w;u,v)$ as $h \to 0$. This will play a role later on in the a posteriori error analysis. Various estimates such as described in (2.4), (2.5), and (2.8) have been established in several literatures on finite volume element methods (see for example [6, 7, 9] and references therein).

2.2. **A variational equation for the approximation.** In a similar fashion to the spatial variable, we partition $[0,T]$ into a collection of subintervals $\mathcal{I}_k$, such that $I_n = [t_{n-1}, t_n] \in \mathcal{I}_k$ with time step $k_n = t_n - t_{n-1}$ and $[0,T] = \cup_{n=1}^N I_n$, where $k = \max\{k_n : 1 \le n \le N\}$. We denote the jump of a function $w(\cdot,t)$ across t_n by $[w]_n = w_n^+ - w_n^-$, where $w_n^+ = \lim_{s\to t_n^+} w(\cdot,s)$ and $w_n^- = \lim_{s\to t_n^-} w(\cdot,s)$. On every space-time slab $[a,b]\times I_n$, the approximate solution is sought in a functional space that contains functions that are piecewise linear polynomial in spatial variable and polynomial of degree q in temporal variable. In particular, we define

$$\mathcal{W}_h^q(I_n) = \Big\{ v : [a,b]\times I_n \to \mathbb{R} : w(z,t) = \sum_{j=0}^q t^j v_{j,n}(z),\ \text{with } v_{j,n} \in \mathcal{X}_h \Big\}. \quad (2.12)$$

We denote by $\mathcal{W}_h^q$ the space of functions defined on $[a,b]\times[0,T]$ such that restriction of $w \in \mathcal{W}_h^q$ to $[a,b]\times I_n$ belongs to $\mathcal{W}_h^q(I_n)$. The approximation amounts to finding $\widetilde{u} \in \mathcal{W}_h^q$ that is governed by

$$\begin{cases} \displaystyle\sum_{n=1}^N R_{h,n}(\widetilde{u};\widetilde{u},w) = 0 \quad \text{for every } w \in \mathcal{W}_h^q, \\[2ex] \langle \widetilde{u}_0^-, \chi\rangle = \langle u_0, \chi\rangle \quad \text{for every } \chi \in \mathcal{X}_h, \end{cases} \quad (2.13)$$

where

$$R_{h,n}(\widetilde{u};\widetilde{u},w) = \int_{I_n}\Big(\langle\partial_t\vartheta(\widetilde{u}),J_hw\rangle + A_h(\widetilde{u};\widetilde{u},w) + g_a(t)w(a,t)\Big)\mathrm{d}t \\ + \langle[\vartheta(\widetilde{u})]_{n-1},J_hw^+_{n-1}\rangle. \tag{2.14}$$

Notice that the first equation in (2.13) is a global formulation in that the integration is over $(0,T)$. While an implementation based on this formulation is possible, it is perhaps more amenable to construct an implementation that is local over I_n for $n=1,\cdots,N$. The corresponding equation for this formulation can be derived from (2.13) by choosing $w\in\mathcal{W}_h^q$ such that $w|_{(a,b)\times I_n}=v\in\mathcal{W}_h^q(I_n)$ and it is zero everywhere else, which yields

$$R_{h,n}(\widetilde{u};\widetilde{u},v)=0 \quad \text{for every } v\in\mathcal{W}_h^q(I_n). \tag{2.15}$$

2.3. **Some examples.** In what follows, we describe two specific examples that transform (2.15) into computable algebraic schemes.

2.3.1. *FVEM in space - dG0 in time.* Here $\tilde{u}\in\mathcal{W}_h^0$, i.e.,

$$\widetilde{u}\Big|_{I_n} = \widetilde{u}^+_{n-1} = \widetilde{u}^-_n = v_{0,n}\in\mathcal{X}_h, \tag{2.16}$$

which for every $w_0\in\mathcal{X}_h$ is governed by

$$k_nA_h(v_{0,n};v_{0,n},w_0) + w_0(a)\int_{I_n}g_a(t)\,\mathrm{d}t + \langle\vartheta(v_{0,n})-\vartheta(\widetilde{u}^-_{n-1}),J_hw_0\rangle = 0, \tag{2.17}$$

for $n=1,\cdots,N$. Notice that (2.17) is mimicking the Backward Euler difference scheme with $v_{0,n}\in\mathcal{X}_h$ being the unknown function to be solved. In particular, setting $(U_{0,n},U_{1,n},\cdots,U_{M-1,n})=\boldsymbol{U}_n\in\mathbb{R}^M$, such that

$$v_{0,n}=\sum_{j=0}^{M-1}U_{j,n}\phi_j, \tag{2.18}$$

then $\boldsymbol{U}_n$ is governed by

$$G(\boldsymbol{U}_n)=\boldsymbol{0}, \tag{2.19}$$

where $G:\mathbb{R}^M\to\mathbb{R}^M$ such that $G_i:\mathbb{R}^M\to\mathbb{R}$, for $i=0,1,\cdots,M-1$, is constructed from the left hand side of (2.17) by replacing w_0 by ϕ_i. Here, the dependence on $\boldsymbol{U}_n$ is realized through (2.18). Clearly (2.19) is a nonlinear algebraic system of equations governing $\boldsymbol{U}_n$.

2.3.2. *FVEM in space - dG1 in time.* Here $\tilde{u}\in\mathcal{W}_h^1$, i.e.,

$$\widetilde{u}\Big|_{I_n} = v_{0,n}+tv_{1,n},\ t\in I_n,\ v_{0,n},v_{1,n}\in\mathcal{X}_h, \tag{2.20}$$

which implies that $\widetilde{u}^+_{n-1}=v_{0,n}+t_{n-1}v_{1,n}$ and $\widetilde{u}^-_n=v_{0,n}+t_nv_{1,n}$. We may equivalently write

$$\widetilde{u}\Big|_{I_n} = \frac{t_n-t}{k_n}\,\widetilde{u}^+_{n-1}+\frac{t-t_{n-1}}{k_n}\,\widetilde{u}^-_n,\ t\in I_n,\ \widetilde{u}^+_{n-1},\widetilde{u}^-_n\in\mathcal{X}_h. \tag{2.21}$$

Choosing $w=\frac{t_n-t}{k_n}\,\psi^+_{n-1}$ with $\psi^+_{n-1}\in\mathcal{X}_h$, and using integration by parts along with acknowledging some cancellations,

$$\int_{I_n}\langle\partial_t\vartheta(\widetilde{u}),J_hw\rangle\,\mathrm{d}t + \langle[\vartheta(\widetilde{u})]_{n-1},J_hw^+_{n-1}\rangle = k_n^{-1}\int_{I_n}\langle\vartheta(\widetilde{u})-\vartheta(\widetilde{u}^-_{n-1}),J_h\psi^+_{n-1}\rangle\,\mathrm{d}t.$$

In a similar fashion, using $w = \frac{t-t_{n-1}}{k_n}\,\psi_n^-$ with $\psi_n^- \in \mathcal{X}_h$ yields

$$\int_{I_n} \langle \partial_t \vartheta(\widetilde{u}), J_h w\rangle \,\mathrm{d}t + \langle [\vartheta(\widetilde{u})]_{n-1}, J_h w_{n-1}^+\rangle = k_n^{-1} \int_{I_n} \langle \vartheta(\widetilde{u}_n^-) - \vartheta(\widetilde{u}), J_h \psi_n^-\rangle \,\mathrm{d}t.$$

Thus $\widetilde{u} \in \mathcal{W}_h^1$ satisfies (2.21) and for every $n = 1, \cdots, N$, it is governed by

$$\begin{cases} \displaystyle\int_{I_n} \langle \vartheta(\widetilde{u}) - \vartheta(\widetilde{u}_{n-1}^-), J_h \psi_{n-1}^+\rangle \,\mathrm{d}t + \\ \qquad \displaystyle\int_{I_n} (t_n - t)\Big(A_h(\widetilde{u}; \widetilde{u}, \psi_{n-1}^+) + g_a(t)\psi_{n-1}^+(a)\Big)\mathrm{d}t = 0, \\ \displaystyle\int_{I_n} \langle \vartheta(\widetilde{u}_n^-) - \vartheta(\widetilde{u}), J_h \psi_n^-\rangle \,\mathrm{d}t + \\ \qquad \displaystyle\int_{I_n} (t - t_{n-1})\Big(A_h(\widetilde{u}; \widetilde{u}, \psi_n^-) + g_a(t)\psi_n^-(a)\Big)\mathrm{d}t = 0, \end{cases} \tag{2.22}$$

where

$$\widetilde{u}_{n-1}^+ = \sum_{j=0}^{M-1} U_{j,n-1}^+ \phi_j, \ \ \widetilde{u}_n^- = \sum_{j=0}^{M-1} U_{j,n}^- \phi_j. \tag{2.23}$$

Setting

$$(U_{0,n-1}^+, U_{1,n-1}^+, \cdots, U_{M-1,n-1}^+) = \boldsymbol{U}_{n-1}^+ \in \mathbb{R}^M$$

and

$$(U_{0,n}^-, U_{1,n}^-, \cdots, U_{M-1,n}^-) = \boldsymbol{U}_n^- \in \mathbb{R}^M,$$

and $\boldsymbol{U}_n = (\boldsymbol{U}_{n-1}^+, \boldsymbol{U}_n^-) \in \mathbb{R}^{2M}$, then (2.22) yields

$$G(\boldsymbol{U}_n) = \boldsymbol{0}, \tag{2.24}$$

where $G : \mathbb{R}^{2M} \to \mathbb{R}^{2M}$, with $G = (G^+, G^-)$, and $G^+ : \mathbb{R}^{2M} \to \mathbb{R}^M$ and $G^- : \mathbb{R}^{2M} \to \mathbb{R}^M$ such that $G_i^+ : \mathbb{R}^{2M} \to \mathbb{R}$ and $G_i^- : \mathbb{R}^{2M} \to \mathbb{R}$ are respectively constructed from the left hand side of (2.22) by setting $\psi_{n-1}^+ = \psi_n^- = \phi_i$, for $i = 0, 1, \cdots, M-1$.

