Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(8): 24152–24178.
DOI:10.3934/math.2026975
Received: 31 May 2026
Revised: 29 July 2026
Accepted: 30 July 2026
Published: 07 August 2026

Research article

An upwind discontinuous finite volume element method on triangular meshes for the two-dimensional unsaturated soil water movement problem

Fang Wang, Fan Chen*, Changqing Lv and Xiaoyan Qin

Department of Mathematics and Statistics, Zaozhuang University, Zaozhuang 277160, China

* **Correspondence:** Email: chenfan00@126.com.

Abstract: The mathematical model for unsaturated soil water movement is essentially the Richards equation, which is a time-dependent nonlinear partial differential equation. This paper proposes an upwind discontinuous finite volume element method (DFVEM) on triangular meshes for solving two-dimensional unsaturated soil water flow problems. By employing an upwind strategy for the convection term, the method effectively avoids numerical dispersion and non-physical oscillations while preserving the local mass conservation and flexibility of the standard DFVEM. The error estimates are derived in the L^2 norm and $|||\cdot|||_{1,h}$ norm. Numerical experiments are conducted to verify the theoretical analysis and demonstrate the efficiency of the proposed method.

Keywords: unsaturated soil water movement; upwind discontinuous finite volume element method; error estimate
Mathematics Subject Classification: 65M12, 65M15, 65M60

1. Introduction

The unsaturated soil water movement problem refers to the flow of soil water when the soil is not completely saturated, and it is an important form of fluid movement in porous media. The prediction of unsaturated soil water movement holds significant importance in scientific and engineering applications such as water resources, petroleum reservoirs, and multiphase flow models. With the growing demand for describing, understanding, and predicting system dynamics of various complexity levels, the development of advanced mathematical models and solvers with precise predictive capabilities remains a long-term necessity.

In this paper, we consider the following unsaturated flow model. We assume that the soil is homogeneous. Let the x-axis be horizontal to the right and the z-axis is vertically downward. According to Darcy's law and the continuity principle, the problem of unsaturated soil water movement can be reduced to the following model [2]: For any $t > 0$, find the volumetric water content

$Q(x,z,t): \Omega \times (0,T] \to \mathbb{R}$ that satisfies

$$\frac{\partial Q}{\partial t} - \frac{\partial}{\partial x}(D(Q)\frac{\partial Q}{\partial x}) - \frac{\partial}{\partial z}(D(Q)\frac{\partial Q}{\partial z}) + \frac{\partial K(Q)}{\partial z} = S_r, (x,z) \in \Omega, t \in (0,T], \tag{1.1}$$

with the initial condition

$$Q(x,z,0) = Q_0,\ (x,z) \in \Omega,$$

and the boundary conditions

$$Q = 0,\ \text{on } \partial\Omega,$$

where Q_0 represents the initial moisture content and Q_s represents the saturated moisture content. $K(Q)$ represents the hydraulic conductivity, $D(Q)$ represents the diffusivity of soil water, and S_r indicates root water absorption. The volumetric moisture content Q satisfies $Q_r \leq Q(z,t) \leq Q_s$, where Q_r is the residual water content and $0 < Q_s < 1$. Additionally, $D(Q)$ and $K(Q)$ belong to $C^2([Q_r, Q_s])$, with uniformly bounded first- and second-order derivatives.

The mathematical model for unsaturated soil water movement, the Richards equation, is a time-dependent nonlinear partial differential equation. Based on the reliability of this problem and its practical significance in meteorology, agricultural environmental engineering, hydrodynamics, etc., in recent years, various numerical methods have been investigated, such as the finite difference method [9], finite element method (FEM) [6], discontinuous Galerkin method (DG) [10], etc. Li et al. [13] proposed generalized difference approaches for unsaturated soil water flow problems. Addressing computational scalability, Li and Zhou [14] developed mass-conservative domain decomposition strategies for multi-subdomain simulations. An optimized reduced implicit difference scheme based on singular value decomposition and proper orthogonal decomposition for the two-dimensional unsaturated soil water flow equation was introduced by Di et al. [11]. In addition to the development of spatial discretizations, the efficient solution of the nonlinear algebraic systems arising from the Richards equation has also received considerable attention. For instance, Liu et al. [18] proposed a parallel fully implicit solver based on Newton-Krylov methods with overlapping additive Schwarz preconditioners for variably saturated soil water flows. Their approach demonstrated good scalability on supercomputers, but relied on standard finite volume discretizations without upwind stabilization.

The discontinuous finite volume element method (DFVEM) was first proposed by Ye in [21] for the second-order elliptic equation. This method not only retains the advantages of the DG, but also ensures the local conservation of physical quantities and excellent parallel performance. Furthermore, Ye and Chou [8] improved the DFVEM and obtained the optimal convergence order. Subsequently, Bi [3] employed the DFVEM to solve parabolic equations. At present, the DFVEM has been applied to solve a variety of problems, including elliptic problems [17], parabolic problems [4, 15], Stokes problems [16, 22], and convection diffusion problems [5, 20].

For the specific case of the two-dimensional unsaturated flow problem, we previously proposed in [7] a standard DFVEM with a fully discrete error analysis. The numerical results demonstrate that the standard DFVEM achieves optimal convergence rates when the convection term is not dominant. However, when the convection term becomes dominant, the DFVEM exhibits significant deterioration in accuracy and convergence order. This degradation occurs because convection-dominated flows are known to cause numerical dispersion and non-physical oscillations in standard finite volume and finite element schemes [12]. These phenomena are particularly problematic in unsaturated flow applications involving sharp wetting fronts, where accurate front capture is essential.

To overcome this limitation, we introduce an upwind strategy into the DFVEM framework. The upwind flux is evaluated based on the sign of $b_e(Q) = \frac{\partial F}{\partial Q} \cdot \boldsymbol{n}_e$ at edge midpoints P_{jl}, ensuring that the numerical scheme selects the physically correct upstream information. This treatment guarantees stability and accuracy even for convection-dominated problems, effectively capturing sharp wetting fronts without spurious oscillations—a critical advantage over the standard DFVEM in [7]. Additionally, we provide new theoretical lemmas (Lemmas 3.8–3.10) that specifically address the consistency and stability of the upwind bilinear form, and we validate our method using the physically realistic van Genuchten-Mualem model.

The primary contributions of this work are threefold: (i) development of an upwind DFVEM on triangular meshes for the two-dimensional unsaturated flow problem, which overcomes the convection-dominated limitation of the standard DFVEM; (ii) derivation of a complete a priori error analysis for the fully discrete scheme, including new stability estimates for the upwind treatment of the convection term; (iii) numerical validation using the physically realistic van Genuchten-Mualem model, including demonstration of the method's ability to capture sharp wetting fronts without oscillations.

The outline of this paper is as follows: In Section 2, we construct the upwind discontinuous finite volume element scheme. After the definition of some projection operators, some lemmas are presented in Section 3. In Section 4, an error analysis for the full discrete scheme is derived. Numerical experiments are presented in Section 5 to validate the accuracy and efficiency of our proposed method.

2. Upwind discontinuous finite volume element scheme

In this section, we present the variational formulation of the model problem (1.1). Subsequently, a fully discrete upwind discontinuous finite volume element scheme is constructed.

First, some notations and definitions are given. Consider the standard Sobolev space $H^s(\Omega)$ for $0 \leq s < \infty$. Let $(\cdot,\cdot)_s$ and $\|\cdot\|_s$ denote the inner product and norm on $H^s(\Omega)$, respectively. When $s = 0$, $H^0(\Omega)$ coincides with the space of square-integrable functions $L^2(\Omega)$, $(\cdot,\cdot)_0 = (\cdot,\cdot)$ and $\|\cdot\|_0 = \|\cdot\|$. Define $H_0^1(\Omega)$ as the subspace of $H^1(\Omega)$ comprising functions with vanishing trace on $\partial\Omega$ and $\|\cdot\|_\infty$ as the norm in $L^\infty(\Omega)$. Throughout this article, C denotes a generic positive constant that may not necessarily be the same at each occurrence.

We assume that $\Omega \subset R^2$ is a suitably smooth and sufficiently large bounded domain. On the boundary $\partial\Omega$ of Ω, the initial moisture content Q_0 remains unchanged. As shown in Figure 1, we define the dual partition T_h^* of T_h for the test function space. Let T_h be a triangulation of Ω, and $h(K)$ represents the side length of the triangular element $K \in T_h$. Γ denotes the set of all the element edges of the original split T_h, $\Gamma^0 = \Gamma\backslash\partial\Omega$, and $h = \max\limits_{K\in T_h} h(K)$. We divide each $K \in T_h$ into three blocks by connecting the barycenter and the three corners of the triangle. Let T_h^* consist of all these triangles T_1, T_2, and T_3.

Define the following broken Sobolev space on the original partition T_h:

$$H^m(T_h) = \{v \in L^2(\Omega) : v|_K \in H^m(K), \forall K \in T_h\},$$

$$W^{m,\infty}(T_h) = \{w \in L^2(\Omega) : w|_K \in W^{m,\infty}(K), \forall K \in T_h\},$$

and its norm

$$\|v\|_{m,h} = (\sum_{i=0}^{m} |v|_{i,h}^2)^{1/2}, |v|_{i,h} = (\sum_{K\in T_h} |v|_{H^i(K)}^2)^{1/2},$$

$$\|v\|_{m,\infty,h} = \sup_{K\in T_h} \|v\|_{m,\infty,K},$$

where $H^i(K)$, $W^{m,\infty}(K)$ is the standard Sobolev space defined on element K, and m is a positive integer.

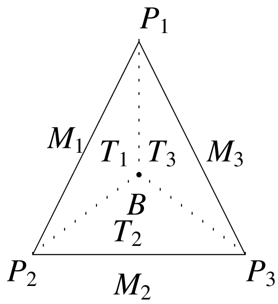

Figure 1. Original subdivision and dual subdivision.

Define a finite-dimensional trial function space on the original partition T_h,

$$U_h = \{u_h \in L^2(\Omega) : u_h|K \in P_1(K), \forall K \in T_h\},$$

and the finite-dimensional test function space on the dual partition T_h^*

$$V_h = \{v_h \in L^2(\Omega) : v_h|T \in P_0(T), \forall T \in T_h^*\},$$

where $P_l(K)$ denotes a polynomial set whose number of times defined on element K is less than or equal to l ($l = 0, 1$).

Let $U(h) = U_h + H^2(\Omega) \cap H_0^1(\Omega)$, and define a mapping [21] $\gamma_h : U(h) \to V_h$,

$$\gamma_h v_h|_T = \frac{1}{h_e} \int_e v_h|_T ds, \quad \forall T \in T_h^*,$$

where h_e is the length of the edge e.

