Numerical Solution of the Viscous Surface Wave with Discontinuous Galerkin Method

Lei Wu and Chi-Wang Shu

Abstract. We consider an incompressible viscous flow without surface tension in a finite-depth domain of two dimensions, with free top boundary and fixed bottom boundary. This system is governed by the Navier-Stokes equations in this moving domain and the transport equation on the moving boundary. In this paper, we construct a stable numerical scheme to simulate the evolution of this system by discontinuous Galerkin method, and discuss the error analysis of the fluid under certain assumptions. Our formulation is mainly based on the geometric structure introduced in [10], [11] and [20], and the natural energy estimate, which is rarely used in the numerical study of this system before.

Keywords: Stability, Free Boundary, Navier-Stokes equation.

1. Introduction

We consider an incompressible viscous flow in the moving domain

\begin{equation}
\Omega(t) = \{ y = (y_1, y_2) \in \Sigma \times \mathbb{R} : -1 < y_2 < \eta(y_1, t) \},
\end{equation}

where \( \Sigma = \mathbb{T} \) for which \( \mathbb{T} \) denotes the 1-torus. We denote the initial domain \( \Omega(0) = \Omega_0 \). For each \( t \), the flow is described by its velocity and pressure \((u, p) : \Omega(t) \mapsto \mathbb{R}^2 \times \mathbb{R} \) which satisfies the incompressible Navier-Stokes equations

\begin{equation}
\begin{cases}
\partial_t u + u \cdot \nabla u + \nabla p = \nu \Delta u + S & \text{in } \Omega(t), \\
\nabla \cdot u = 0 & \text{in } \Omega(t), \\
(p I - \nu \nabla)(u) \mu = g \eta \mu & \text{on } \{ y_2 = \eta(y_1, t) \}, \\
u = 0 & \text{on } \{ y_2 = -1 \}, \\
\partial_\eta \eta = u_2 - u_1 \partial_{y_2} \eta & \text{on } \{ y_2 = \eta(y_1, t) \}, \\
u(t = 0) = u_0 & \text{in } \Omega_0, \\
\eta(t = 0) = \eta_0 & \text{on } \Sigma,
\end{cases}
\end{equation}

for \( \mu \) the outward-pointing unit normal vector on \( \{ y_2 = \eta \} \), \( I \) the \( 2 \times 2 \) identity matrix, \((\nabla u)_{ij} = \partial_i u_j + \partial_j u_i\) the symmetric gradient of \( u \), \( g \) the gravitational constant, \( \nu > 0 \) the viscosity and \( S = S(t, y_1, y_2) \) an external source term. The fifth equation in (1.2) implies the free surface is convected with the fluid. Note in (1.2), we have shifted the actual pressure \( \bar{p} \) by the constant atmosphere pressure \( p_{atm} \) according to \( p = \bar{p} + g y_2 - p_{atm} \).

We always assume the natural condition that there exists a positive number \( \delta \) such that \( \eta_0 + 1 \geq \delta > 0 \) on \( \Sigma \), which means the initial free surface is strictly separated from the bottom.

Traditionally, based on the handling of the free surface, this type of problems can be solved via moving-grid technique as in [17], marker-and-cell method as in [12], volume-of-fluid method as in [13] and level-set method as in [9]. In each case, finite difference method, finite volume method and finite element method can be applied to solve the Navier-Stokes equation in \( \Omega(t) \). However, to the best of authors’ knowledge, there is very little in the literature on solving this problem with discontinuous Galerkin method other than [9]. On the other hand, in spite of many computational tests presented in the literature, there is very little discussion on the stability and convergence of the numerical scheme. In this paper, we employ the idea from [10], [11] and [20], to construct a stable numerical scheme and give detailed analysis.

Our central idea is to flatten the free surface via a coordinates transform. First, we define a fixed domain

\begin{equation}
\Omega = \{ x = (x_1, x_2) \in \Sigma \times \mathbb{R} \mid -1 < x_2 < 0 \},
\end{equation}

2000 Mathematics Subject Classification. 35Q30, 35R35, 74S05.
Research Supported by DOE Grant DE-FG02-08ER25863 and NSF Grant DMS-1112700.
for which we write the coordinates \( x \in \Omega \). In this slab, we take \( \Sigma : \{x_2 = 0\} \) as the upper boundary and \( \Sigma_b : \{x_2 = -1\} \) as the lower boundary.