2.4. **A variational crime by numerical integrations.** The preceding description is a derivation of algebraic equations governing the approximation that is faithful to the variational equation (2.15) and the choice of polynomial degree of the temporal variable. Still, for a completely implementable scheme, one must rely on further approximation of the integrations appeared in (2.17) and (2.22). In the current setting, there are two integrations that need to be approximated: the spatial integration $\langle \cdot, \cdot \rangle$ and the temporal integration $\int_{I_n} \cdot \,\mathrm{d}t$. Utilization of various numerical integration techniques are pretty common. In particular, in the standard finite element methods for typical steady state problems, forms/functionals in the variational equations, which are expressed as integrations of spatial variables, are approximated by various Gaussian quadratures. This is clearly applicable for $\langle \cdot, \cdot \rangle$. To minimize the associated pollution to the global accuracy of the approximation, the numerical integrations must be chosen such that the degree of their errors is of similar order to the errors corresponding to W_h^q.

Furthermore, what is more crucial in this case is how $\int_{I_n} \cdot \,\mathrm{d}t$ is to be approximated. We note that the only temporal integration in (2.17) is the one associated with the Neumann condition $g_a(t)$, and for this a right hand point rule resulting in

$$\int_{I_n} g_a(t) \,\mathrm{d}t \approx k_n g_a(t_n)$$

is adequate.

Derivation of numerical integrations for (2.22) is a bit more involved. A viable option is the following two point Gaussian quadrature

$$\begin{aligned}&\int_{I_n} f(t)\,\mathrm{d}t \approx \frac{k_n}{2}\sum_{\ell=1}^{2} f(t_{\ell,n}),\\ &\text{where } t_{1,n} = -\frac{k_n}{2\sqrt{3}} + \frac{t_{n-1}+t_n}{2} \text{ and } t_{2,n} = \frac{k_n}{2\sqrt{3}} + \frac{t_{n-1}+t_n}{2}.\end{aligned} \tag{2.25}$$

With this, set

$$\begin{aligned}\widetilde{u}_{1,n} &= \widetilde{u}(\cdot,t_{1,n}) = \gamma\,\widetilde{u}_{n-1}^{+} + (1-\gamma)\,\widetilde{u}_{n}^{-},\\ \widetilde{u}_{2,n} &= \widetilde{u}(\cdot,t_{2,n}) = (1-\gamma)\,\widetilde{u}_{n-1}^{+} + \gamma\,\widetilde{u}_{n}^{-},\end{aligned} \tag{2.26}$$

where $\gamma = \frac{1+\sqrt{3}}{2\sqrt{3}}$. The resulting approximations $G_{n,i}^{+} \approx G_i^{+}$ and $G_{n,i}^{-} \approx G_i^{-}$ are expressed as

$$G_{n,i}^{+}(\boldsymbol{U}_n) = \frac{1}{2}\sum_{\ell=1}^{2}\langle\vartheta(\widetilde{u}_{\ell,n}) - \vartheta(\widetilde{u}_{n-1}^{-}),\phi_i^{*}\rangle + k_n c_\ell^{+}\big(A_h(\widetilde{u}_{\ell,n};\widetilde{u}_{\ell,n},\phi_i) + g_a(t_{\ell,n})\delta_{i0}\big)$$

$$G_{n,i}^{-}(\boldsymbol{U}_n) = \frac{1}{2}\sum_{\ell=1}^{2}\langle\vartheta(\widetilde{u}_{n}^{-}) - \vartheta(\widetilde{u}_{\ell,n}),\phi_i^{*}\rangle + k_n c_\ell^{-}\big(A_h(\widetilde{u}_{\ell,n};\widetilde{u}_{\ell,n},\phi_i) + g_a(t_{\ell,n})\delta_{i0}\big)$$

where $c_l^{+} = c_r^{-} = \gamma$, $c_r^{+} = c_l^{-} = 1-\gamma$, and δ_{ij} is the usual Kronecker delta.

The approach proceeds with the construction of algebraic equations for $(\widetilde{u}_{1,n}, \widetilde{u}_{2,n})$ where $\widetilde{u}_n^{-}$ appearing on the second equation above is represented as

$$\widetilde{u}_n^{-} = \frac{1-\gamma}{1-2\gamma}\widetilde{u}_{1,n} - \frac{\gamma}{1-2\gamma}\widetilde{u}_{2,n}, \tag{2.27}$$

which is obtained from (2.26). Thus, with $(G_{n,i}^{+}, G_{n,i}^{-})$ replacing (G_i^{+}, G_i^{-}), (2.24) is solved to get $(\widetilde{u}_{1,n}, \widetilde{u}_{2,n})$, after which $\widetilde{u}_n^{-}$ is recovered from (2.27).

3. An adjoint-based a posteriori error analysis.

In many realistic situations, it is often desirable to achieve an acceptable level of accuracy of a numerical approximation in some quantities of interest. Relevant examples include average water content over a certain region and at some time instances or the water content at some locations. Along this line of argument, it may be computationally infeasible as well as very inefficient to attempt to control the error in a global fashion when all that is required is accuracy on those aforementioned quantities. A practical alternative is to estimate the error of the numerical approximation in the specified quantity of interest, whose representation is expressed as a functional of u:

$$[Q(u)](T) = \langle\vartheta(u(\cdot,T)),\psi_T\rangle + \int_0^T \big(\langle\vartheta(u),\psi\rangle + u(a,t)\psi_a\big)\,\mathrm{d}t, \tag{3.1}$$

for given data $\psi_T : (a,b) \to \mathbb{R}$, $\psi : (a,b)\times(0,T) \to \mathbb{R}$, and $\psi_a : (0,T) \to \mathbb{R}$. If one wants to quantify the (averaged) water content at time $t = T$, then (3.1) uses $\psi = 0$, $\psi_a = 0$, and ψ_T is set to be a piecewise constant function in the spatial variable that reflects the desired nature of the average quantity. On the other hand, if an accumulated water content is the quantity of interest to be approximated, then $\psi_T = 0$, $\psi_a = 0$, and $\psi = 1$ in (3.1).

To derive the error in approximating $Q(u)$, we use a generalized Green's function that solves the adjoint problem corresponding to a special choice of (adjoint) data

ψ_T, ψ, and ψ_a, as illustrated by the description in the previous paragraph. As alluded to earlier, the formulation of an adjoint problem broadens the applications of Green's functions (see for example [19, 15, 1] and references therein). However, an adjoint operator formally corresponds to a linear operator. Since Richards Equation and the associated numerical approximation are nonlinear problem, we must perform a linearization, after which the adjoint problem is built to correspond to that linearized representation.