To facilitate the theoretical analysis, take $Q_0 = 0$. Let $F(Q) = \mathbf{p} \cdot K(Q)$, and we have

$$\frac{\partial K(Q)}{\partial z} = \nabla \cdot F(Q), \quad \mathbf{p} = (0, 1).$$

So, (1.1) can be written as

$$\frac{\partial Q}{\partial t} - \frac{\partial}{\partial x}(D(Q)\frac{\partial Q}{\partial x}) - \frac{\partial}{\partial z}(D(Q)\frac{\partial Q}{\partial z}) + \nabla \cdot F(Q) = S_r. \tag{2.1}$$

Multiplying (2.1) by $v_h \in V_h$, integrating over the dual element, then applying Green's formula, we obtain

$$(\frac{\partial Q}{\partial t}, v_h) - \sum_{T\in T_h^*} \int_{\partial T} D(Q)\nabla Q \cdot \mathbf{n} v_h \, \mathrm{d}s + \sum_{T\in T_h^*} \int_{\partial T} F(Q) \cdot \mathbf{n} v_h \, \mathrm{d}s = (S_r, v_h), \tag{2.2}$$

where $\mathbf{n}$ is the element outside normal of the boundary ∂T of the dual element $T \in T_h^*$, and

$$\sum_{T\in T_h^*}\int_{\partial T} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s=\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s+\sum_{K\in T_h}\int_{\partial T} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s, \tag{2.3}$$

$$\sum_{T\in T_h^*}\int_{\partial T} F(Q)\cdot\mathbf{n}v_h\,\mathrm{d}s=\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} F(Q)\cdot\mathbf{n}v_h\,\mathrm{d}s+\sum_{K\in T_h}\int_{\partial T} F(Q)\cdot\mathbf{n}v_h\,\mathrm{d}s, \tag{2.4}$$

where $P_4 = P_1, P_5 = P_2$, and $P_6 = P_3$.

Assume $e = \partial K_1 \bigcap \partial K_2$, where K_1, K_2 are two elements of T_h, $\mathbf{n}_1, \mathbf{n}_2$ are the element normal vectors of the boundary e pointing to K_1, K_2, respectively. For a scalar function p and vector function $\mathbf{q}$, we define the average $\{\cdot\}$ and the jump $[\cdot]$ on e as follows [1]:

If $e \in \Gamma_0$ and $e \subset \partial K_1 \cap \partial K_2$, then

$$[v]|_e = v|_{\partial K_1}\mathbf{n}_1 + v|_{\partial K_2}\mathbf{n}_2,\;\; \{v\}|_e = \frac{1}{2}(v|_{\partial K_1} + v|_{\partial K_2}),$$

$$[\mathbf{w}]|_e = \mathbf{w}|_{\partial K_1}\cdot\mathbf{n}_1 + \mathbf{w}|_{\partial K_2}\cdot\mathbf{n}_2,\;\; \{\mathbf{w}\}|_e = \frac{1}{2}(\mathbf{w}|_{\partial K_1} + \mathbf{w}|_{\partial K_2}).$$

If $e \in \Gamma\backslash\Gamma_0$ and $e \subset \partial K$, then

$$[v]|_e = v|_{\partial K}\mathbf{n}_k,\;\; \{v\}|_e = v|_{\partial K},$$

$$[\mathbf{w}]|_e = \mathbf{w}|_{\partial K}\cdot\mathbf{n}_k,\;\; \{\mathbf{w}\}|_e = \mathbf{w}|_{\partial K}.$$

According to the definition of the jump and average, we have

$$\sum_{K\in T_h}\int_{\partial K} v\mathbf{w}\cdot\mathbf{n}ds=\sum_{e\in\Gamma}\int_e[v]\cdot\{\mathbf{w}\}ds+\sum_{e\in\Gamma_0}\int_e\{v\}[\mathbf{w}]ds. \tag{2.5}$$

By (2.5), notice that $[D(Q)\nabla Q\cdot\mathbf{n}v_h]\mid_e= 0$, $\forall e \in \Gamma_0$, thus we have

$$\sum_{K\in T_h}\int_{\partial K} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s=\sum_{e\in\Gamma}\int_e\{D(Q)\nabla Q\}\cdot[v_h]\,\mathrm{d}s. \tag{2.6}$$

Then, (2.3) becomes

$$\sum_{T\in T_h^*}\int_{\partial T} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s=\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s+\sum_{e\in\Gamma}\int_e\{D(Q)\nabla Q\}\cdot[v_h]\,\mathrm{d}s. \tag{2.7}$$

Using (2.7), (2.2) can be written as

$$\begin{aligned}(\frac{\partial Q}{\partial t},v_h)&-\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} D(Q)\nabla Q\cdot\mathbf{n}v_h\,\mathrm{d}s-\sum_{e\in\Gamma}\int_e\{D(Q)\nabla Q\}\cdot[v_h]\,\mathrm{d}s\\&+\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} F(Q)\cdot\mathbf{n}v_h\,\mathrm{d}s+\sum_{K\in T_h}\int_{\partial K} F(Q)\cdot\mathbf{n}v_h\,\mathrm{d}s=(S_r,v_h).\end{aligned} \tag{2.8}$$

Define the bilinear form

$$a(w;u,v) = -\sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} D(w)\nabla u\cdot \mathbf{n}v\,\mathrm{d}s - \sum_{e\in\Gamma}\int_e \{D(w)\nabla u\}\cdot[v]\,\mathrm{d}s,$$
$$b(u,v) = \sum_{K\in T_h}\int_{\partial K} F(u)\cdot \mathbf{n}v\,\mathrm{d}s,$$
$$d_h(u,v) = \sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i} F(u)\cdot \mathbf{n}v\,\mathrm{d}s.$$

Now, we can obtain the variational form of the problem (1.1): Seek $Q:[0,T]\to H_0^1(\Omega)$ such that

$$\begin{cases} (a)(\dfrac{\partial Q}{\partial t}, v_h) + a(Q;Q,v_h) + b(Q,v_h) + d_h(Q,v_h) = (S_r, v_h), \forall v_h\in V_h, \\ (b)Q(x,z,0)=0. \end{cases} \tag{2.9}$$

According to the definition of the transfer operator γ_h, we know the relationship between the trial function space U_h and the test function space V_h: $\gamma_h U_h = V_h$.

So, (2.9) can be rewritten as

$$\begin{cases} (a)(\dfrac{\partial Q}{\partial t}, \gamma_h v_h) + a(Q;Q,\gamma_h v_h) + b(Q,\gamma_h v_h) + d_h(Q,\gamma_h v_h) = (S_r, \gamma_h v_h), \quad \forall v_h\in V_h, \\ (b)Q(x,z,0)=0. \end{cases} \tag{2.10}$$

Next, we introduce the upwind term of $b(Q,v_h)$. As shown in Figure 2,

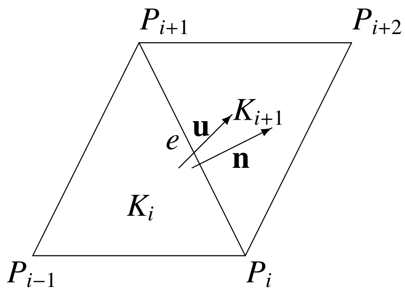

Figure 2. Schematic diagram of the upwind approach.

Consider $K(0)=0$, i.e., $F(Q)=\mathbf{0}$. Rewrite $F(Q)$ [12] to

$$F(Q) = \int_0^Q \frac{\partial F(\overline{Q})}{\partial \overline{Q}}\,\mathrm{d}\overline{Q}.$$

Use P_{jl} to indicate the midpoint of the adjacent node P_j and P_l. $\mathbf{n}_e$ represents the element outside normal of element K,

$$\beta_{jl}^{+}(Q) = \int_0^Q max(0, \frac{\partial F(\overline{Q})}{\partial \overline{Q}}\cdot \mathbf{n}_e)\,\mathrm{d}\overline{Q},$$
$$\beta_{jl}^{-}(Q) = \int_0^Q max(0, -\frac{\partial F(\overline{Q})}{\partial \overline{Q}}\cdot \mathbf{n}_e)\,\mathrm{d}\overline{Q}.$$

For $Q_h, v_h \in U_h$, we get

$$b_h(Q_h, \gamma_h v_h) = \sum_{e\in\Gamma} [\gamma_h v_h] \int_e [\beta^+_{jl}(Q_h(P_{jl})) - \beta^-_{jl}(Q_h(P_{jl}))]\, \mathrm{d}s.$$

Denote

$$b_e(Q) = \frac{\partial F(\overline{Q})}{\partial \overline{Q}} \cdot \mathbf{n}_e.$$

Let $H(r)$ be the Heaviside function, satisfying

$$H(r) = \begin{cases} 1 & \text{when } r \geq 0\,, \\ 0 & \text{when } r < 0\,. \end{cases}$$

Then there is

$$b_h(Q_h, \gamma_h v_h) = \sum_{e\in\Gamma_0} [[\gamma_h v_h]] \int_e [H(b_e(Q_h)) F(Q_h(P_{jl}))|_{K_1} \cdot \mathbf{n}_e + (1 - H(b_e(Q))) F(Q_h(P_{jl}))|_{K_2} \cdot \mathbf{n}_e]\, \mathrm{d}s.$$

Now we can get the semi-discrete upwind discontinuous finite volume element scheme of the problem (1.1): Seek $Q_h : [0, T] \to U_h$ such that

$$\begin{cases} (a)\ (\dfrac{\partial Q_h}{\partial t}, \gamma_h v_h) + A(Q_h; Q_h, \gamma_h v_h) + b(Q_h, \gamma_h v_h) + d_h(Q_h, \gamma_h v_h) = (S_r, \gamma_h v_h), \quad \forall v_h \in U_h, \\ (b)\ Q_h(x, z, 0) = 0, \end{cases} \tag{2.11}$$

where

$$\begin{aligned} A(Q_h; Q_h, \gamma_h v_h) = &- \sum_{K\in T_h} \sum_{i=1}^{3} \int_{P_{i+1}BP_i} D(Q_h)\nabla Q_h \cdot \mathbf{n}\gamma_h v_h ds - \sum_{e\in\Gamma} \int_e \{D(Q_h)\nabla Q_h\}[\gamma_h v_h] ds \\ &- \sum_{e\in\Gamma} \int_e \{D(Q_h)\nabla v_h\}[\gamma_h Q_h] ds + \alpha \sum_{e\in\Gamma} [\gamma_h Q_h][\gamma_h v_h] \end{aligned}$$

and α is a penalty parameter.

Because Q is the solution of the problem (1.1),

$$[\gamma_h Q]|_e = 0, \quad a(Q; Q, \gamma_h v_h) = A(Q; Q, \gamma_h v_h).$$

Therefore,

$$\begin{cases} (a)\ (\dfrac{\partial Q}{\partial t}, \gamma_h v_h) + A(Q; Q, \gamma_h v_h) + b(Q, \gamma_h v_h) + d_h(Q, \gamma_h v_h) = (S_r, \gamma_h v_h),\ \forall v_h \in U_h, \\ (b)\ Q(x, z, 0) = 0. \end{cases} \tag{2.12}$$

Next, we present the fully discrete upwind discontinuous finite volume scheme. Let $\Delta t = \dfrac{T}{N}$ be the time step, where N is a positive integer. $t^j = j\Delta t$. Replace $\partial_t Q_h^j = \dfrac{Q_h^j - Q_h^{j-1}}{\Delta t}$ with $\dfrac{\partial Q_h}{\partial t}$. By employing the backward Euler method for temporal discretization, we obtain the fully discrete upwind discontinuous finite volume element scheme: Seek $Q_h^n \in U_h$, $(n = 0, 1, \cdots, N)$ such that

$$\begin{cases} (a)\ (\dfrac{Q_h^n - Q_h^{n-1}}{\Delta t}, \gamma_h v_h) + A(Q_h^n; Q_h^n, \gamma_h v_h) + b(Q_h^n, \gamma_h v_h) + d_h(Q_h^n, \gamma_h v_h) = (S_r^n, \gamma_h v_h),\ \forall v_h \in U_h, \\ (b)\ Q_h(x, z, 0) = 0. \end{cases} \tag{2.13}$$

Remark 2.1. *The upwind strategy introduced here is essential for problems where the convection term $\frac{\partial K(Q)}{\partial z}$ is dominant. In such cases, standard finite volume schemes without upwinding may produce spurious oscillations and fail to capture sharp fronts accurately. The upwind flux evaluation based on the sign of $b_e(Q) = \frac{\partial F(\overline{Q})}{\partial \overline{Q}} \cdot \mathbf{n_e}$ ensures that information propagates in the correct physical direction, thereby enhancing the stability of the scheme. This is a key improvement over the standard DFVEM in [7], which does not employ upwinding and consequently suffers from accuracy degradation in convection-dominated regimes.*

3. Some lemmas

First, we give the definition of the discrete norm [21]:

$$|||u|||_{1,h}^2 = |u|_{1,h}^2 + \sum_{e\in\Gamma} [\gamma_h u]_e^2 + \sum_{K\in T_h} h_K^2 |u|_{2,K}^2,\ \forall u \in U(h),$$

where $|u|_{1,h}^2 = \sum\limits_{K\in T_h} |u|_{1,K}^2$.