Consider the geometric transform from \( \Omega \) to \( \Omega(t) \), which is first introduced in \([2]\) and further extended in \([10]\) and \([20]\):

\[
\Phi : (x_1, x_2) \mapsto (x_1, x_2 + \eta(1 + x_2)) = (y_1, y_2).
\]

We may directly verify this transform maps \( \Omega \) into \( \Omega(t) \) with the Jacobi matrix

\[
\nabla \Phi = \begin{pmatrix} 1 & 0 \\ A & J \end{pmatrix},
\]

and the transform matrix

\[
\mathcal{A} = ((\nabla \Phi)^{-1})^T = \begin{pmatrix} 1 & -AK \\ 0 & K \end{pmatrix},
\]

where

\[
\tilde{b} = 1 + x_2, \quad A = \partial_1 \eta \tilde{b}, \quad J = 1 + \eta, \quad K = 1/J.
\]

Here we denote the derivative with respect to \( x_1 \) as \( \partial_1 \) and with respect to \( x_2 \) as \( \partial_2 \). Define the transformed operators as follows:

\[
\begin{align*}
(\nabla_{\mathcal{A}} f)_i &= \mathcal{A}_{ij} \partial_j f, \\
\nabla_{\mathcal{A}} \cdot \vec{g} &= \mathcal{A}_{ij} \partial_j g_i, \\
\Delta_{\mathcal{A}} f &= \nabla_{\mathcal{A}} \cdot \nabla_{\mathcal{A}} f, \\
\mathcal{N} &= (-\partial_1 \eta, 1), \\
\chi &= \partial_1 \eta, \\
(\mathcal{D}_{\mathcal{A}} u)_{ij} &= \mathcal{A}_{ik} \partial_k u_j + \mathcal{A}_{jk} \partial_k u_i, \\
S_{\mathcal{A}}(p, u) &= pI - \mathcal{D}_{\mathcal{A}} u,
\end{align*}
\]

where the summation index should be understood in the Einstein convention. If we extend the divergence \( \nabla_{\mathcal{A}} \cdot \) to act on symmetric tensors in the natural way, then a straightforward computation reveals \( \nabla_{\mathcal{A}} \cdot u = 0 \).

In our new coordinates, the original system \((1.2)\) becomes

\[
\begin{aligned}
\begin{cases}
\partial_t u - \partial_1 \tilde{b} K \partial_2 u + u \cdot \nabla_{\mathcal{A}} u - \nu \Delta_{\mathcal{A}} u + \nabla_{\mathcal{A}} p = S & \text{in } \Omega, \\
\nabla_{\mathcal{A}} \cdot u = 0 & \text{in } \Omega, \\
S_{\mathcal{A}}(p, u) \mathcal{N} = g \eta \mathcal{N} & \text{on } \Sigma, \\
uu &= 0 & \text{on } \Sigma_b, \\
\partial_2 \eta &= u \cdot \mathcal{N} & \text{on } \Sigma, \\
\eta(x', 0) &= \eta_0(x') & \text{on } \Sigma.
\end{cases}
\end{aligned}
\]

Since \( \mathcal{A} \) depends on \( \eta \) through the transform, most of the operators in the Navier-Stokes equations are related to the free surface \( \eta \). Hence, the Navier-Stokes equations and the transport equation are essentially coupled.

Based on \([10]\), the equation \((1.8)\) with \( S = 0 \) possesses a natural energy equality as follows:

\[
\int_\Omega J(t) |u(t)|^2 + g \int_\Sigma |\eta(t)|^2 + \nu \int_0^t \int_\Omega J(s) |\mathcal{D}_{\mathcal{A}} u(s)|^2 \, ds = \int_\Omega J(0) |u(0)|^2 + g \int_\Sigma |\eta(0)|^2.
\]

Hence, we try to construct a numerical scheme to recover this energy stability for the numerical solution.

In the following, we refer to the term “continuous case” when we consider the exact solution triple \((u, p, \eta)\) which is sufficiently smooth. On the other hand, we refer to the term “discrete case” when we consider the numerical solution triple \((u_h, p_h, \eta_h)\).

Throughout this paper, \( C > 0 \) denotes a positive constant that only depends on the parameters \( \Omega, g \) and \( \nu \) of the problem, and the exact solution \((u, p, \eta)\). It is referred as universal and can change from one inequality to another. When we write \( C(z) \), it means a positive constant depending on the quantity \( z \). \( a \lesssim b \) denotes \( a \leq Cb \), where \( C \) is a universal constant as defined above.
The method we discuss in this paper belongs to the class of discontinuous Galerkin (DG) methods. DG methods are finite element methods which use completely discontinuous piecewise polynomial solution and test spaces. They are particularly useful for convection dominated wave problems, and have the advantage of flexibility in adaptivity and efficient parallel implementation. We refer to the references [3, 4, 5, 18] for more details.