The nonlinearity in Richards Equation stems from $\theta(u)$ and $\kappa(u)$, and that is where the linearization effort is concentrated on. To this end, letting

$$\widetilde{u}_\sigma = \widetilde{u} + \sigma(u - \widetilde{u}), \text{ for } \sigma \in [0,1], \tag{3.2}$$

the Mean Value Theorem for integral gives

$$\vartheta(u) - \vartheta(\widetilde{u}) = \overline{\vartheta'}(u - \widetilde{u}), \text{ where } \overline{\vartheta'} = \int_0^1 \vartheta'(\widetilde{u}_\sigma)\,\mathrm{d}\sigma. \tag{3.3}$$

Furthermore, setting $F : H^1(a,b) \to \mathbb{R}$ by

$$F(w) = \kappa(w)\partial_z(w - z), \tag{3.4}$$

its Fréchet derivative is $F'(w) : H^1(a,b) \to \mathbb{R}$ such that

$$F'(w)v = \kappa(w)\partial_z v + (\kappa'(w)\partial_z(w - z))v \tag{3.5}$$

Utilizing again the Mean Value Theorem for integral, one gets

$$F(u) - F(\widetilde{u}) = \int_0^1 F'(u_\sigma)(u - \widetilde{u})\,\mathrm{d}\sigma = \overline{\kappa}\partial_z(u - \widetilde{u}) + \overline{\boldsymbol{v}}(u - \widetilde{u}), \tag{3.6}$$

where

$$\overline{\kappa} = \int_0^1 \kappa(\widetilde{u}_\sigma)\,\mathrm{d}\sigma \text{ and } \overline{\boldsymbol{v}} = \int_0^1 \kappa'(\widetilde{u}_\sigma)\partial_z(\widetilde{u}_\sigma - z)\,\mathrm{d}\sigma. \tag{3.7}$$

At this stage, we are in a position to formulate the adjoint problem. For $t \in [T, 0]$, let $\varphi(\cdot, t) \in H^1_{\mathrm{D}}$ satisfy

$$\begin{cases} -\langle w, \overline{\vartheta'}\partial_t\varphi\rangle + \langle \partial_z w, \overline{\kappa}\partial_z\varphi\rangle + \langle w, \overline{\boldsymbol{v}}\partial_z\varphi\rangle = \langle w, \overline{\vartheta'}\psi\rangle + w(a,t)\psi_a, \ t < T, \\ \langle w(\cdot,T), \overline{\vartheta'}\varphi(\cdot,T)\rangle = \langle w(\cdot,T), \overline{\vartheta'}\psi_T\rangle, \end{cases} \tag{3.8}$$

for every $w(\cdot,t) \in H^1_{\mathrm{D}}$. Here φ is solution to the adjoint problem, which is governed by a linear problem as stated in (3.8). The two theorems below state the quantification of error in the approximation of $Q(u)$, which is written in terms of residuals of $\widetilde{u}$ weighted against φ. In what follows, we use

$$\begin{aligned} R_n(\widetilde{u}; \widetilde{u}, w) = \int_{I_n} \big(\langle \partial_t\vartheta(\widetilde{u}), w\rangle + A(\widetilde{u}; \widetilde{u}, w) + g_a(t)w(a,t)\big)\,\mathrm{d}t \\ + \langle [\vartheta(\widetilde{u})]_{n-1}, w^+_{n-1}\rangle. \end{aligned} \tag{3.9}$$

Theorem 3.1. *For $\widetilde{u} \in \mathcal{W}^q_h$ in (2.13) and u in (2.2),*

$$[Q(u)](T) - [Q(\widetilde{u})](T) = E_0 + E_1 + E_2 + E_3, \tag{3.10}$$

where

$$\begin{aligned} E_0 &= \langle \vartheta(u_0) - \vartheta(\widetilde{u}_0^-), \varphi_0\rangle, \ E_1 = -\sum_{n=1}^N R_{h,n}(\widetilde{u}; \widetilde{u}, \varphi), \\ E_2 &= -\sum_{n=1}^N \varepsilon_{A,n}(\widetilde{u}; \widetilde{u}, \varphi)\,\mathrm{d}t, \ E_3 = -\sum_{n=1}^N \varepsilon_{h,n}(\widetilde{u}; \varphi), \end{aligned} \tag{3.11}$$

with $R_{h,n}(\widetilde{u};\widetilde{u},\varphi)$ *as expressed in* (2.14), *and*

$$\begin{aligned}\varepsilon_{A,n}(\widetilde{u};\widetilde{u},w) &= \int_{I_n}\varepsilon_A(\widetilde{u};\widetilde{u},w)\,\mathrm{d}t,\\ \varepsilon_{h,n}(\widetilde{u};w) &= \int_{I_n}\langle\partial_t\vartheta(\widetilde{u}),w-J_hw\rangle\,\mathrm{d}t+\langle[\vartheta(\widetilde{u})]_{n-1},(w-J_hw)^+_{n-1}\rangle.\end{aligned}\tag{3.12}$$

Proof. Substitute $w=e=u-\widetilde{u}$ in (3.8) so that

$$\begin{aligned}&-\langle e,\overline{\vartheta'}\partial_t\varphi\rangle+\langle\partial_z e,\overline{\kappa}\partial_z\varphi\rangle+\langle e,\overline{\boldsymbol{v}}\partial_z\varphi\rangle\\ &\quad=-\langle\vartheta(u)-\vartheta(\widetilde{u}),\partial_t\varphi\rangle+\langle\kappa(u)\partial_z(u-z)-\kappa(\widetilde{u})\partial_z(\widetilde{u}-z),\partial_z\varphi\rangle\\ &\quad=-\Big(\langle\partial_t\vartheta(\widetilde{u}),\varphi\rangle+A(\widetilde{u};\widetilde{u},\varphi)+g(t)\varphi(a,t)\Big)-\partial_t\langle\vartheta(u)-\vartheta(\widetilde{u}),\varphi\rangle,\end{aligned}\tag{3.13}$$

where we have used the first equation in (2.2). Since

$$\langle e,\overline{\vartheta'}\psi\rangle+e(a,t)\psi_a=\langle\vartheta(u)-\vartheta(\widetilde{u}),\psi\rangle+(u(a,t)-\widetilde{u}(a,t))\psi_a,\tag{3.14}$$

integration of (3.8) over I_n yields

$$\begin{aligned}\langle\vartheta(u_n)-\vartheta(\widetilde{u}_n^-),\varphi_n^-\rangle+\int_{I_n}\big(\langle\vartheta(u)-\vartheta(\widetilde{u}),\psi\rangle+(u-\widetilde{u}))(a,t)\psi_a\big)\,\mathrm{d}t\\ =\langle\vartheta(u_{n-1})-\vartheta(\widetilde{u}_{n-1}^-),\varphi_{n-1}^+\rangle-R_n(\widetilde{u};\widetilde{u},\varphi),\end{aligned}\tag{3.15}$$

where R_n is as expressed in (3.9). Next we sum up (3.15) over $n=1,\cdots,N$ and take advantage of the continuity of φ in t to get

$$[Q(u)](T)-[Q(\widetilde{u})](T)=\langle\vartheta(u_0)-\vartheta(\widetilde{u}_0^-),\varphi_0\rangle-\sum_{n=1}^{N}R_n(\widetilde{u};\widetilde{u},\varphi).\tag{3.16}$$

The residual $R_n(\widetilde{u};\widetilde{u},\varphi)$ can be decomposed into three components, which is obtained by adding and subtracting $R_{h,n}(\widetilde{u};\widetilde{u},\varphi)$:

$$R_n(\widetilde{u};\widetilde{u},\varphi)=R_{h,n}(\widetilde{u};\widetilde{u},\varphi)+\delta R_n(\widetilde{u};\widetilde{u},\varphi),\tag{3.17}$$

where

$$\begin{aligned}\delta R_n(\widetilde{u};\widetilde{u},w)&=R_n(\widetilde{u};\widetilde{u},w)-R_{h,n}(\widetilde{u};\widetilde{u},w)\\ &=\int_{I_n}\langle\partial_t\vartheta(\widetilde{u}),w-J_hw\rangle\,\mathrm{d}t+\langle[\vartheta(\widetilde{u})]_{n-1},(w-J_hw)^+_{n-1}\rangle\\ &\quad+\int_{I_n}(A(\widetilde{u};\widetilde{u},w)-A_h(\widetilde{u};\widetilde{u},w))\,\mathrm{d}t\\ &=\varepsilon_{h,n}(\widetilde{u};w)+\int_{I_n}\varepsilon_A(\widetilde{u};\widetilde{u},w)\,\mathrm{d}t\\ &=\varepsilon_{h,n}(\widetilde{u};w)+\varepsilon_{A,n}(\widetilde{u};\widetilde{u},w).\end{aligned}\tag{3.18}$$

Putting(3.18) to (3.17) and in turn to (3.16) completes the proof. □

Theorem 3.2. *For* $\widetilde{u}\in\mathcal{W}_h^q$ *in* (2.15) *and* u *in* (2.2),

$$[Q(u)](T)-[Q(\widetilde{u})](T)=\mathcal{E}_0+\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3,\tag{3.19}$$

where

$$\begin{aligned}\mathcal{E}_0 &= \langle \vartheta(u_0) - \vartheta(\widetilde{u}_0^-), \varphi_0 \rangle,\ \ \mathcal{E}_1 = -\sum_{n=1}^{N} R_n(\widetilde{u}; \widetilde{u}, \varphi - \pi_h^q \varphi),\\ \mathcal{E}_2 &= -\sum_{n=1}^{N} \varepsilon_{A,n}(\widetilde{u}; \widetilde{u}, \Pi_h^q \varphi),\ \ \mathcal{E}_3 = -\sum_{n=1}^{N} \varepsilon_{h,n}(\widetilde{u}; \pi_h^q \varphi),\end{aligned} \tag{3.20}$$

and $\pi_h^q \varphi \in \mathcal{W}_h^q$ *is the usual projection of* φ *onto* $\mathcal{W}_h^q$.