The following trace inequality can be found in [1]. If e is the edge of K with length h_e, then

$$||u||_e^2 \le C(h_e^{-1}||u||_K^2 + h_e|u|_{1,K}^2), \forall u \in H^2(K). \tag{3.1}$$

Lemma 3.1 ([21]). *The operator γ_h is self-adjoint with respect to the L^2-inner product, that is,*

$$(u_h, \gamma_h v_h) = (v_h, \gamma_h u_h),\ \ \forall u_h, v_h \in U_h.$$

We define

$$|||u_h|||_0 = (u_h, \gamma_h u_h)^{1/2},$$

then $|||\cdot|||_0$ and $||\cdot||$ are equivalent, and

$$||\gamma_h u_h|| = ||u_h||,\ \ \forall u_h \in U_h.$$

Lemma 3.2 ([3]). *The operator γ_h satisfies the following properties:*

$$\int_K (w_h - \gamma_h w_h) dx = 0,\ \forall w_h \in U(h),\ \forall K \in T_h,$$

$$\int_e (w_h - \gamma_h w_h) ds = 0,\ \forall w_h \in U(h),\ \forall e \in \partial K,\ \forall K \in T_h,$$

$$[w_h] = 0 \Longrightarrow\ [\gamma_h w_h] = 0,\ \forall w_h \in U(h),$$

$$||\gamma_h w_h - w_h||_{0,K} \le C h_K |w_h|_{1,K},\ \forall w_h \in U(h),\ \forall K \in T_h.$$

Lemma 3.3 ([21]). *There exists a positive constant β independent of h such that*

$$\beta|||u_h|||_{1,h}^2 \leq A(q; u_h, \gamma_h u_h), \quad \forall u_h \in U_h.$$

Lemma 3.4 ([1]). *For $\forall u_h, v_h \in U_h$, there exists a positive constant C independent of h, so we have*

$$A(q; u_h, \gamma_h v_h) \leq C|||u_h|||_{1,h}|||v_h|||_{1,h}.$$

Lemma 3.5 ([21]). *There exists a positive constant C independent of h such that*

$$|A(q; u_h, \gamma_h v_h) - A(q; v_h, \gamma_h u_h)| \leq Ch|||u_h|||_{1,h}|||v_h|||_{1,h}, \quad \forall u_h \in U_h,$$

$$|A(p; u_h, \gamma_h v_h) - A(q; u_h, \gamma_h v_h)| \leq C|u_h|_{1,\infty}(||p - q|| + h|||p - q|||_{1,h}))|||v_h|||_{1,h}, \quad \forall u_h \in U_h.$$

Lemma 3.6 ([3]). *There exists a positive constant C independent of h such that*

$$h|||u_h|||_{1,h} \leq C||u_h||, \quad \forall u_h \in U_h,$$

$$||u_h|| \leq C|||u_h|||_{1,h}, \quad \forall u_h \in U_h.$$

To obtain the error estimates, we use Ritz-Galerkin projections of the solution of the steady state problem corresponding to (2.9). Given $Q \in H^1(\Omega)$, we introduce the projection $R_h Q : [0, T] \to U_h$ of Q onto U_h satisfying

$$A(Q; Q, \gamma_h v_h) = A(Q; R_h Q, \gamma_h v_h), \qquad \forall v_h \in U_h. \tag{3.2}$$

According to the relevant theory of Ritz projection R_h, the following interpolation properties are obtained.

$$\begin{cases} (a) \quad ||Q - R_h Q|| \leq Ch^2||Q||_3, \\ (b) \quad ||(Q - R_h Q)_t|| \leq Ch^2||Q||_{1,3,2}, \\ (c) \quad |||Q - R_h Q|||_{1,h} \leq Ch||Q||_2, \\ (d) \quad |||(Q - R_h Q)_t|||_{1,h} \leq Ch||Q||_{1,2,2}. \end{cases} \tag{3.3}$$

Lemma 3.7. *If $Q, Q_t \in W^{2,\infty}(\Omega) \bigcap H^3(\Omega)$, there exist positive constants C_0, C_1 independent of h such that*

$$|R_h Q|_{1,\infty} \leq C_0,$$

$$|R_h Q_t|_{1,\infty} \leq C_1.$$

Proof. Let $Q_I \in U_h$ be the interpolation of Q. It is well known that

$$|Q - Q_I|_{s,p,k} \leq Ch^{2-s}|Q|_{2,p,k}, \qquad \forall K \in T_h, s = 0, 1, 1 \leq p \leq \infty. \tag{3.4}$$

Based on the definition of the energy norm, we have

$$|||Q - Q_I|||_{1,h} \leq Ch|Q|_2. \tag{3.5}$$

Using Lemmas 3.3 and 3.4, we have

$$\beta|||R_h Q - Q_I|||_{1,h}^2 \leq A(Q; R_h Q - Q_I, \gamma_h(R_h Q - Q_I))$$

$$
\begin{aligned}
&= A(Q; Q - Q_I, \gamma_h(R_h Q - Q_I)) \\
&\le C |||Q - Q_I|||_{1,h} \cdot |||R_h Q - Q_I|||_{1,h}.
\end{aligned}
$$

Using triangle inequalities, inverse inequalities, and (3.4), we get

$$
|R_h Q|_{1,\infty} \le |R_h Q - Q_I|_{1,\infty} + |Q_I - Q|_{1,\infty} + |Q|_{1,\infty}.
$$

From (3.5),

$$
|||R_h Q - Q_I|||_{1,h} \le Ch|Q|_2.
$$

So,

$$
|R_h Q|_{1,\infty} \le C(|Q|_2 + |Q|_1).
$$

Similarly, we can prove $|R_h Q_t|_{1,\infty} \le C_1$. □

Lemma 3.8. *There exists a positive constant C independent of h such that*

$$
|b_h(Q, \gamma_h v_h) - b(Q, \gamma_h v_h)| \le Ch |||v_h|||_{1,h} |Q|_{1,h}, \quad \forall v_h \in U_h.
$$

Proof. Due to $[[\gamma_h v_h]] \cdot \mathbf{n}_e = [\gamma_h v_h]$, for any $e \in \partial\Omega$, we have $Q = Q_0 = 0$, and by the continuity of the true solution, we know that

$$
F(Q) = F(Q), [F(Q)] = 0, \forall e \in \partial\Gamma_0.
$$

According to (2.11) and the definition of bilinear form $b_h(Q, \gamma_h v_h)$, we have

$$
\begin{aligned}
b(Q, \gamma_h v_h) &= \sum_{K \in T_h} \int_{\partial K} F(Q) \cdot \mathbf{n} \gamma_h v_h ds \\
&= \sum_{e \in \Gamma} \int_e \{F(Q)\}[\gamma_h v_h] ds + \sum_{e \in \Gamma_0} \int_e [F(Q)]\{\gamma_h v_h\} ds \\
&= \sum_{e \in \Gamma_0} \int_e \{F(Q)\}[\gamma_h v_h] ds, \\
b_h(Q, \gamma_h v_h) &= \sum_{e \in \Gamma_0} [[\gamma_h v_h]] \int_e [H(be(Q_h)) F(Q_h(P_{jl}))|_{K_1} \cdot \mathbf{n}_e + (1 - H(be(Q))) F(Q_h(P_{jl}))|_{K_2} \cdot \mathbf{n}_e] \, \mathrm{d}s \\
&= \sum_{e \in \Gamma_0} \int_e [\gamma_h v_h][H(be(Q_h)) F(Q_h(P_{jl}))|_{K_1} \cdot \mathbf{n}_e + (1 - H(be(Q))) F(Q_h(P_{jl}))|_{K_2} \cdot \mathbf{n}_e] \, \mathrm{d}s.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality, trace inequality, and linear interpolation properties, we obtain

$$
\begin{aligned}
& b_h(Q, \gamma_h v_h) - b(Q, \gamma_h v_h) \\
&= \sum_{e \in \Gamma_0} \int_e [\gamma_h v_h][H(be(Q)) F(Q(P_{jl}))|_{K_1} \cdot \mathbf{n}_e + (1 - H(be(Q))) F(Q(P_{jl}))|_{K_2} \cdot \mathbf{n}_e - F(Q)] \, \mathrm{d}s \\
&\le \sum_{e \in \Gamma_0} \int_e [\gamma_h v_h][F(Q(P_{jl})) - F(Q)] \, \mathrm{d}s \\
&= \sum_{e \in \Gamma_0} \int_e [\gamma_h v_h] \int_{Q(x)}^{Q(P_{jl})} \frac{\partial F(\overline{Q})}{\partial \overline{Q}} d\overline{Q} \, \mathrm{d}s.
\end{aligned}
$$

By the definition of discrete norm in [21], (3.1), and linear interpolation properties,

$$
\begin{aligned}
& b_h(Q,\gamma_h v_h) - b(Q,\gamma_h v_h) \\
\le & (\sum_{e\in\Gamma_0}\int_e h_e^{-1}[\gamma_h v_h]^2\,\mathrm{d}s)^{1/2}(\sum_{e\in\Gamma_0}\int_e h_e(\int_{Q(x)}^{Q(P_{jl})}\frac{\partial F(\overline{Q})}{\partial\overline{Q}}d\overline{Q})^2\,\mathrm{d}s)^{\frac{1}{2}} \\
\le & (\sum_{e\in\Gamma_0}\int_e h_e^{-1}[\gamma_h v_h]^2\,\mathrm{d}s)^{1/2}\cdot(\sum_{e\in\Gamma_0}\int_e h_e\|\frac{\partial F(\overline{Q})}{\partial\overline{Q}}\|_{0,\infty}^2\cdot|Q(P_{jl})-Q(x)|^2\,\mathrm{d}s)^{\frac{1}{2}} \\
\le & C(\sum_{e\in\Gamma_0}[\gamma_h v_h]^2)^{\frac{1}{2}}(\sum_{e\in\Gamma_0}\int_e h_e|Q(P_{jl})-Q(x)|^2\,\mathrm{d}s)^{\frac{1}{2}} \\
\le & C|||\gamma_h v_h|||_{1,h}\cdot(\sum_{e\in\Gamma_0}\|Q(P_{jl})-Q(x)\|_{0,K}^2 + h_e^2|Q(P_{jl})-Q(x)|_{1,K}^2)^{\frac{1}{2}} \\
\le & Ch|||\gamma_h v_h|||_{1,h}\cdot|Q|_{1,h}.
\end{aligned}
$$

□

Lemma 3.9. *For any $Q_h\in U_h$, there exists a positive constant C independent of h,*

$$|b_h(Q_h,\gamma_h v_h)| \le C|||v_h|||_{1,h}|Q_h|_{1,h},\quad \forall v_h\in U_h.$$

Proof. Since $[[\gamma_h v_h]]\cdot\mathbf{n}_e = [\gamma_h v_h]$, according to the definition of the bilinear form b_h, we have

$$
\begin{aligned}
b_h(Q_h,\gamma_h v_h) =& \sum_{e\in\Gamma_0}\int_e[\gamma_h v_h](H(be(Q_h))F(Q_h(P_{jl}))|_{K_1} + (1-H(be(Q_h)))F(Q_h(P_{jl}))|_{K_2})\,\mathrm{d}s \\
& -\sum_{e\in\Gamma_0}\int_e\{\gamma_h F(Q_h)\}[\gamma_h v_h]\,\mathrm{d}s + \sum_{e\in\Gamma_0}\int_e\{\gamma_h F(Q_h)\}[\gamma_h v_h]\,\mathrm{d}s \\
=& K_1 + K_2.
\end{aligned}
\tag{3.6}
$$

For K_2, we obtain

$$
\begin{aligned}
K_2 \le & \sum_{e\in\Gamma_0}(\int_e\{\gamma_h F(Q_h)\}^2\,\mathrm{d}s)^{1/2}\cdot(\int_e[\gamma_h v_h]^2\,\mathrm{d}s)^{\frac{1}{2}} \\
\le & \sum_{e\in\Gamma_0}(\int_e\|\gamma_h F(Q_h)\|_{0,K}^2 + h_e^2|\gamma_h F(Q_h)|_{1,K}^2\,\mathrm{d}s)^{1/2}\cdot\sum_{e\in\Gamma_0}(\int_e h_e^{-1}[\gamma_h v_h]^2\,\mathrm{d}s)^{\frac{1}{2}} \\
\le & C|||v_h|||_{1,h}\cdot\|Q_h\|.
\end{aligned}
$$

For K_1, since

$$\{\gamma_h F(Q_h)\}|_e = \frac{1}{2}\gamma_h F(Q_h)_{K_1}\cdot\mathbf{n}_e + \frac{1}{2}\gamma_h F(Q_h)_{K_2}\cdot\mathbf{n}_e,$$

we get

$$
\begin{aligned}
K_1 =& \sum_{e\in\Gamma_0}\int_e[\gamma_h v_h](H(be(Q_h))F(Q_h(P_{jl}))|_{K_1} + (1-H(be(Q_h)))F(Q_h(P_{jl}))|_{K_2} \\
& -(\frac{1}{2}\gamma_h F(Q_h)_{K_1}\cdot\mathbf{n}_e + \frac{1}{2}\gamma_h F(Q_h)_{K_2}\cdot\mathbf{n}_e)\,\mathrm{d}s.
\end{aligned}
$$

When $\mathbf{n}_e \cdot \dfrac{\partial F(\overline{Q_h})}{\partial \overline{Q_h}} > 0$, at the outflow boundary,

$$\begin{aligned}
K_1 &= \sum_{e\in\Gamma_0} \int_e [\gamma_h v_h](\frac{1}{2}F(Q_h(P_{jl}))|_{K_1} - \frac{1}{2}F(Q_h(P_{jl}))|_{K_2})\,\mathrm{d}s \\
&\le \frac{1}{2}\sum_{e\in\Gamma_0}(\int_e h_e^{-1}[\gamma_h v_h]^2\,\mathrm{d}s)^{\frac{1}{2}} \cdot (\int_e h_e(F(Q_h(P_{jl}))|_{K_1} - F(Q_h(P_{jl}))|_{K_2})^2\,\mathrm{d}s)^{\frac{1}{2}} \\
&\le C|||v_h|||_{1,h} \cdot ||Q_h||.
\end{aligned}$$

When $\mathbf{n}_e \cdot \dfrac{\partial F(\overline{Q_h})}{\partial \overline{Q_h}} < 0$, on the inflow boundary,

$$\begin{aligned}
K_1 &= \sum_{e\in\Gamma_0} \int_e [\gamma_h v_h][\frac{1}{2}F(Q_h(P_{jl}))|_{K_2} - \frac{1}{2}F(Q_h(P_{jl}))|_{K_1}]\,\mathrm{d}s \\
&\le \frac{1}{2}\sum_{e\in\Gamma_0}(\int_e h_e^{-1}[\gamma_h v_h]^2\,\mathrm{d}s)^{1/2} \cdot (\int_e h_e(F(Q_h(P_{jl}))|_{K_2} - F(Q_h(P_{jl}))|_{K_1})^2\,\mathrm{d}s)^{\frac{1}{2}} \\
&\le C|||v_h|||_{1,h} \cdot ||Q_h||.
\end{aligned}$$

Substituting K_1 and K_2 into (3.6) yields

$$b_h(Q_h, \gamma_h v_h) \le C|||v_h|||_{1,h} \cdot ||Q_h||.$$

□

Lemma 3.10. *There exists a positive constant C independent of h such that*

$$|d_h(Q, \gamma_h v_h) - d_h(Q_h, \gamma_h v_h)| \le C|||v_h|||_{1,h}(|||Q - Q_h|||_{1,h} + ||Q - Q_h||), \quad \forall v_h \in U_h.$$

Proof. According to the definition of the bilinear form $d_h(Q_h, v_h)$ and Green's formula,

$$\begin{aligned}
&|d_h(Q, \gamma_h v_h) - d_h(Q_h, \gamma_h v_h)| \\
=& \sum_{K\in T_h}\sum_{i=1}^{3}\int_{P_{i+1}BP_i}(F(Q) - F(Q_h)) \cdot \mathbf{n}\gamma_h v_h)\,\mathrm{d}s \\
=& \sum_{K\in T_h}\sum_{i=1}^{3}\int_{T_i}\nabla\cdot(F(Q) - F(Q_h))\cdot\gamma_h v_h\,\mathrm{d}x - \sum_{K\in T_h}\int_{\partial K}(F(Q) - F(Q_h))\cdot\mathbf{n}\gamma_h v_h\,\mathrm{d}s \\
=& \sum_{K\in T_h}\int_K \nabla\cdot(F(Q) - F(Q_h))\cdot(\gamma_h v_h - v_h)\,\mathrm{d}x + \sum_{K\in T_h}\int_K \nabla\cdot(F(Q) - F(Q_h))v_h\,\mathrm{d}x \\
&- \sum_{K\in T_h}\int_{\partial K}(F(Q) - F(Q_h))\cdot\mathbf{n}v_h\,\mathrm{d}s - \sum_{K\in T_h}\int_{\partial K}(F(Q) - F(Q_h))\cdot\mathbf{n}(\gamma_h v_h - v_h)\,\mathrm{d}s \\
=& H_1 + H_2 + H_3 + H_4.
\end{aligned} \tag{3.7}$$

Next, we sequentially estimate the terms H_1–H_4. Using the Holder inequality, Cauchy-Schwarz inequality, (2.2), and Lemma 3.2, we provide

$$|H_1| \le (\sum_{K\in T_h}||\nabla\cdot(F(Q) - F(Q_h))||_K^2)^{\frac{1}{2}} \cdot (\sum_{K\in T_h}||\gamma_h v_h - v_h||_K^2)^{\frac{1}{2}}$$

$$
\begin{aligned}
&\leq \|\nabla \cdot (F(Q) - F(Q_h)\| \cdot \|\gamma_h v_h - v_h\| \\
&\leq C|Q - Q_h|_{1,h} \cdot h|v_h|_{1,h} \\
&\leq Ch|\!|\!|v_h|\!|\!| \cdot |\!|\!|Q - Q_h|\!|\!|_{1,h}.
\end{aligned}
$$

Applying the Holder inequality, Cauchy-Schwarz inequality, trace inequality, (2.2), and Lemma 3.2, we have

$$
\begin{aligned}
|H_2| &\leq C \sum_{K\in T_h} (\|(F(Q) - F(Q_h)\|_{0,\partial K}^2)^{\frac{1}{2}} \cdot (\sum_{K\in T_h} \|\gamma_h v_h - v_h\|_{0,\partial K}^2)^{\frac{1}{2}} \\
&\leq C(\sum_{K\in T_h} h_e^{-1}\|Q - Q_h\|_{0,K}^2 + h_e|Q - Q_h|_{1,K}^2)^{\frac{1}{2}} \cdot \|\gamma_h v_h - v_h\|_e \\
&\leq C(h|Q - Q_h|_{1,h} + \|Q - Q_h\|)|\!|\!|v_h|\!|\!|_{1,h} \\
&\leq C\|Q - Q_h\|\|v_h\|_{1,h}.
\end{aligned}
$$

Similarly, we can also get H_3 and H_4 as follows:

$$
\begin{aligned}
|H_3| &\leq C(h|Q - Q_h|_{1,h} + \|Q - Q_h\|)|\!|\!|v_h|\!|\!|_{1,h} \leq C|\!|\!|v_h|\!|\!|_{1,h} \cdot \|Q - Q_h\|, \\
|H_4| &\leq C(h|Q - Q_h|_{1,h} + \|Q - Q_h\|)|\!|\!|v_h|\!|\!|_{1,h} \leq C|\!|\!|v_h|\!|\!|_{1,h} \cdot \|Q - Q_h\|.
\end{aligned}
$$

Combining H_1–H_4, we obtain

$$
|d_h(Q, \gamma_h v_h) - d_h(Q_h, \gamma_h v_h)| \leq C|\!|\!|v_h|\!|\!|_{1,h}(|\!|\!|Q - Q_h|\!|\!|_{1,h} + \|Q - Q_h\|), \quad \forall v_h \in U_h.
$$

□

4. Convergence analysis

In this section, we will establish a priori error estimates for the fully discrete scheme.

Let

$$
\rho^j = Q^j - R_h Q^j, \eta^j = R_h Q^j - Q_h^j.
$$

The true solution Q satisfies the equation

$$
(Q_t^j, \gamma_h v_h) + A(Q^j; Q^j, \gamma_h v_h) + b(Q^j, \gamma_h v_h) + d_h(Q^j, \gamma_h v_h) = (S_r^j, \gamma_h v_h). \tag{4.1}
$$

Subtracting (2.13) from (4.1), we have

$$
\begin{aligned}
&(Q_t^j - \frac{Q_h^j - Q_h^{j-1}}{\Delta t}, \gamma_h v_h) + A(Q^j; Q^j, \gamma_h v_h) - A(Q_h^j; Q_h^j, \gamma_h v_h) + b(Q^j, \gamma_h v_h) - b_h(Q_h^j, \gamma_h v_h) + d_h(Q^j - Q_h^j, \gamma_h v_h) \\
=&0,
\end{aligned} \tag{4.2}
$$

where

$$
\begin{aligned}
&A(Q^j; Q^j, \gamma_h v_h) - A(Q_h^j; Q_h^j, \gamma_h v_h) = A(Q^j; \eta^j, \gamma_h v_h) + A(Q^j; Q_h^j, \gamma_h v_h) - A(Q_h^j; Q_h^j, \gamma_h v_h), \\
&Q_t^j - \frac{Q_h^j - Q_h^{j-1}}{\Delta t} = Q_t^j + \partial_t \eta^j - \partial_t R_h Q^j.
\end{aligned}
$$

So, we can get the error equation

$$\begin{aligned}(\partial_t\eta^j,\gamma_h v_h)+A(Q^j;\eta^j,\gamma_h v_h)=&(\partial_t R_h Q^j-Q_t^j,\gamma_h v_h)+A(Q_h^j;Q_h^j,\gamma_h v_h)-A(Q^j;Q_h^j,\gamma_h v_h)\\&+b_h(Q_h^j,\gamma_h v_h)-b(Q^j,\gamma_h v_h).+d_h(Q_h^j-Q^j,\gamma_h v_h).\end{aligned}\tag{4.3}$$

The following theorem establishes a priori error estimates for the fully discrete scheme in the energy norm