Proof. Most derivation steps follow the proof of Theorem 3.1 up to (3.16):

$$[Q(u)](T) - [Q(\widetilde{u})](T) = \langle \vartheta(u_0) - \vartheta(\widetilde{u}_0^-), \varphi_0 \rangle - \sum_{n=1}^{N} R_n(\widetilde{u}; \widetilde{u}, \varphi). \tag{3.21}$$

At this stage, we intend to insert (2.15), which is valid when the test function is $\pi_h^q \varphi \in \mathcal{W}_h^q$. To do so, add and subtract $R_n(\widetilde{u}; \widetilde{u}, \pi_h^q \varphi)$ so that

$$\begin{aligned} R_n(\widetilde{u}; \widetilde{u}, \varphi) &= R_n(\widetilde{u}; \widetilde{u}, \varphi - \pi_h^q \varphi) + R_n(\widetilde{u}; \widetilde{u}, \pi_h^q \varphi)\\ &= R_n(\widetilde{u}; \widetilde{u}, \varphi - \pi_h^q \varphi) + R_n(\widetilde{u}; \widetilde{u}, \pi_h^q \varphi) - R_{h,n}(\widetilde{u}; \widetilde{u}, \pi_h^q \varphi)\\ &= R_n(\widetilde{u}; \widetilde{u}, \varphi - \pi_h^q \varphi) + \varepsilon_{A,n}(\widetilde{u}; \widetilde{u}, \pi_h^q \varphi) + \varepsilon_{h,n}(\widetilde{u}; \pi_h^q \varphi), \end{aligned} \tag{3.22}$$

where similar equation to (3.18) has been used, and $\varepsilon_{A,n}$ and $\varepsilon_{h,n}$ are as in (3.12). Putting all these results back to (3.21) completes the proof. □

4. **Numerical examples.** Several numerical examples are presented in this section to achieve two goals: 1) to investigate the global/norm-based accuracy of the proposed approximation, and 2) to validate the robustness of error estimators that are derived from Theorem 3.1 and Theorem 3.2. While the former cannot satisfactorily substitute for a rigorous a priori error analysis, at least it should give an illustrative indicator on the global convergence property of the approximation. With respect to the latter, we only concentrate on the estimators' accuracy in predicting error and their capability to decompose it into relevant components. Various pertinent applications of the proposed error estimators to other aspects in numerical simulation of Richards Equation, such as its role in adaptivity, will be a subject of future work.

A uniform set of discretization parameters $h_i = h = (b-a)/M$ and $k_n = k = T/N$ is used to construct the algebraic equations (2.19). The time marching is executed by solving this system using the standard Newton's method of iteration.

4.1. **A problem with closed form solution.** While the proposed procedures enjoy a flexibility in their implementation, a closed form solution of Richards Equation is needed for the purpose of assessing their performance. As alluded to in the introduction, it is only on a very rare occasion that a closed form solution of Richards Equation is available. One such instance is when the constitutive relations are expressed as

$$\kappa(u) = \kappa_s e^{\alpha u},\ \text{ and } \vartheta(u) = \vartheta_r + (\vartheta_s - \vartheta_r) e^{\alpha u}, \tag{4.1}$$

where κ_s is the saturated hydraulic conductivity, ϑ_r and ϑ_s are respectively the residual and saturated water content, and α is the reciprocal of vertical height associated with the capillary fringe. When $g_\star(t) = g_\star = \text{constant}$, $\star = a, b$, then the closed form solution can be expressed as a series representation:

$$u(z,t) = \alpha^{-1} \ln(\kappa_s^{-1} w(z,t)), \tag{4.2}$$

where

$$w(z,t) = C_1 + C_2 e^{\alpha z} + e^{\alpha z/2} \sum_{n=1}^{\infty} w_n(t)\phi_n(z), \text{ with} \tag{4.3}$$

$$\begin{aligned} w_n(t) &= w_n(0)e^{-\mu_n t}, \ \mu_n = \frac{\kappa_s(\alpha^2 + 4\lambda_n^2)}{4\alpha(\vartheta_s - \vartheta_r)} > 0, \\ w_n(0) &= \frac{1}{\langle \phi_n, \phi_n \rangle} \int_0^L \left(\kappa_s e^{\alpha u_0(z)} - C_1 - C_2 e^{\alpha z}\right) e^{-\alpha z/2} \phi_n(z)\, \mathrm{d}z. \end{aligned} \tag{4.4}$$

The pair $\{\phi_n, \lambda_n\}_{n=1}^{\infty}$ constitutes an eigenfunction and an eigenvalue that satisfies

$$\begin{cases} -\phi_n'' = \lambda_n^2 \phi_n \text{ in } (a,b), \\ \tilde{\mathcal{B}}_a \phi_n = 0, \ \tilde{\mathcal{B}}_b \phi_n = 0, \end{cases} \tag{4.5}$$

where $\tilde{\mathcal{B}}_\star$ are boundary conditions for w, which are appropriately derived from $\mathcal{B}_\star$ via the relation $w(z,t) = \kappa_s e^{\alpha u(z,t)}$. The constants $\{C_1, C_2\}$ are obtained from imposing the boundary conditions $\mathcal{B}_\star u = g_\star$.

Two examples are considered in the numerical experiments whose data are listed in Table 4.1. Solution profiles of these examples are shown in Figure 4.1 and Figure 4.2. The axes on these figures are flipped to follow the plotting style for profiles associated with Richards Equation (see for example [25, 24]). The initial condition is

$$u_0(z) = \alpha^{-1} \ln(\kappa_s^{-1} f(z)), \tag{4.6}$$

where for Ex. 1,

$$\begin{aligned} f(z) &= C_1 + C_2 e^{\alpha z} + A e^{\alpha z/2} \sin(\lambda_1 z), \\ \lambda_1 &= \pi/b, \ C_2 = \frac{\kappa_s(e^{\alpha g_a} - e^{\alpha g_b})}{1 - e^{\alpha b}}, \ C_1 = \kappa_s e^{\alpha g_a} - C_2, \\ A &= \frac{4\lambda_1(\kappa_s e^{-\alpha(65+b/2)} - C_1 e^{-\alpha b/2} - C_2 e^{\alpha b/2})}{((\alpha/2)^2 + \lambda_1^2)b}, \end{aligned} \tag{4.7}$$

and for Ex. 2,

$$\begin{aligned} f(z) &= C_1 + C_2 e^{\alpha z} + e^{-\alpha(b-z)/2} \sum_{n=1}^{6000} A_n \sin(\lambda_n(b-z)), \\ &\lambda_n \text{ is governed by } \tan(\lambda_n b) + \frac{2\lambda_n}{\alpha} = 0, \\ C_1 &= -g_a, \ C_2 = \frac{\kappa_s e^{\alpha g_b} - C_1}{e^{\alpha b}}, \\ A_n &= \frac{\alpha \cosh(\alpha b/2)\sin(\lambda_n b) - 2\lambda_n \cos(\lambda_n b)\sinh(\alpha b/2)}{(\alpha/2)^2 + \lambda_n^2} \frac{4\lambda_n g_a}{2\lambda_n b - \sin(2\lambda_n)}. \end{aligned} \tag{4.8}$$

4.2. **An accuracy assessment of the approximation.** In this subsection, a set of numerical experiments to investigate accuracy of the approximation is presented. We solve the two examples whose data are listed in Table 4.1.

Table 4.2 and Table 4.3 list the errors of approximation $\vartheta(\widetilde{u}(T))$ in $L^2(a,b)$-norm for Ex. 1 and Ex. 2, respectively. Four different number of elements ($M = 12, 24, 48, 96$) and four different number of time steps ($N = 1, 2, 4, 8$) are used to collect the error data in Table 4.2, while for error data in Table 4.3, $M = 5, 10, 20, 40$ and $N = 4, 8, 16, 32$ are used.

	Ex. 1	Ex. 2
(a, b)	$(0, 60)$ cm	$(0, 2)$ m
T	1000 s	0.5 hr
$\mathcal{B}_a u$	$u(a, t)$	$\kappa(u)\partial_z(u - z)\big\|_{(a,t)}$
$\mathcal{B}_b u$	$u(b, t)$	$u(b)$
g_a	-65 cm	-0.15 m/hr
g_b	0 cm	0 m
u_0	(4.6) & (4.7)	(4.6) & (4.8)
α	0.01 cm^{-1}	4 m^{-1}
κ_s	0.001 cm/s	0.1 m/hr
θ_s	0.3	0.6
θ_r	0.08	0.02

Table 4.1: Data for all examples with closed form solution

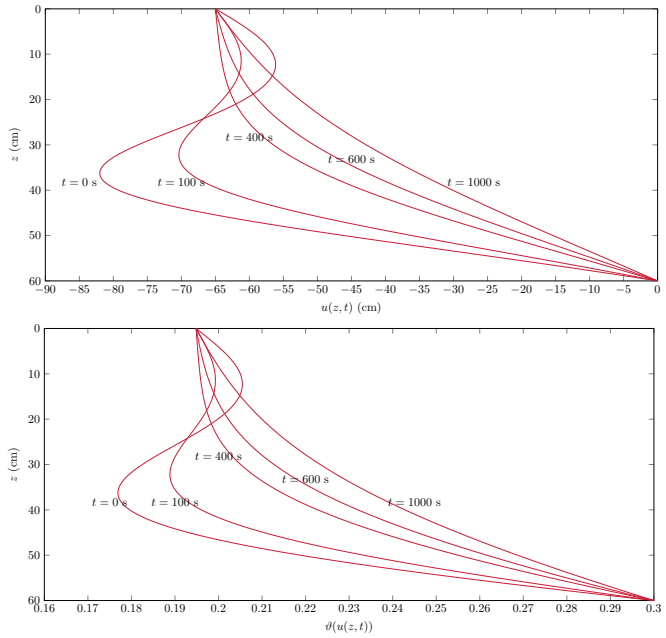

FIGURE 4.1. Ex. 1: $u(z, t)$ (top) and $\vartheta(u(z, t))$ (bottom)

First, the accuracy of FVEM-dG1 generally outperforms that of FVEM-dG0, which is especially evident when solving Ex. 1 (see Table 4.2). The approximation error for Ex. 1 seems to be dominated by component of the temporal discretization. For a fixed N, refining M does not quite improve the accuracy. However, for a fixed M, refining N by two roughly reduces the error by two for FVEM-dG0 and by seven to ten for FVEM-dG1, especially for larger M. A strikingly different result is observed for Ex. 2 (see Table 4.3), for which the spatial discretization error component is more dominant. For a fixed N, the error of FVEM-dG1 shows a quadratic convergence with respect to M. On the other hand, when N is still

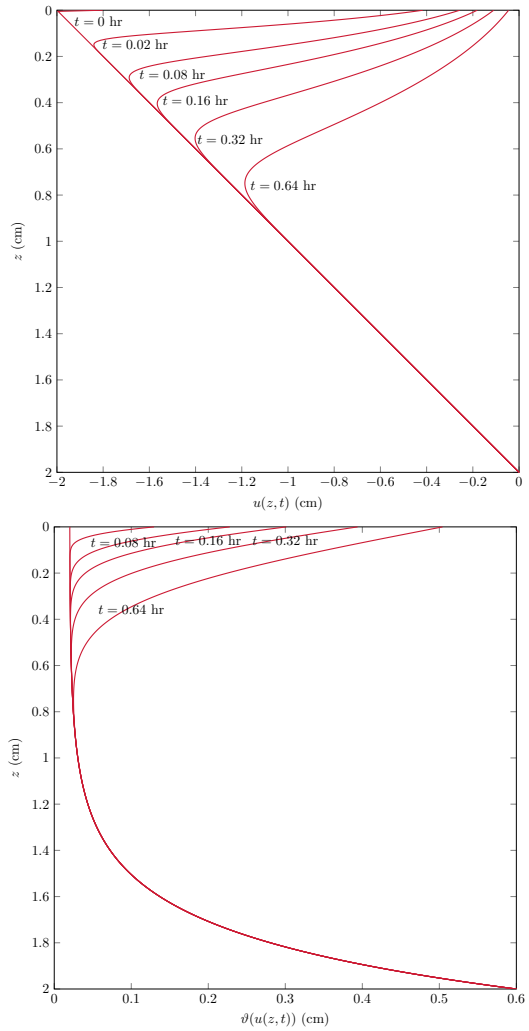

FIGURE 4.2. Ex. 2: $u(z,t)$ (top) and $\vartheta(u(z,t))$ (bottom)

small, the error of FVEM-dG0 resembles a first order convergence with respect to M. As N is increased, a better convergence rate is obtained.

4.3. **Performance of the a posteriori error estimators.** As mentioned, the error equation for a quantity of interest Q stated in Theorems 3.1 and 3.2 can be used to derive fully computable error estimators for the approximate solution $\widetilde{u}$. Notice that adjoint equation (3.8) is formulated based on linearization that utilizes Mean Value Theorem on the path $\widetilde{u}_\sigma = \widetilde{u}+\sigma(u-\widetilde{u})$, for $\sigma \in [0,1]$, cf. (3.2). Since in reality u is not available, calculation of solution to the adjoint equation (3.8) must be done using the only available information, namely, the approximate solution $\widetilde{u}$, so in practice, $\widetilde{u}_\sigma \approx \widetilde{u}$. The adjoint φ is approximated preferably using higher order approximation than the one used to produce $\widetilde{u}$.

To test the proposed error estimators, we consider two quantities of interest:

M	N	FVEM-dG0	FVEM-dG1
12	1	58.0082e-03	14.5267e-03
24	1	58.1606e-03	15.1200e-03
48	1	58.1991e-03	15.2688e-03
96	1	58.2087e-03	15.3060e-03
12	2	32.5818e-03	0.6939e-03
24	2	32.5477e-03	0.8597e-03
48	2	32.5396e-03	0.9086e-03
96	2	32.5376e-03	0.9211e-03
12	4	17.3786e-03	0.2823e-03
24	4	17.2600e-03	0.0916e-03
48	4	17.2310e-03	0.1222e-03
96	4	17.2238e-03	0.1344e-03
12	8	9.0728e-03	0.3586e-03
24	8	8.9160e-03	0.0793e-03
48	8	8.8779e-03	0.0154e-03
96	8	8.8685e-03	0.0153e-03

Table 4.2: Ex. 1: Error of $\vartheta(\widetilde{u}(T))$ quantified in $L^2(a,b)$-norm

M	N	FVEM-dG0	FVEM-dG1
5	4	4.9527e-02	5.0631e-02
10	4	2.0846e-02	1.8016e-02
20	4	1.0534e-02	0.4128e-02
40	4	0.8508e-02	0.0992e-02
5	8	4.9824e-02	5.0641e-02
10	8	1.8868e-02	1.7991e-02
20	8	0.6907e-02	0.4088e-03
40	8	0.4671e-02	0.0965e-02
5	16	5.0159e-02	5.0645e-02
10	16	1.8208e-02	1.7988e-02
20	16	0.5197e-02	0.4079e-02
40	16	0.2637e-02	0.0959e-02
5	32	5.0382e-02	5.0647e-02
10	32	1.8027e-02	1.7988e-02
20	32	0.4504e-02	0.4077e-02
40	32	0.1646e-02	0.0958e-02

Table 4.3: Ex. 2: Error of $\vartheta(\widetilde{u}(T))$ quantified in $L^2(a,b)$-norm

- The spatial average of water content at time T, which is represented as

$$[Q(u)](T) = \frac{1}{b-a}\int_a^b \vartheta(u(z,T))\,\mathrm{d}z. \tag{4.9}$$

To calculate the adjoint solution associated with this quantity, the corresponding adjoint data is $\psi_T = 1/(b-a)$, and the rest are zero.

- The total average of water content over $(a,b)\times(0,T)$, which is expressed as

$$[Q(u)](T)=\frac{1}{T}\int_0^T\frac{1}{b-a}\int_a^b\vartheta(u(z,t))\,\mathrm{d}z\,\mathrm{d}t. \tag{4.10}$$

The corresponding adjoint data is $\psi=1/(b-a)/T$ and the rest are zero.

The true values of these quantities of interest for the two examples are listed in Table 4.4.

	Ex. 1	Ex. 2
Q in (4.9)	0.231831624739998887	0.129975678959476710
Q in (4.10)	0.221196137487056291	0.111225678959476525

Table 4.4: True Value of Quantities of Interest

The approximate solution is $\widetilde{u}\in\mathcal{W}_h^0$ (FVEM-dG0). The adjoint solution is solved by continuous and piecewise quadratic finite element in spatial variable and continuous piecewise linear in temporal variable. Profiles of φ corresponding to each of these quantities of interest for each of the problems are shown in Figure 4.3 and Figure 4.4, respectively.