Theorem 4.1. *If Q is the solution of problem* (2.12)*, $\{Q_h^i\}_{i=1}^N$ is the solution of problem* (2.13)*, then when Δt and h are small enough, there exists a constant C independent of h such that*

$$\sup_{0\le i\le N}|||Q^i-Q_h^i|||_{1,h}\le C(\Delta t+h).$$

Proof. Taking $v_h=\partial_t\eta^j$ in (4.3),

$$\begin{aligned}&(\partial_t\eta^j,\gamma_h\partial_t\eta^j)+A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)=(\partial_t R_h Q^j-Q_t^j,\gamma_h\partial_t\eta^j)+[A(Q_h^j;Q_h^j,\gamma_h\partial_t\eta^j)\\&-A(Q^j;Q_h^j,\gamma_h\partial_t\eta^j)]+[b_h(Q_h^j,\gamma_h\partial_t\eta^j)-b(Q^j,\gamma_h\partial_t\eta^j)]+d_h(Q_h^j-Q^j,\gamma_h\partial_t\eta^j).\end{aligned}\tag{4.4}$$

Notice that

$$(\partial_t\eta^j,\gamma_h\partial_t\eta^j)=|||\partial_t\eta^j|||_0^2,\tag{4.5}$$

and

$$\begin{aligned}A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)=&A(Q^j;\eta^j,\gamma_h(\frac{\eta^j-\eta^{j-1}}{\Delta t}))\\=&\frac{1}{\Delta t}A(Q^j;\eta^j,\gamma_h(\eta^j-\eta^{j-1}))\\=&\frac{1}{2\Delta t}[A(Q^j;\eta^j-\eta^{j-1},\gamma_h(\eta^j-\eta^{j-1}))+A(Q^j;\eta^j+\eta^{j-1},\gamma_h(\eta^j-\eta^{j-1}))]\\\ge&\frac{1}{2\Delta t}A(Q^j;\eta^j+\eta^{j-1},\gamma_h(\eta^j-\eta^{j-1}))\\\ge&\frac{1}{2\Delta t}[A(Q^j;\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^{j-1},\gamma_h\eta^{j-1})]-\frac{1}{2}[A(Q^j;\partial_t\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)].\end{aligned}\tag{4.6}$$

Substituting (4.5) and (4.6) into the left-hand side of (4.4), we obtain

$$\begin{aligned}&|||\partial_t\eta^j|||_0^2+\frac{1}{2\Delta t}[A(Q^j;\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^{j-1},\gamma_h\eta^{j-1})]\\\le&(\partial_t R_h Q^j-Q_t^j,\gamma_h\partial_t\eta^j)+\frac{1}{2}[A(Q^j;\partial_t\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)]+A(Q_h^j;Q_h^j,\gamma_h\partial_t\eta^j)\\&-A(Q^j;Q_h^j,\gamma_h\partial_t\eta^j)+b_h(Q_h^j,\gamma_h\partial_t\eta^j)-b(Q^j,\gamma_h\partial_t\eta^j)+d_h(Q_h^j-Q^j,\gamma_h\partial_t\eta^j).\end{aligned}\tag{4.7}$$

Simplify the last three lines of (4.7),

$$\begin{aligned}&A(Q_h^j;Q_h^j,\gamma_h\partial_t\eta^j)-A(Q^j;Q_h^j,\gamma_h\partial_t\eta^j)\\=&\frac{1}{\Delta t}[A(Q_h^j;Q_h^j,\gamma_h\eta^j)-A(Q_h^j;Q_h^j,\gamma_h\eta^{j-1})]\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{\Delta t}[A(Q^j;Q_h^j,\gamma_h\eta^j)-A(Q^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})]\\
&+[A(Q^j;\partial_t Q_h^j,\gamma_h\eta^{j-1})-A(Q_h^j;\partial_t Q_h^j,\gamma_h\eta^{j-1}]\\
&+\frac{1}{\Delta t}[A(Q_h^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})-A(Q_h^j;Q_h^{j-1},\gamma_h\eta^{j-1})\\
&+A(Q^j;Q_h^{j-1},\gamma_h\eta^{j-1})-A(Q^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})],\\
&b_h(Q_h^j,\gamma_h\partial_t\eta^j)-b(Q^j,\gamma_h\partial_t\eta^j)\\
=&\frac{1}{\Delta t}[b_h(Q_h^j,\gamma_h\eta^j)-b_h(Q^{j-1},\gamma_h\eta^{j-1})]-b_h(\partial_t Q_h^j,\gamma_h\eta^j)\\
&-\frac{1}{\Delta t}[b(Q^j,\gamma_h\eta^j)-b(Q^{j-1},\gamma_h\eta^{j-1})]+b(\partial_t Q^j,\gamma_h\eta^j)\\
=&\frac{1}{\Delta t}[b_h(Q_h^j,\gamma_h\eta^j)-b_h(Q^{j-1},\gamma_h\eta^{j-1})-b(Q^j,\gamma_h\eta^j)+b(Q^{j-1},\gamma_h\eta^{j-1})]\\
&+b(\partial_t Q^j,\gamma_h\eta^j)-b_h(\partial_t Q_h^j,\gamma_h\eta^j),\\
&d_h(Q_h^j-Q^j,\gamma_h\partial_t\eta^j)\\
=&\frac{1}{\Delta t}[d_h(Q_h^j-Q^j,\gamma_h\eta^j)-d_h(Q_h^{j-1}-Q^{j-1},\gamma_h\eta^{j-1})]-d_h(\partial_t(Q_h^j-Q^j),\gamma_h\eta^j),
\end{aligned}
$$

so (4.7) can be rewritten as

$$
\begin{aligned}
&|||\partial_t\eta^j|||_0^2+\frac{1}{2\Delta t}[A(Q^j;\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^{j-1},\gamma_h\eta^{j-1})]\\
\le&(\partial_t R_h Q^j-Q_t^j,\gamma_h\partial_t\eta^j)+\frac{1}{2}[A(Q^j;\partial_t\eta^j,\gamma_h\eta^j)-A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)]\\
&+\frac{1}{\Delta t}[A(Q_h^j;Q_h^j,\gamma_h\eta^j)-A(Q_h^j;Q_h^j,\gamma_h\eta^{j-1})]-\frac{1}{\Delta t}[A(Q^j;Q_h^j,\gamma_h\eta^j)-A(Q^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})]\\
&+[A(Q^j;\partial_t Q_h^j,\gamma_h\eta^{j-1})-A(Q_h^j;\partial_t Q_h^j,\gamma_h\eta^{j-1}]\\
&+\frac{1}{\Delta t}[A(Q_h^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})-A(Q_h^j;Q_h^{j-1},\gamma_h\eta^{j-1})\\
&+A(Q^j;Q_h^{j-1},\gamma_h\eta^{j-1})-A(Q^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})]\\
&+\frac{1}{\Delta t}[b_h(Q_h^j,\gamma_h\eta^j)-b_h(Q^{j-1},\gamma_h\eta^{j-1})]-b_h(\partial_t Q_h^j,\gamma_h\eta^j)\\
&-\frac{1}{\Delta t}[b(Q^j,\gamma_h\eta^j)-b(Q^{j-1},\gamma_h\eta^{j-1})]+b(\partial_t Q^j,\gamma_h\eta^j)\\
&+\frac{1}{\Delta t}[b_h(Q_h^j,\gamma_h\eta^j)-b_h(Q^{j-1},\gamma_h\eta^{j-1})-b(Q^j,\gamma_h\eta^j)+b(Q^{j-1},\gamma_h\eta^{j-1})]\\
&+b(\partial_t Q^j,\gamma_h\eta^j)-b_h(\partial_t Q_h^j,\gamma_h\eta^j)-d_h(\partial_t(Q_h^j-Q^j),\gamma_h\eta^j)\\
&+\frac{1}{\Delta t}[d_h(Q_h^j-Q^j,\gamma_h\eta^j)-d_h(Q_h^{j-1}-Q^{j-1},\gamma_h\eta^{j-1})].
\end{aligned}
\tag{4.8}
$$

Multiplying both sides of (4.8) by Δt and summing up j from 1 to i, we have

$$
\Delta t\sum_{j=1}^{i}|||\partial_t\eta^j|||_0^2+\frac{1}{2}A(Q^i;\eta^i,\gamma_h\eta^i)\le M_1+M_2+\cdots+M_9,
$$

where

$$
\begin{aligned}
M_1 =&\Delta t\sum_{j=1}^{i}(\partial_t R_h Q^j - Q_t^j, \gamma_h\partial_t\eta^j),\\
M_2 =&\frac{\Delta t}{2}\sum_{j=1}^{i}[A(Q^j;\partial_t\eta^j,\gamma_h\eta^j) - A(Q^j;\eta^j,\gamma_h\partial_t\eta^j)],\\
M_3 =&[A(Q_h^i;Q_h^i,\gamma_h\eta^i) - A(Q^i;Q_h^i,\gamma_h\eta^i)],\\
M_4 =&\Delta t\sum_{j=1}^{i}[A(Q^j;\partial_t Q_h^j,\gamma_h\eta^{j-1}) - A(Q_h^j;\partial_t Q_h^j,\gamma_h\eta^{j-1})],\\
M_5 =&\sum_{j=1}^{i}[A(Q_h^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1}) - A(Q_h^j;Q_h^{j-1},\gamma_h\eta^{j-1})\\
&+ A(Q^j;Q_h^{j-1},\gamma_h\eta^{j-1}) - A(Q^{j-1};Q_h^{j-1},\gamma_h\eta^{j-1})],\\
M_6 =&\Delta t\sum_{j=1}^{i}[b_h(\partial_t Q^j - Q_h^j,\gamma_h\eta^j) + b(\partial_t Q^j,\gamma_h\eta^j) - b_h(\partial_t Q^j,\gamma_h\eta^j)],\\
M_7 =&[b_h(R_h Q^i,\gamma_h\eta^i) - b_h(\eta^i,\gamma_h\eta^i) - b(Q^i,\gamma_h\eta^i)],\\
M_8 =&d_h(Q_h^i - Q^i,\gamma_h\eta^i),\\
M_9 =&\Delta t\sum_{j=1}^{i}[d_h(Q_h^j - Q^j,\gamma_h\eta^j)].
\end{aligned}
$$

Next, we estimate the right-hand terms M_1–M_9. For M_1,

$$
\begin{aligned}
|(\partial_t R_h Q^j - Q_t^j,\gamma_h\partial_t\eta^j)| \le&|(\partial_t Q^j - Q_t^j,\gamma_h\partial_t\eta^j)| + |(\partial_t\rho^j - Q_t^j,\gamma_h\partial_t\eta^j)| \le \|\partial_t Q^j - Q_t^j\|^2 + \|\partial_t\rho^j\|^2 + |||\partial_t\eta^j|||_0^2,\\
\|\partial_t\rho^j\| =&\|\frac{\rho^j-\rho^{j-1}}{\Delta t}\| \le \frac{C}{\Delta t}\int_{t^{j-1}}^{t^j}\|\rho_t\|\,\mathrm{d}\tau \le \frac{Ch^4}{\Delta t}\int_{t^{j-1}}^{t^j}\|Q\|_{1,3,2}\,\mathrm{d}\tau \le Ch^4\|Q\|_{L^\infty(H^3(\Omega))},\\
\|\partial_t Q^j - Q_t^j\| =&\|\frac{1}{\Delta t}\int_{t^{j-1}}^{t^j}(t^j - t)Q_{tt}\,\mathrm{d}\tau\| \le C\int_{t^{j-1}}^{t^j}\|Q_{tt}\|\,\mathrm{d}\tau \le C\Delta t\|Q_{tt}\|_{L^\infty(L^2(\Omega))}.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality and Young's inequality, we get

$$
|M_1| \le \Delta t\sum_{j=1}^{n}|||\partial_t\eta^j|||_0^2 + C(\Delta t)^2\|Q_{tt}\|_{L^\infty(L^2\Omega)}^2 + Ch^4\Delta t\|Q\|_{L^\infty(H^3(\Omega))}^2.
$$

For M_2, using Lemmas 3.1 and 3.5, we have

$$
|M_2| \le C\frac{\Delta t}{2}\sum_{j=1}^{i}h|||\partial_t\eta^j|||_{1,h}|||\eta^j|||_{1,h} \le C\Delta t\sum_{j=1}^{i}|||\partial_t\eta^j|||_0|||\eta^j|||_{1,h} \le \Delta t\sum_{j=1}^{n}|||\partial_t\eta^j|||_0^2 + C\Delta t\sum_{j=1}^{i}|||\eta^j|||_{1,h}^2.
$$

For M_3–M_5, we will use the following propositions:

$$
|\partial_t Q_h^j|_{1,\infty} \le C_1, \qquad j = 0,1,\cdots,N, \tag{4.9}
$$

$$|Q_h^j|_{1,\infty} \leq C_2, \qquad j = 0, 1, \cdots, N, \tag{4.10}$$

the detailed proof will be given later.