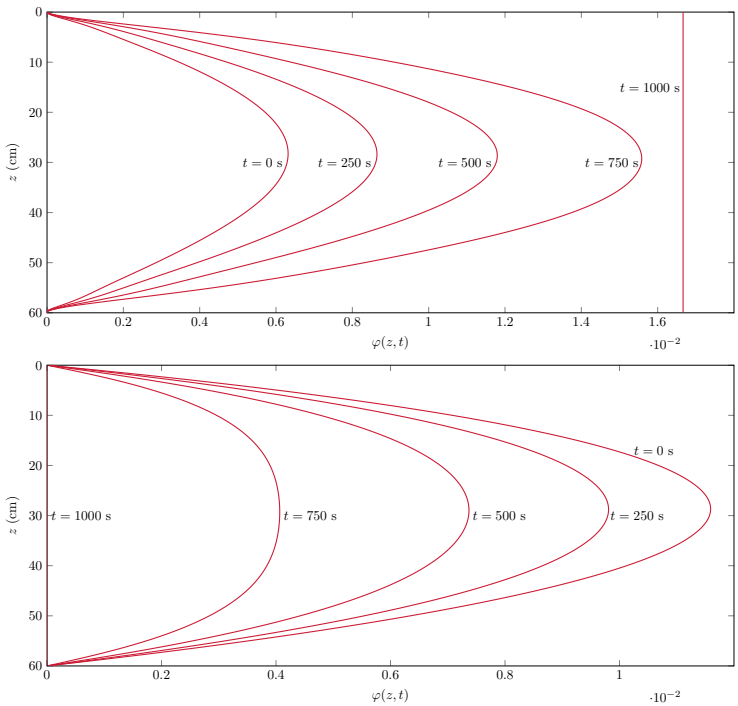

FIGURE 4.3. Ex. 1: $\varphi(z,t)$ for Q in (4.9) (top) and for Q in (4.10) (bottom). Each of them is obtained from numerical approximation of (3.8), with $\widetilde{u}\in\mathcal{W}_h^0$, $h=(b-a)/96$, and $k=T/8$.

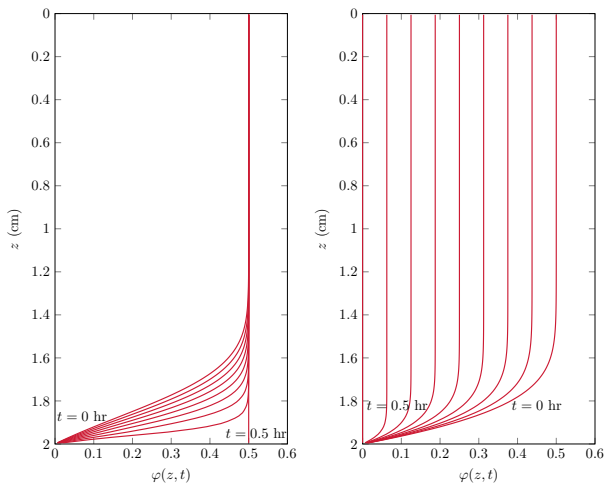

FIGURE 4.4. Ex. 2: $\varphi(z,t)$ for Q in (4.9) (left) and for Q in (4.10) (right). Each of them is obtained from numerical approximation of (3.8), with $\widetilde{u} \in \mathcal{W}_h^0$, $h = (b-a)/50$, and $k = T/8$.

Table 4.5 and Table 4.6 demonstrate performance of the error estimator for the above quantities of interest when solving Ex. 1. In these tables, the error has been decomposed according to the components of error listed in Theorem 3.1. As in the accuracy assessment results, four different number of elements ($M = 12, 24, 48, 96$) and four different number of time steps ($N = 1, 2, 4, 8$) are used. The last column, which is labeled by Eff. denotes the ratio of error estimator (Err. Est.) to the actual error, so the closer Eff. is to 1 indicates a more accurate error estimator.

Notice that values in the tables give an indication that the error in Q is dominated by the contribution from temporal discretization, as refinement of the spatial mesh for a fixed time step only causes negligible reduction in Err. Est. Reducing the time step by two seems to reduce the error by two, which demonstrates an asymptotic first order convergence with respect to the time step, i.e., Err. Est. $= \mathcal{O}(k)$. Prominence of temporal discretization effect to the total error makes sense due to the longer time simulation.

The component E_0 quantifies the quality of representation of the true initial condition u_0 in the simulation. Representation of u_0 is realized through the projection of u_0 onto $\mathcal{X}_h$. Thus, E_0 measures the discrepancy attributed to this choice, which shows a second order convergence with respect to h (i.e., $E_0 = \mathcal{O}(h^2)$) for a fixed time step. Comparison of values in the tables indicates that relative contribution of this component to the overall error is less significant.

The component E_1, which measures the residual of the finite volume element discretization weighted by the adjoint solution and integrated over time, shows a decrease with respect to time step as well. In fact, E_1 is clearly the main contributor to the total error with asymptotic behavior $E_1 = \mathcal{O}(k)$. The components E_2 and E_3 measures the discrepancy between the variational setting associated with finite volume element and that of standard continuous Galerkin finite element. In the realm of a priori error analysis, these two components are bounded in terms of the spatial mesh size h (see Proposition 2.1 and (2.5)). For every fixed N, there is a reduction of E_2 and E_3 as h is refined, with mostly $E_2 = \mathcal{O}(h)$ in Table 4.5, $E_2 =$

$\mathcal{O}(h^2)$ in Table 4.6, and $E_3 = \mathcal{O}(h^2)$ in both tables. However, these two components are less dominant in relative comparison to E_1. Notice also that E_0 has a different sign than the rest of components. A capability to distinguish components of error and to recognize potential cancellation is arguably one of the strong advantages of the adjoint-based error estimation techniques. Finally, as illustrated by the Eff., as refinement is performed, the error estimator gives a more accurate prediction.

Next we utilize the proposed error estimator to the numerical solution of Ex. 2. Table 4.7 and Table 4.8 show the breakdown of error for each of the quantities of interest. Again, following the accuracy assessment results, four different number of elements ($M = 5, 10, 20, 40$) and four different number of time steps ($N = 4, 8, 16, 32$) are used. As in the results for Ex. 1, the proposed estimator performs really well in predicting the error. However, upon a closer observation, the detailed situation is quite different from what happened in Ex. 1. In Table 4.7 (error associated with Q in (4.9)), we notice that the error caused by discretization of the spatial variable is more dominant, so refining the spatial mesh reduces the error. In particular, E_3 is seen to be the main contributor to the total error with asymptotic behavior $E_3 = \mathcal{O}(h^2)$, which results in Err. Est. $= \mathcal{O}(h^2)$. Notice also that time step refinement does not seem to improve the accuracy.

The result in Table 4.8 shows that E_1 and E_3 are two competing error components with $E_1 = \mathcal{O}(k)$ and $E_3 = \mathcal{O}(h^2)$. Due to their different sign, they tend to cancel each other, especially when M and N are still smaller. This in turn lowers the magnitudes of total error. However, as N is increased, E_3 tends to be more dominant than E_1, especially for small M. Since there is an intertwinement of the error components stemming from temporal and spatial variables, Err. Est. in this table does not indicate a clear pattern of convergence with respect to any discretization parameter. However, as the two discretizations are simultaneously refined, there is an observable reduction.

M	N	E_0	E_1	E_2	E_3	Err. Est.	Eff.
12	1	-5.5601e-05	6.6298e-03	2.6571e-04	2.2485e-05	6.8623e-03	1.149
24	1	-1.3843e-05	6.6974e-03	1.3440e-04	5.3567e-06	6.8233e-03	1.139
48	1	-3.4568e-06	6.7142e-03	6.7378e-05	1.3220e-06	6.7794e-03	1.131
96	1	-8.6396e-07	6.7184e-03	3.3706e-05	3.2943e-07	6.7515e-03	1.126
12	2	-4.4365e-05	3.8268e-03	7.1941e-05	3.6142e-05	3.8906e-03	1.120
24	2	-1.1046e-05	3.8548e-03	3.9509e-05	8.5556e-06	3.8918e-03	1.120
48	2	-2.7579e-06	3.8617e-03	2.0874e-05	2.1051e-06	3.8820e-03	1.120
96	2	-6.8914e-07	3.8635e-03	1.0730e-05	5.2410e-07	3.8740e-03	1.115
12	4	-3.9779e-05	2.0010e-03	2.5003e-05	4.3906e-05	2.0310e-03	1.067
24	4	-9.9212e-06	2.0128e-03	1.2165e-06	1.0291e-05	2.0254e-03	1.069
48	4	-2.4774e-06	2.0158e-03	6.9258e-06	2.5084e-06	2.0227e-03	1.069
96	4	-6.1911e-07	2.0165e-03	3.7657e-06	6.2266e-07	2.0203e-03	1.068
12	8	-3.8042e-05	1.0175e-03	1.9493e-05	4.6704e-05	1.0457e-03	1.037
24	8	-9.4764e-06	1.0227e-03	5.0975e-06	1.1390e-05	1.0297e-03	1.036
48	8	-2.3676e-06	1.0240e-03	2.3602e-06	2.7559e-06	1.0268e-03	1.036
96	8	-5.9172e-07	1.0244e-03	1.3885e-06	6.8063e-07	1.0258e-03	1.036

Table 4.5: Ex. 1: Performance of the Error Estimator in Theorem 3.1 for Q in (4.9)