For M_3 and M_5, it follows from the Cauchy-Schwarz inequality and (4.10) that

$$\begin{aligned}
|M_3| &\leq C|Q_h^i|_{1,\infty}(\|Q^i - Q_h^i\| + h|\!|\!|Q^i - Q_h^i|\!|\!|_{1,h})|\!|\!|\eta^i|\!|\!|_{1,h} \\
&\leq C|\!|\!|\eta^i|\!|\!|_{1,h}(\|\rho^i\| + \|\eta^i\| + h|\!|\!|\rho^i|\!|\!|_{1,h} + h|\!|\!|\eta^i|\!|\!|_{1,h}) \\
&\leq C|\!|\!|\eta^i|\!|\!|_{1,h}^2 + C(\|\rho^i\|^2 + \|\eta^i\|^2 + h^2|\!|\!|\rho^i|\!|\!|_{1,h}^2),
\end{aligned}$$

$$\begin{aligned}
|M_5| \leq\ & C\sum_{j=1}^{i} |Q_h^{j-1}|_{1,\infty}(\|Q^{j-1} - Q_h^{j-1}\| + h|\!|\!|Q^{j-1} - Q_h^{j-1}|\!|\!|_{1,h})|\!|\!|\eta^{j-1}|\!|\!|_{1,h} \\
&+C\sum_{j=1}^{i} |Q_h^{j-1}|_{1,\infty}(\|Q^{j} - Q_h^{j}\| + h|\!|\!|Q^{j} - Q_h^{j}|\!|\!|_{1,h})|\!|\!|\eta^{j-1}|\!|\!|_{1,h} \\
\leq\ & C\sum_{j=1}^{i} |\!|\!|\eta^{j-1}|\!|\!|_{1,h}(\|\rho^j\| + \|\eta^j\| + h|\!|\!|\rho^j|\!|\!|_{1,h} + h|\!|\!|\eta^j|\!|\!|_{1,h}) \\
&+C\sum_{j=1}^{i} |\!|\!|\eta^{j-1}|\!|\!|_{1,h}(\|\rho^{j-1}\| + \|\eta^{j-1}\| + h|\!|\!|\rho^{j-1}|\!|\!|_{1,h} + h|\!|\!|\eta^{j-1}|\!|\!|_{1,h}) \\
\leq\ & C\sum_{j=1}^{i} |\!|\!|\eta^j|\!|\!|_{1,h}^2 + C(\|\rho^j\|^2 + \|\eta^j\|^2 + h^2|\!|\!|\rho^j|\!|\!|_{1,h}^2).
\end{aligned}$$

For M_4, it follows from the Cauchy-Schwarz inequality and (4.9) that

$$\begin{aligned}
|M_4| &\leq C\Delta t\sum_{j=1}^{i} |\partial_t Q_h^j|_{1,\infty}(\|Q^j - Q_h^j\| + h|\!|\!|Q^j - Q_h^j|\!|\!|_{1,h})|\!|\!|\eta^{j-1}|\!|\!|_{1,h} \\
&\leq C\Delta t\sum_{j=1}^{i} |\!|\!|\eta^{j-1}|\!|\!|_{1,h}(\|\rho^j\| + \|\eta^j\| + h|\!|\!|\rho^j|\!|\!|_{1,h} + h|\!|\!|\eta^j|\!|\!|_{1,h}) \\
&\leq C\Delta t\sum_{j=1}^{i} |\!|\!|\eta^j|\!|\!|_{1,h}^2 + C(\|\rho^j\|^2 + \|\eta^j\|^2 + h^2|\!|\!|\rho^j|\!|\!|_{1,h}^2).
\end{aligned}$$

For M_6, using Lemmas 3.8 and 3.9, and

$$\|\partial_t Q^j\| \leq \frac{C}{\Delta t}\int_{t^{j-1}}^{t^j} \|Q_t\| \leq C\|Q\|_{L^\infty(L^2(\Omega))},$$

we obtain

$$\begin{aligned}
|M_6| \leq\ & C\Delta t\sum_{j=1}^{i} |\!|\!|\eta^{j-1}|\!|\!|_{1,h}(\|\partial_t\rho^j\| + \|\partial_t\eta^j\| + Ch|\!|\!|\eta^j|\!|\!|_{1,h}\|Q_t\|_{L^\infty(H^3(\Omega))}) \\
\leq\ & C\Delta t\sum_{j=1}^{i}(|\!|\!|\eta^j|\!|\!|_{1,h}^2 + Ch^4\|Q\|_{L^\infty(L^2(\Omega))}^2 + Ch^2\|Q_t\|_{L^\infty(L^2(\Omega))}^2) + |\!|\!|\partial_t\eta^j|\!|\!|_0^2.
\end{aligned}$$

For M_7, applying the definition of $b(u, v)$, Young's inequality, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|M_7| &\leq C|\!|\!|\eta^i|\!|\!|_{1,h}(\|R_h Q^i\| + \|\eta^i\|) + Ch|\!|\!|\eta^i|\!|\!|_{1,h}\|Q^i\|_{1,h} \\
&\leq C(|\!|\!|\eta^i|\!|\!|_{1,h}^2 + h^4|Q|_2^2 + \|\eta^i\|^2 + h^2\|Q^i\|_{1,h}^2).
\end{aligned}$$

Applying Lemma 3.10, the triangle inequality, and ϵ inequality to get estimates for M_8 and M_9,

$$
\begin{aligned}
|M_8| &\le C\|\!|\eta^i|\!\|_{1,h}(\|Q^i - Q_h^i\| + h\|\!|Q^i - Q_h^i|\!\|_{1,h}) \\
&\le C\sum_{j=1}^{i}\|\!|\eta^j|\!\|_{1,h}^2 + C(\|\rho^j\|^2 + \|\eta^j\|^2 + h^2\|\!|\rho^j|\!\|_{1,h}^2), \\
|M_9| &\le C\Delta t\sum_{j=1}^{i}\|\!|\eta^j|\!\|_{1,h}(h\|\!|\partial_t(Q^j - Q_h^j)|\!\|_{1,h} + \|\partial_t(Q^j - Q_h^j)\|) \\
&\le C\Delta t\sum_{j=1}^{i}\|\!|\eta^j|\!\|_{1,h}(\|\partial_t\rho^j\| + \|\partial_t\eta^j\| + h\|\!|\partial_t\rho^j|\!\|_{1,h} + h\|\!|\partial_t\eta^j|\!\|_{1,h}) \\
&\le C\Delta t\sum_{j=1}^{i}\|\!|\eta^j|\!\|_{1,h}(\|\partial_t\rho^j\| + \|\!|\partial_t\eta^j|\!\|_0) \\
&\le C\Delta t\sum_{j=1}^{i}(\|\!|\eta^j|\!\|_{1,h}^2 + Ch^4\|Q\|_{L^\infty(H^3(\Omega))}^2 + \|\!|\partial_t\eta^j|\!\|_0^2).
\end{aligned}
$$

Applying Lemma 3.3, we get

$$
\begin{aligned}
&\Delta t\sum_{j=1}^{i}\|\!|\partial_t\eta^j|\!\|_0^2 + \beta\|\!|\eta^j|\!\|_{1,h}^2 \\
\le& C(\|\rho^i\|^2 + \|\eta^i\|^2 + h^2\|\!|\rho^i|\!\|_{1,h}^2 + h^4|Q|_2^2 + h^2|Q|_1^2) + C\|\!|\eta^i|\!\|_{1,h}^2 \\
&+ C\Delta t\sum_{j=1}^{i}\|\!|\eta^j|\!\|_{1,h}^2 + \sum_{j=1}^{i}(\|\rho^j\|^2 + \|\eta^j\|^2 + h^2\|\!|\rho^j|\!\|_{1,h}^2) \\
&+ C\Delta t\sum_{j=1}^{i}(h^4\|Q\|_{L^\infty(L^2(\Omega))}^2 + h^4\|Q\|_{L^\infty(H^3(\Omega))}^2 + h^2\|Q_t\|_{L^\infty(L^2(\Omega))}^2 + \epsilon\|\!|\partial_t\eta^j|\!\|_0^2).
\end{aligned}
$$

By Gronwall's inequality and (3.3), we supply

$$
\|\!|\eta^i|\!\|_{1,h} \le C(h + \Delta t). \tag{4.11}
$$

Combine the triangle inequality with (3.3),

$$
\|\!|Q^i - Q_h^i|\!\|_{1,h} \le C(\Delta t + h).
$$

Finally, we give the proof of the proposition (4.9) and (4.10).
In (4.4), take $j = 1$, then

$$
\begin{aligned}
&(\partial_t\eta^1, \gamma_h\partial_t\eta^1) + A(Q^1; \eta^1, \gamma_h\partial_t\eta^1) \\
=&(\partial_t R_h Q^1 - Q_t^1, \gamma_h\partial_t\eta^1) + [A(Q_h^1; Q_h^1, \gamma_h\partial_t\eta^1) - A(Q^1; Q_h^1, \gamma_h\partial_t\eta^1)] \\
&+ [b(Q_h^1, \gamma_h\partial_t\eta^1) - b(Q^1, \gamma_h\partial_t\eta^1)] + d_h(Q_h^1, \gamma_h\partial_t\eta^1).
\end{aligned} \tag{4.12}
$$

Since η^0=0, $\partial_t\eta^1 = \dfrac{1}{\Delta t}\eta^1$. Similar to (4.11), we have

$$
\Delta t\|\!|\partial_t\eta^1|\!\|_0^2 + \beta\|\!|\eta^1|\!\|_{1,h}^2 \le C(h^4 + \Delta t^2) + C\|\!|\eta^1|\!\|_{1,h}^2.
$$

So,

$$(\Delta t|||\partial_t\eta^1|||_0^2)^{\frac{1}{2}} \le C(h^2 + \Delta t).$$

It follows from the triangle inequality and Lemma 3.7 that

$$||\partial_t Q_h^1||_{1,\infty} \le ||\partial_t\eta^1||_{1,\infty} + ||\partial_t R_h Q^1||_{1,\infty} \le C\Delta t^{-\frac{1}{2}}(h^2 + \Delta t) + C_1 \le C_1 \quad (h \to 0).$$

So, $||\partial_t Q_h^1||_{1,\infty}$ is bounded.

Similarly, summing up (4.4) from $j = 1$ to i, it follows from Lemma 3.3 and Gronwall's inequality that

$$\sum_{j=1}^{i} ||\partial_t Q_h^j||_0^2 + |||\eta^i|||_{1,h}^2 \le C(\Delta t^2 + h^2).$$

We obtain

$$\begin{aligned} ||\partial_t Q_h^i||_{1,\infty}^2 &\le ||\partial_t\eta^i||_{1,\infty}^2 + ||\partial_t R_h Q^i||_{1,\infty} \le C|lnh|^{\frac{1}{2}}(\sum_{j=1}^{i} ||\partial_t\eta^j||_0^2)^{\frac{1}{2}} + ||\partial_t R_h Q^i||_{1,\infty} \\ &\le C|lnh|^{\frac{1}{2}}(\Delta t)^{-\frac{1}{2}}(\Delta t + h) + C \le C \quad (h \to 0). \end{aligned}$$

Therefore, using the mathematical induction, we can prove (4.9).

Then, we give the proof of the proposition (4.10).

Assume $C_2 = 1 + \sup_{[0,T]} |R_h Q_t|_{1,\infty}$. By mathematical induction, when $t = 0$, $Q_h^0 - R_h Q^0 = 0$, so obviously there is

$$|Q_{ht}^0|_{1,\infty} \le \sup_{[0,T]} |R_h Q_t|_{1,\infty} < C_2.$$

For $i = 0, 1, \cdots, k-1$, assume (4.10) holds, then we can get

$$|||\eta^i|||_{1,h} \le C(h + \Delta t).$$

For $i = k$,

$$|Q_{ht}^j|_{1,\infty} \le |Q_{ht}^n - R_h Q_t^j|_{1,\infty} + |R_h Q_t^i|_{1,\infty} \le C|lnh|^{\frac{1}{2}}|Q_{ht}^j - R_h Q_t^j|_{1,h} + |R_h Q_t^j|_{1,\infty} \le C_2.$$

While h is small enough and $\Delta t = O(h)$, we obtain $|Q_{ht}^j|_{1,\infty} \le C_2$. This shows that the mathematical induction holds. Hence, the proof of (4.10) is completed. □

The error estimate for the fully discrete scheme in the L^2 norm is given in the next theorem.

Theorem 4.2. *Under the conditions of Theorem 4.1, we have*

$$\sup_{0 \le i \le N} ||Q^i - Q_h^i|| \le C(\Delta t + h).$$

Proof. The conclusion follows directly from Lemma 3.6 and Theorem 4.1. □

Remark 4.1. *The proof of Theorem 4.2 follows the standard duality argument for the L^2 norm. Taking $v_h = \eta^j$ in the error Eq* (4.3) *and estimating the resulting terms using Lemmas 3.3–3.10 yields a first-order convergence rate. However, a crucial difficulty arises in the estimate of the convection term. The upwind bilinear form $b_h(\cdot,\cdot)$ introduces a consistency error of order h (see Lemma 3.8):*

$$|b_h(Q^j, \gamma_h\eta^j) - b(Q^j, \gamma_h\eta^j)| \leq Ch\|\eta^j\|_{1,h}|Q|_{1,h}.$$

This h-dependent error limits the L^2 norm error estimate derived by the standard duality argument to first-order accuracy, i.e., $O(h)$. Nevertheless, our numerical experiments in Section 5 consistently demonstrate a second-order convergence rate in the L^2 norm (see Tables 2–8). This suggests that the upwind scheme may possess a superconvergence property in practice and that our theoretical estimate is not sharp. A refined analysis, possibly exploiting additional cancellation or stability properties of the upwind bilinear form, is needed to close this gap. We leave this as an interesting open problem for future research.

5. Numerical experiments

In this section, we perform numerical experiments to verify the theoretical analysis and demonstrate the efficiency of the proposed upwind scheme.

5.1. Experiment 1

Consider the two-dimensional domain $\Omega = [0,1] \times [0,1]$ for the system (1.1), where the final time $T = 1$. All the right-hand side terms are selected according to the analytical solution $Q = tsin\pi xsin\pi z$, with $0 \leq Q \leq 1$. The space step is h, the time step is Δt, and the penalty parameter is $\alpha = 40$.

The nonlinear algebraic system (2.13) is solved using the **Picard iteration** at each time step. Given the solution at the previous time level Q_h^{n-1}, we seek $Q_h^n \in U_h$ satisfying

$$\left(\frac{Q_h^n - Q_h^{n-1}}{\Delta t}, \gamma_h v_h\right) + A(Q_h^n; Q_h^n, \gamma_h v_h) + b_h(Q_h^n, \gamma_h v_h) + d_h(Q_h^n, \gamma_h v_h) = (S_r^n, \gamma_h v_h), \quad \forall v_h \in U_h.$$

The Picard iteration proceeds as follows. At the k-th iteration ($k = 0, 1, 2, \ldots$), given the previous iteration $Q_h^{n,(k)}$, we solve the following **linearized** problem for $Q_h^{n,(k+1)}$:

$$\left(\frac{Q_h^{n,(k+1)} - Q_h^{n-1}}{\Delta t}, \gamma_h v_h\right) + A(Q_h^{n,(k)}; Q_h^{n,(k+1)}, \gamma_h v_h) + b_h(Q_h^{n,(k)}, \gamma_h v_h) + d_h(Q_h^{n,(k)}, \gamma_h v_h) = (S_r^n, \gamma_h v_h),$$

where only the coefficients $D(Q_h^{n,(k)})$, $K(Q_h^{n,(k)})$, and $F(Q_h^{n,(k)})$ are evaluated using the previous iteration. The bilinear form $A(Q_h^{n,(k)}; Q_h^{n,(k+1)}, \gamma_h v_h)$ is linear in the unknown $Q_h^{n,(k+1)}$.

The iteration is initialized with $Q_h^{n,(0)} = Q_h^{n-1}$. The linear system at each Picard iteration is solved using a direct solver for small-scale problems or the conjugate gradient method preconditioned by incomplete LU factorization for larger problems. The iteration is terminated when the relative residual satisfies

$$\frac{\|\mathcal{F}(Q_h^{n,(k+1)})\|}{\|\mathcal{F}(Q_h^{n,(0)})\|} < \text{TOL},$$

where $\mathcal{F}(\cdot)$ denotes the nonlinear residual, and we set $\text{TOL} = 10^{-8}$ in all numerical experiments.

Remark 5.1. *The Picard iteration is preferred over Newton's method in this context for several reasons: (i) It does not require the computation of the Jacobian matrix, which is complex for the upwind DFVEM formulation; (ii) it is simpler to implement and more robust for the nonlinear Richards equation; (iii) the mild nonlinearity of the problem ensures rapid convergence without the need for line search or damping strategies commonly required by Newton's method.*

First, to show the convergence orders in time, we fix the mesh size $h = \frac{1}{32}$ and $D(Q) = Q + 1$, $K(Q) = 2Q^{10}$. In Table 1, numerical results demonstrate first-order temporal convergence in both the L^2 and energy norms, consistent with our theoretical analysis.

Table 1. Upwind DFVEM solutions and convergence orders with $h = \frac{1}{32}$.

Δt	L^2 error	order	H^1 error	order
1/2	0.3164		0.0651	
1/4	0.2028	0.8335	0.0314	1.0519
1/8	0.1075	0.9157	0.0161	0.9637

Next, we take $D(Q) = Q + 1$, $K(Q) = cQ^{10}$. The results are shown in Tables 2–4. From these tables, we can find that the optimal convergence orders for error can be obtained by our proposed method.

Table 2. Upwind DFVEM solutions and convergence orders with $c = \frac{1}{40}$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2139		1.5188	
1/4	1/8	0.0516	2.0515	0.7759	0.9690
1/8	1/32	0.0105	2.2970	0.3628	1.0967
1/16	1/128	0.0023	2.1907	0.1724	1.0734

Table 3. Upwind DFVEM solutions and convergence orders with $c = 1$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2144		1.5207	
1/4	1/8	0.0520	2.0437	0.7779	0.9671
1/8	1/32	0.0107	2.2809	0.3632	1.0988
1/16	1/128	0.0024	2.1565	0.1725	1.0742

Table 4. Upwind DFVEM solutions and convergence orders with $c = 2$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2156		1.5263	
1/4	1/8	0.0535	2.0107	0.7837	0.9617
1/8	1/32	0.0112	2.2560	0.3644	1.1048
1/16	1/128	0.0026	2.1069	0.1727	1.0773

To demonstrate the superiority of the proposed upwind scheme over the standard DFVEM in [7], we perform a direct comparison on the manufactured solution test case. The standard DFVEM is implemented by replacing the upwind bilinear form $b_h(\cdot,\cdot)$ with the standard form $b(\cdot,\cdot)$ defined in (2.9), while keeping all other discretization parameters identical.

The numerical experiments in Chen and Xu [7] reveal that the standard DFVEM exhibits significant deterioration in accuracy when the convection term becomes dominant. For $c = 1/40$ (convection weak), the L^2 convergence order is approximately 2.19; for $c = 1$ (convection dominant), it drops to approximately 1.04; and for $c = 2$ (strongly convection dominant), it drops to approximately 0.77.

Table 5 presents the comparative L^2 errors and convergence orders for both methods under identical mesh and time-step settings. The results demonstrate that when the convection term is dominant, the standard DFVEM suffers from severe accuracy degradation, while the upwind DFVEM maintains optimal convergence rates.

Table 5. Comparison of L^2 errors and convergence orders between the standard DFVEM and the upwind DFVEM.

c	Δt	Standard DFVEM [7]		Upwind DFVEM	
		L^2 error	Order	L^2 error	Order
1/40	1/2	2.1390e-01	–	2.1390e-01	–
	1/4	5.1600e-02	2.05	5.1600e-02	2.05
	1/8	1.0500e-02	2.30	1.0500e-02	2.30
	1/16	2.3000e-03	2.19	2.3000e-03	2.20
1	1/2	2.1440e-01	–	2.1440e-01	–
	1/4	5.3400e-02	2.01	5.2000e-02	2.04
	1/8	1.0900e-02	2.29	1.0700e-02	2.28
	1/16	5.3000e-03	1.04	2.4000e-03	2.16
2	1/2	2.1560e-01	–	2.1560e-01	–
	1/4	5.6100e-02	1.94	5.3500e-02	2.01
	1/8	1.3800e-02	2.02	1.1200e-02	2.26
	1/16	7.8100e-02	0.77	2.6000e-03	2.11

Considering that the analytical solution Q discussed in problem (1.1) of this paper has a value range of $[0,1]$, we take $D(Q) = Q^8 + 1, K(Q) = cQ^{10}$. Tables 6–8 show the convergence orders for error. The results demonstrate that the upwind discontinuous finite volume element scheme achieves optimal convergence rates in the L^2 norm and the energy norm.

Table 6. Upwind DFVEM solutions and convergence orders with $c = \frac{1}{40}$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2508		1.5153	
1/4	1/8	0.0629	1.9954	0.7779	0.9619
1/8	1/32	0.0125	2.3311	0.3644	1.0941
1/16	1/128	0.0025	2.3219	0.1726	1.0781

Table 7. Upwind DFVEM solutions and convergence orders with $c = 1$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2512		1.5178	
1/4	1/8	0.0639	1.9749	0.7832	0.9545
1/8	1/32	0.0129	2.3084	0.3653	1.1003
1/16	1/128	0.0027	2.2563	0.1728	1.0800

Table 8. Upwind DFVEM solutions and convergence orders with $c = 2$.