M	N	E_0	E_1	E_2	E_3	Err. Est.	Eff.
5	4	-3.6102e-05	-6.2164e-08	-3.6456e-07	1.0290e-02	1.0253e-02	0.977
10	4	-1.3302e-05	3.3879e-09	-4.6917e-09	3.9111e-03	3.8978e-03	0.993
20	4	-5.7876e-06	9.9041e-10	-2.2824e-10	1.0090e-03	1.0032e-03	0.998
40	4	-2.7053e-06	5.6955e-10	-3.8568e-11	2.4965e-04	2.4695e-04	0.999
5	8	-3.6107e-05	-4.3218e-08	-3.5113e-07	1.1133e-02	1.1096e-02	0.974
10	8	-1.3302e-05	1.8737e-10	-4.7768e-10	4.3243e-03	4.3110e-03	0.993
20	8	-5.7876e-06	1.9063e-11	-6.0320e-12	1.0842e-03	1.0784e-03	0.998
40	8	-2.7053e-06	5.5338e-12	-4.5001e-13	2.6644e-04	2.6373e-02	0.999
5	16	-3.6121e-05	6.8028e-08	-4.5382e-07	1.1595e-02	1.1559e-02	0.973
10	16	-1.3302e-05	1.6801e-11	-8.8583e-11	4.5811e-03	4.5678e-03	0.992
20	16	-5.7876e-06	5.2323e-13	-3.2271e-13	1.1290e-03	1.1232e-03	0.998
40	16	-2.7053e-06	4.0884e-14	-2.5154e-14	2.7647e-04	2.7376e-04	0.999
5	32	-3.6089e-05	-4.4295e-08	4.5726e-08	1.1837e-02	1.1801e-02	0.972
10	32	-1.3302e-05	2.5330e-12	-2.9891e-11	4.7280e-03	4.7144e-03	0.992
20	32	-5.7876e-06	2.6583e-14	-9.4480e-14	1.1545e-03	1.1487e-03	0.998
40	32	-2.7053e-06	-1.0436e-14	-1.8777e-14	2.8211e-04	2.7941e-04	0.999

Table 4.7: Ex. 2: Performance of the Error Estimator in Theorem 3.1 for Q in (4.9)

M	N	E_0	E_1	E_2	E_3	Err. Est.	Eff.
12	1	-7.6641e-05	-5.9702e-03	6.5348e-06	2.7273e-05	-6.0131e-03	1.290
24	1	-1.9082e-05	-5.9818e-03	1.6059e-06	6.7751e-06	-5.9925e-03	1.290
48	1	-4.7653e-06	-5.9847e-03	3.9971e-07	1.6907e-06	-5.9874e-03	1.290
96	1	-1.1910e-06	-5.9854e-03	9.9817e-08	4.2248e-07	-5.9861e-03	1.290
12	2	-7.4772e-05	-3.0291e-03	1.6147e-05	2.5436e-05	-3.0623e-03	1.123
24	2	-1.8616e-05	-3.0342e-03	3.9483e-06	6.3250e-06	-3.0426e-03	1.123
48	2	-4.6488e-06	-3.0355e-03	9.8092e-07	1.5784e-06	-3.0376e-03	1.123
96	2	-1.1619e-06	-3.0358e-03	2.4483e-07	3.9441e-07	-3.0363e-03	1.123
12	4	-7.4397e-05	-1.5584e-03	2.3449e-05	2.4137e-05	-1.5852e-03	1.057
24	4	-1.8523e-05	-1.5609e-03	5.7189e-06	6.0235e-06	-1.5677e-03	1.057
48	4	-4.6257e-06	-1.5615e-03	1.4158e-06	1.5036e-06	-1.5632e-03	1.057
96	4	-1.1561e-06	-1.5617e-03	3.5295e-07	3.7571e-07	-1.5621e-03	1.057
12	8	-7.4326e-05	-7.9524e-04	2.7824e-05	2.3365e-05	-8.1837e-04	1.027
24	8	-1.8504e-05	-7.9647e-04	6.8634e-06	5.8679e-06	-8.0225e-04	1.027
48	8	-4.6210e-06	-7.9678e-04	1.6962e-06	1.4661e-06	-7.9824e-04	1.027
96	8	-1.1549e-06	-7.9686e-04	4.2208e-07	3.6638e-07	-7.9722e-04	1.027

Table 4.6: Ex. 1: Performance of the Error Estimator in Theorem 3.1 for Q in (4.10)

Tables 4.9 to 4.12 present the decomposition of error for Ex. 1 and Ex. 2 into various components as dictated by Theorem 3.2. The columns for $\mathcal{E}_0 = E_0$, Err. Est., and Eff. are not included since they are the same as in the corresponding columns in Tables 4.5 to 4.8, respectively. The component $\mathcal{E}_1$ seems to be the main contributor to the total error. For results associated with Ex. 1 (see Tables 4.9 and 4.10), the asymptotic behavior is roughly $\mathcal{E}_1 = \mathcal{O}(k)$. For Ex. 2 with Q as stated in (4.9) (see

M	N	E_0	E_1	E_2	E_3	Err. Est.	Eff.
5	4	-3.6102e-05	-4.6874e-03	-7.6435e-08	7.2999e-03	2.5763e-03	0.921
10	4	-1.3302e-05	-4.6875e-03	-8.4587e-10	3.3265e-03	-1.3743e-03	1.026
20	4	-5.7876e-06	-4.6875e-03	-3.4359e-11	9.8963e-04	-3.7036e-03	1.001
40	4	-2.7054e-06	-4.6875e-03	-4.2724e-12	2.5707e-04	-4.4331e-03	1.000
5	8	-3.6102e-05	-2.3437e-03	-4.6414e-08	7.2458e-03	4.8659e-03	0.950
10	8	-1.3302e-05	-2.3437e-03	-6.5517e-11	3.5887e-03	1.2316e-03	0.961
20	8	-5.7876e-06	-2.3437e-03	-7.0707e-13	1.0899e-03	-1.2596e-03	1.004
40	8	-2.7054e-06	-2.3437e-03	-4.9054e-14	2.8339e-04	-2.0631e-03	1.000
5	16	-3.6105e-05	-1.1719e-03	-7.9249e-08	7.1892e-03	5.9812e-03	0.956
10	16	-1.3302e-05	-1.1719e-03	-1.0129e-11	3.7626e-03	2.5774e-03	0.977
20	16	-5.7876e-06	-1.1719e-03	-6.6506e-14	1.1637e-03	-1.3957e-05	1.823
40	16	-2.7054e-06	-1.1719e-03	-1.1789e-14	3.0287e-04	-8.7171e-04	1.001
5	32	-3.6103e-05	-5.8593e-04	-2.1637e-08	7.1541e-03	6.5320e-03	0.959
10	32	-1.3302e-06	-5.8594e-04	-3.1392e-12	3.8643e-03	3.2650e-03	0.979
20	32	-5.7876e-06	-5.8594e-04	-4.6973e-14	1.2132e-03	6.2149e-04	0.988
40	32	-2.7054e-06	-5.8594e-04	-1.1491e-14	3.1628e-04	-2.7236e-04	1.002

Table 4.8: Ex. 2: Performance of the Error Estimator in Theorem 3.1 for Q in (4.10)

M	N	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$
12	1	6.9104e-03	7.5953e-06	0
24	1	6.8352e-03	1.8846e-06	6.9389e-18
48	1	6.7824e-03	4.6724e-07	3.4694e-18
96	1	6.7523e-03	1.1620e-07	6.9389e-18
12	2	3.8392e-03	6.1045e-06	8.9582e-05
24	2	3.8789e-03	1.5228e-06	2.2428e-05
48	2	3.8787e-03	3.7986e-07	5.6090e-06
96	2	3.8732e-03	9.4827e-08	1.4024e-06
12	4	1.9444e-03	5.6811e-06	1.1982e-04
24	4	2.0038e-03	1.4190e-06	3.0033e-05
48	4	2.0173e-03	3.5462e-07	7.5129e-06
96	4	2.0189e-03	8.8625e-08	1.8785e-06
12	8	9.4839e-04	5.5176e-06	1.2981e-04
24	8	1.0052e-03	1.3769e-06	3.2575e-05
48	8	1.0206e-03	3.4421e-07	8.1499e-06
96	8	1.0243e-03	8.6050e-08	2.0378e-06

Table 4.9: Ex. 1: Decomposition of Error according to Theorem 3.2 for Q in (4.9)

Table 4.11), the asymptotic behavior is $\mathcal{E}_1 = \mathcal{O}(h^2)$. However, it is not quite the case for Q in (4.10) (see Table 4.12).