Δt	h	L^2 error	order	H^1 error	order
1/2	1/2	0.2526		1.5252	
1/4	1/8	0.0669	1.9168	0.7990	0.9327
1/8	1/32	0.0138	2.2773	0.3680	1.1185
1/16	1/128	0.0031	2.1543	0.1733	1.0864

Therefore, the upwind DVFEM is an effective approach for solving nonlinear problems. Although our numerical experiments show a second-order convergence rate in the L^2 norm for the nonlinear problem, the theoretical analysis in this paper only establishes a first-order estimate. This discrepancy indicates that the current duality argument is suboptimal for the upwind scheme, and a more refined analysis is worthy of further investigation. Nonetheless, the numerical results confirm the practical efficiency and accuracy of the proposed method.

5.2. Experiment 2

For the nonlinear soil hydraulic functions, we use the van Genuchten-Mualem model [19]:

$$S_e(Q) = \frac{Q - Q_r}{Q_s - Q_r},\ K(Q) = K_s S_e^{\frac{1}{2}}[1 - (1 - S_e^{\frac{1}{m}})^m]^2,$$

$$D(Q) = \frac{(1-m)K_s}{\theta m(Q_s - Q_r)} S_e^{\frac{1}{2}-m}((1 - S_e^{\frac{1}{m}})^{-m} + (1 - S_e^{\frac{1}{m}})^m - 2).$$

We consider the sandy soil from the USDA soil texture classification with the following parameters: $Q_s = 0.43$, $Q_r = 0.045$, $\theta = 0.145$, $n = 2.68$, $m = 1 - \frac{1}{n}$, and $K_s = 712.8$. For this soil type, the Richards equation is strongly convection-dominated.

The numerical simulations are conducted over the computational domain $\Omega = [0, 8]^2$ using the upwind DFVEM, with mesh size $h = \frac{1}{16}$, time step $\Delta t = 0.001$, and different final times $T = 0.25, 0.5, 1$. We set $S_r = 0$. Figure 3 exhibits the spatial distribution of soil moisture content for the various final times T. As depicted in Figure 3, the wetting front of the soil expands progressively with increasing infiltration duration, indicating a continuous increase in soil moisture content. The numerical result is in good agreement with the physical behavior observed in practical scenarios.

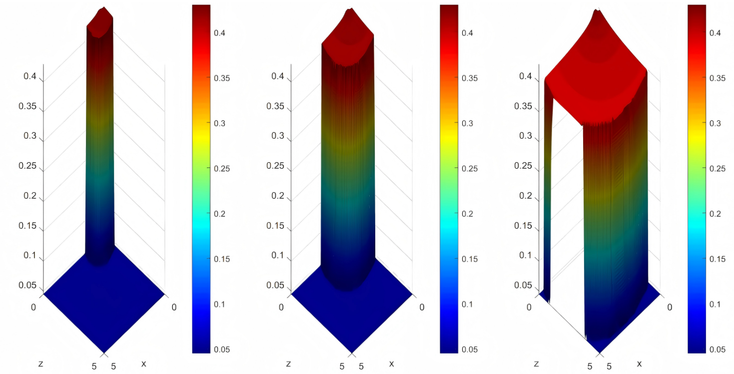

Figure 3. The upwind DFVEM solution Q_h^n with $T = 0.25$ (Left), $T = 0.5$ (Center), and $T = 1$ (Right).

We also verify the mass conservation property of the proposed upwind DFVEM. For the van Genuchten-Mualem problem, the total water content is computed as

$$M^n = \int_\Omega Q_h^n \, dx = \sum_{K \in T_h} \int_K Q_h^n \, dx,$$

and the relative mass change is defined by $E_M^n = \left| \frac{M^n - M^0}{M^0} \right|$, where M^0 is the initial total water content.

Figure 4 presents the mass conservation verification for the van Genuchten-Mualem problem. The absolute mass (upper panel) increases smoothly and monotonically from 0 to approximately 24.5, which is physically correct due to continuous water injection through the inlet boundary condition $Q = Q_s$. The relative mass (lower panel) increases from 0 to approximately 8.51. For this simulation, the initial mass is $M_0 = 2.88$, and the final mass is $M_{end} = 27.40$, resulting in a relative mass change of approximately 8.51. This confirms that the numerical scheme accurately tracks the cumulative mass increase and that no unphysical mass loss or gain is observed. Table 9 summarizes the maximum relative mass for different mesh sizes, further confirming the mesh-independent mass conservation property.

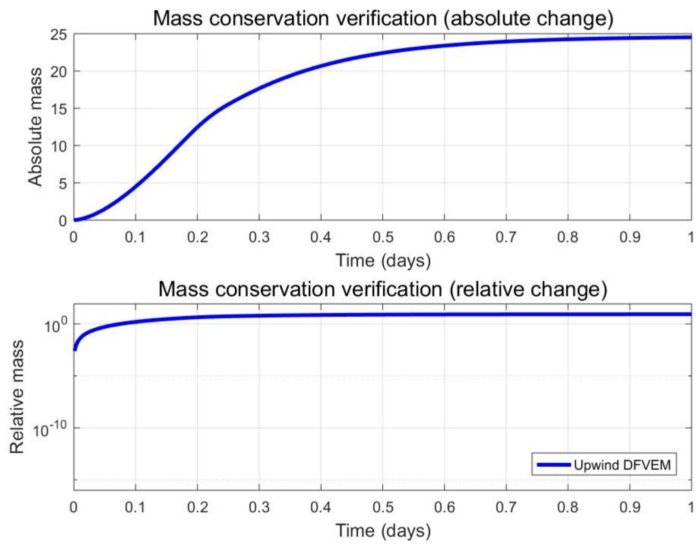

Figure 4. Mass conservation verification for the van Genuchten-Mualem problem. Upper panel: absolute mass change (linear scale). Lower panel: relative mass change (semi-log scale).

Table 9. Maximum relative mass for different mesh sizes.

Mesh size h	Maximum relative mass
$1/2$	8.55
$1/4$	8.54
$1/8$	8.53
$1/16$	8.51

Author contributions

Fang Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing–original draft, Writing–review & editing; Fan Chen: Conceptualization, Methodology, Supervision, Writing–review & editing; Changqing Lv: Software, Validation, Writing–review & editing; Xiaoyan Qin: Software, Validation, Writing–review & editing.

Use of Generative-AI tools declaration

The authors declare they have used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are grateful to the editor and the anonymous reviewers for their insightful comments and suggestions. This work was supported by the Natural Science Foundation of Shandong Province (Nos. ZR2022MA081, ZR2024MA056, ZR2025LZH001, ZR2025MS1036).

Conflict of interest

The authors declare no conflict of interest.

References

1. D. N. Arnold, An interior penalty finite element method with discontinuous elements, *SIAM J. Numer. Anal.*, **19** (1982), 742–760. https://doi.org/10.1137/0719052
2. J. Bear, *Dynamics of fluids in porous media*, Dover: Dover Publications, 2013.
3. C. Bi, J. Geng, Discontinuous finite volume element method for parabolic problems, *Numer. Meth. Part. D. E.*, **26** (2010), 367–383. https://doi.org/10.1002/num.20437
4. D. Breit, T. C. Moyo, P. Öffner, Discontinuous Galerkin methods for the complete stochastic Euler equations, *J. Comput. Phys.*, **541** (2025), 114324. https://doi.org/10.1016/j.jcp.2025.114324
5. R. Bürger, S. Kumar, R. Ruiz-Baier, Discontinuous finite volume element discretization for coupled flow-transport problems arising in models of sedimentation, *J. Comput. Phys.*, **299** (2015), 446–471. https://doi.org/10.1016/j.jcp.2015.07.020

6. F. Chen, M. Cui, C. Zhou, Sequential symmetric interior penalty discontinuous Galerkin method for fully coupled quasi-static thermo-poroelasticity problems, *Comput. Math. Appl.*, **218** (2026), 67–88. https://doi.org/10.1016/j.camwa.2026.06.012

7. F. Chen, Z. Xu, Discontinuous finite volume element method of two-dimensional unsaturated soil water movement problem, *Adv. Differ. Equ.*, **2019** (2019), 478. https://doi.org/10.1186/s13662-019-2395-7

8. S. H. Chou, X. Ye, Unified analysis of finite volume methods for second order elliptic problems, *SIAM J. Numer. Anal.*, **45** (2007), 1639–1653. https://doi.org/10.1137/050643994

9. C. N. Dawson, M. F. Wheeler, C. S. Woodward, A two-grid finite difference scheme for nonlinear parabolic equations, *SIAM J. Numer. Anal.*, **35** (1998), 435–452. https://doi.org/10.1137/S0036142995293493

10. R. Devi, D. N. Pandey, Discontinuous Galerkin methods for nonlinear parabolic delay-equations of nonmonotone type, *J. Sci. Comput.*, **101** (2024), 50. https://doi.org/10.1007/s10915-024-02696-x

11. Z. Di, Z. Luo, Z. Xie, A. Wang, I. M. Navon, An optimizing implicit difference scheme based on proper orthogonal decomposition for the two-dimensional unsaturated soil water flow equation, *Int. J. Numer. Meth. Fluids*, **68** (2012), 1324–1340. https://doi.org/10.1002/fld.2610

12. F. Gao, Y. Yuan, An upwind finite volume element method based on quadrilateral meshes for nonlinear convection-diffusion problems, *Numer. Meth. Part. D. E.*, **25** (2009), 1067–1085. https://doi.org/10.1002/num.20387

13. H. R. Li, Z. D. Luo, Z. H. Xie, J. Zhu, Generalized difference methods and numerical simulation for unsaturated soil water flow problems (Chinese), *Mathematica Numerica Sinica*, **28** (2006), 321–336. https://doi.org/10.12286/jssx.2006.3.321

14. L. Li, Z. Zhou, A mass-conserved domain decomposition method for the unsaturated soil flow water problem, *Adv. Differ. Equ.*, **2019** (2019), 272. https://doi.org/10.1186/s13662-019-2213-2

15. J. Li, J. Zeng, R. Li, An adaptive discontinuous finite volume element method for the Allen-Cahn equation, *Adv. Comput. Math.*, **49** (2023), 55. https://doi.org/10.1007/s10444-023-10031-5

16. R. Li, Y. Gao, W. Yan, Z. Chen, A Crank-Nicolson discontinuous finite volume element method for a coupled non-stationary Stokes-Darcy problem, *J. Comput. Appl. Math.*, **353** (2019), 86–112. https://doi.org/10.1016/j.cam.2018.12.025

17. J. Liu, L. Mu, X. Ye, R. Jari, Convergence of the discontinuous finite volume method for elliptic problems with minimal regularity, *J. Comput. Appl. Math.*, **236** (2012), 4537–4546. https://doi.org/10.1016/j.cam.2012.05.009

18. Y. Liu, H. Yang, Z. Xie, P. Qin, R. Li, Parallel simulation of variably saturated soil water flows by fully implicit domain decomposition methods, *J. Hydrol.*, **582** (2020), 124481. https://doi.org/10.1016/j.jhydrol.2020.124481

19. M. Th. Van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, **44** (1980), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

20. F. Wang, P. Wang, M. Zhao, C. Shan, Z. Yang, The power of modality: improving polyp segmentation with multimodal information, *IET Image Process.*, **20** (2026), e70305. https://doi.org/10.1049/ipr2.70305

21. X. Ye, A new discontinuous finite volume method for elliptic problems, *SIAM J. Numer. Anal.*, **42** (2004), 1062–1072. https://doi.org/10.1137/S0036142902417042

22. X. Ye, A discontinuous finite volume method for the Stokes problems, *SIAM J. Numer. Anal.*, **44** (2006), 183–198. https://doi.org/10.1137/040616759