5. **Conclusion and future work.** This paper investigates the application of adjoint-based a posteriori error analysis for numerical approximation of Richards

M	N	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$
12	1	-5.9364e-03	0	0
24	1	-5.9734e-03	0	0
48	1	-5.9826e-03	0	0
96	1	-5.9849e-03	0	0
12	2	-3.0501e-03	2.5710e-06	5.9998e-05
24	2	-3.0397e-03	6.4271e-07	1.5112e-05
48	2	-3.0369e-03	1.6068e-07	3.7850e-06
96	2	-3.0361e-03	4.0169e-08	9.4670e-07
12	4	-1.6024e-03	4.3468e-06	8.7251e-05
24	4	-1.5722e-03	1.0865e-06	2.1970e-05
48	4	-1.5644e-03	2.7161e-07	5.5023e-06
96	4	-1.5624e-03	6.7902e-08	1.3762e-06
12	8	-8.5030e-04	5.5588e-06	1.0069e-04
24	8	-8.1048e-04	1.3899e-06	2.5351e-05
48	8	-8.0031e-04	3.4750e-07	6.3488e-06
96	8	-7.9774e-04	8.6875e-08	1.5879e-06

Table 4.10: Ex. 1: Decomposition of Error according to Theorem 3.2 for Q in (4.10)

M	N	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$
5	4	1.0289e-02	-7.6351e-08	3.0166e-07
10	4	3.9111e-03	-5.9250e-10	4.6158e-09
20	4	1.0090e-03	-1.7676e-11	1.7694e-10
40	4	2.4965e-04	-1.9490e-12	2.1193e-11
5	8	1.1134e-02	7.3138e-08	-1.3466e-06
10	8	4.3243e-03	-6.0938e-11	1.1118e-09
20	8	1.0842e-03	-4.8331e-13	1.1840e-11
40	8	2.6644e-04	-2.7270e-14	6.6398e-13
5	16	1.1601e-02	8.7291e-08	-5.8431e-06
10	16	4.5811e-03	-1.1412e-11	3.7774e-10
20	16	1.1290e-03	-3.7491e-14	1.0044e-12
40	16	2.7647e-04	-4.7254e-15	1.7396e-14
5	32	1.1833e-02	2.1590e-08	3.9613e-06
10	32	4.7277e-03	-3.8970e-12	1.8867e-10
20	32	1.1545e-03	-1.9623e-14	1.5073e-13
40	32	2.8211e-04	-4.4201e-15	-2.7756e-17

Table 4.11: Ex. 2: Decomposition of Error according to Theorem 3.2 for Q in (4.9)

Equation. Construction of the approximate solution is cast into space-time variational formulation, specifically using the finite volume element in spatial variable and discontinuous Galerkin finite element in temporal variable. The resulting error estimators have the capability to predict components of error in certain quantities of interest that are expressed as a functional of the solution. The two examples give

M	N	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$
5	4	2.6123e-03	-7.7482e-09	1.1954e-07
10	4	-1.3610e-03	-5.0964e-11	1.2934e-09
20	4	-3.6979e-03	-1.4624e-12	4.4327e-11
40	4	-4.4304e-03	-1.6102e-13	5.1326e-12
5	8	4.9021e-03	6.9429e-09	-1.0375e-07
10	8	1.2450e-03	-5.4195e-12	2.1318e-10
20	8	-1.2539e-03	-4.3805e-14	1.8653e-12
40	8	-2.0604e-03	-3.6394e-15	9.6166e-14
5	16	6.0183e-03	1.5114e-08	-1.0262e-06
10	16	2.5907e-03	-1.0270e-12	5.4566e-11
20	16	-8.1695e-06	-9.7700e-15	1.0732e-13
40	16	-8.6900e-04	-2.1094e-15	1.2386e-15
5	32	6.5684e-03	3.4467e-09	-3.0913e-07
10	32	3.2783e-03	-3.6607e-13	2.2585e-11
20	32	6.2727e-04	-8.8020e-15	1.0911e-14
40	32	-2.6966e-04	-2.1545e-15	-4.1980e-16

Table 4.12: Ex. 2: Decomposition of Error according to Theorem 3.2 for Q in (4.10)

a strong indication that the error estimators are robust and capable to predict the error satisfactorily.

As for future work, we are interested in exploring further applications of the error estimators, in particular as to how they are applied to the setting of multidimensional problems. Owing to the various challenges persistent in the approximations of Richards Equation, a utilization of adaptivity is perhaps the only judicious route. Here the adaptivity is multi-faceted, not only as it pertains to local spatial refinement and dynamic time stepping, but also as it relates to determining optimal number of iterations when solving the nonlinear algebraic system. In this regard, the prospect of adjoint-based approach to estimate the components of error seems to be very promising.

Another interesting subject, which is not pursued in the present work, is a rigorous mathematical analysis of the proposed approximation. It must begin with establishing the existence of an approximate solution of (2.15). Here a potentially useful tool is either the Banach Fixed Point Theorem or the Brouwer Fixed Point Theorem. It should then be followed by a careful convergence analysis with the ultimate goal of showing the existence of a limit of the sequence of approximate solutions as $(h, k) \to (0, 0)$, and confirming that the limit satisfies a weak formulation of the Richards Equation. This can then be supplied with a study convergence rate of the approximate solution with respect to h and k.

REFERENCES

[1] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2003.

[2] V. Baron, Y. Coudière and P. Sochala, Adaptive multistep time discretization and linearization based on a posteriori error estimates for the Richards equation, *Appl. Numer. Math.*, **112** (2017), 104–125.

[3] M. Bause and P. Knabner, Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods, *Advances in Water Resources*, **27** (2004), 565–581.
[4] C. Bernardi, L. El Alaoui and Z. Mghazli, *A posteriori* analysis of a space and time discretization of a nonlinear model for the flow in partially saturated porous media, *IMA J. Numer. Anal.*, **34** (2014), 1002–1036.
[5] 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), 1483–1496.
[6] P. Chatzipantelidis, Finite volume methods for elliptic PDE's: A new approach, *M2AN Math. Model. Numer. Anal.*, **36** (2002), 307–324.
[7] P. Chatzipantelidis, V. Ginting and R. D. Lazarov, A finite volume element method for a non-linear elliptic problem, *Numer. Linear Algebra Appl.*, **12** (2005), 515–546.
[8] Z. Chen, G. Huan and Y. Ma, *Computational Methods for Multiphase Flows in Porous Media*, vol. 2 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
[9] S.-H. Chou and Q. Li, Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach, *Math. Comp.*, **69** (2000), 103–120.
[10] B. Cumming, T. Moroney and I. Turner, A mass-conservative control volume-finite element method for solving Richards'equation in heterogeneous porous media, *BIT*, **51** (2011), 845–864.
[11] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, *Computational Differential Equations*, Cambridge University Press, Cambridge, 1996.
[12] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, Introduction to adaptive methods for differential equations, in *Acta Numerica, 1995*, Acta Numer., Cambridge Univ. Press, Cambridge, 1995, 105–158.
[13] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, Introduction to computational methods for differential equations, in *Theory and Numerics of Ordinary and Partial Differential Equations (Leicester, 1994)*, Adv. Numer. Anal., IV, Oxford Univ. Press, New York, 1995, 77–122.
[14] D. Estep, A posteriori error bounds and global error control for approximation of ordinary differential equations, *SIAM J. Numer. Anal.*, **32** (1995), 1–48.
[15] D. J. Estep, M. G. Larson and R. D. Williams, Estimating the error of numerical solutions of systems of reaction-diffusion equations, *Mem. Amer. Math. Soc.*, **146** (2000), no. 696.
[16] R. Eymard, M. Gutnic and D. Hilhorst, The finite volume method for Richards equation, *Comput. Geosci.*, **3** (1999), 259–294.
[17] M. W. Farthing and F. L. Ogden, Numerical solution of Richards' equation: A review of advances and challenges, *Soil Science Society of America Journal*, **81** (2017), 1257–1269.
[18] W. R. Gardner, Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table, *Soil Science*, **85** (1958), 228–232.
[19] M. B. Giles and E. Süli, Adjoint methods for PDEs: A posteriori error analysis and postprocessing by duality, *Acta Numer.*, **11** (2002), 145–236.
[20] P. Jamet, Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain, *SIAM J. Numer. Anal.*, **15** (1978), 912–928.
[21] G. I. Marchuk, *Adjoint Equations and Analysis of Complex Systems*, vol. 295 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1995, Translated from the 1992 Russian edition by Guennadi Kontarev and revised by the author.
[22] F. Marinelli and D. S. Durnford, Semianalytical solution to Richards' equation for layered porous media, *Journal of Irrigation and Drainage Engineering*, **124** (1998), 290–299.
[23] L. A. Richards, Capillary conduction of liquids through porous mediums, *Physics*, **1** (1931), 318–333.
[24] R. Srivastava and T.-C. Jim Yeh, Analytical solutions for one-dimensional, transient infiltration toward the water table in homogeneous and layered soils, *Water Resour. Res.*, **27** (1991), 753–762.
[25] A. W. Warrick, A. Islas and D. O. Lomen, An analytical solution to Richards' equation for time-varying infiltration, *Water Resour. Res.*, **27** (1991), 763–766.

Received November 2020; revised April 2021.

E-mail address: vginting@uwyo.edu