2. Numerical Scheme

2.1. Fundamental Settings. In our construction and analysis of the numerical scheme, we focus on the semi-discrete form of the system. The time discretization is based on the Runge-Kutta method for differential-algebraic equations of index 2, as presented in [16]. We do not discuss the fully-discrete scheme in this paper.

We choose the bulk domain as \( \Omega : (x_1, x_2) \in [0, 1] \times [-1, 0] \) and the surface domain as \( \Sigma : x_1 \in [0, 1] \), which are 1-periodic in the \( x_1 \) direction. The surface mesh is defined by dividing \( \Sigma \) into \( N \) uniform elements with length \( h = 1/N \). The bulk mesh construction can be divided into two steps: First we divide \( \Omega \) into \( N \times N \) uniform squares, where each element is sized \( h \times h \). Then, we divide each square into two right triangles by cutting along the diagonal from the left-up corner to the right-down corner. The bulk mesh is shown in Figure 1.

Remark 2.1. Since the discretization of the weighted differential operators \( \nabla df, \nabla df \cdot \) and \( \Delta df \) depends on \( \eta \), i.e. the surface variable, we need the bulk discretization to match the surface discretization in the vertical direction. Hence, we need to first choose the strip-shape mesh as in Figure 2, whose projection on the upper boundary is exactly the surface mesh. Then in each strip, any triangulation is permitted. For convenience, we made the choice as in Figure 1.

![Figure 1. Mesh Distribution in N = 4](image1)

![Figure 2. Strip-Shape Mesh Distribution in N = 4](image2)

Let \( \mathcal{E}_h \) be the set consisting of all the triangular elements in the bulk mesh and \( \partial \mathcal{E}_h \) be the set of all the sides of the triangular cells. Let \( \mathcal{F}_h \) be the set consisting of all the interval elements in the surface mesh and \( \partial \mathcal{F}_h \) be the set of all the boundary points of intervals.

Define the usual Sobolev space in \( \Omega \) as \( H^k(\Omega) \) and in \( \Sigma \) as \( H^k(\Sigma) \). We define the space \( P_h^k(\Omega) \) where \( f \in P_h^k(\Omega) \) if and only if \( f|_E \in P^k(E) \) for all \( E \in \mathcal{E}_h \) (\( P^k(E) \) denotes the set of polynomials of degree at most \( k \) defined on \( E \)). Similarly, we can define the space \( P_h^k(\Sigma) \).

We define the solution spaces as

\[
\begin{align*}
X_h^k &= \{ u_h \in (L^2(\Omega))^2 : u_h \in (P_h^k(\Omega))^2 \}, \\
M_h^k &= \{ p_h \in L^2(\Omega) : p_h \in P_h^k(\Omega) \}, \\
S_h^k &= \{ \eta_h \in L^2(\Sigma) : \eta_h \in P_h^k(\Sigma) \},
\end{align*}
\]

for \( k \geq 1 \).
For discrete functions \( f_h \in X_h^k \) and \( g_h \in S_h^k \), we may define the \( H_h \) norms as follows:

\[
\| f_h \|_{H_h}^2 = \sum_{E \in \mathcal{E}_h} \| f_h \|_{H^1(E)}^2,
\]

\[
\| g_h \|_{H_h}^2 = \sum_{F \in \mathcal{F}_h} \| g_h \|_{H^1(F)}^2.
\]

Define the \( X_h \) norm for \( v_h \in X_h^k \) as

\[
\| v_h \|_{X_h}^2 = \sum_{E \in \mathcal{E}_h} \left( \| \nabla v_h \|_{L^2(E)}^2 + \int_{\partial E} \frac{1}{h} |v_h|^2 \right).
\]

Note the fact \( X_h \) and \( H_h \) norm can both be defined in the space \( X_h^k \).

Before stating the scheme, we announce the notations used below. For a given boundary belonging to a reference cell \( E \in \mathcal{E}_h \) and a function \( f \) defined in \( \Omega \), \( f^{ext(E)} \) denotes the value of \( f \) read from the exterior direction on the boundary and \( f^{int(E)} \) denotes that read from the interior direction. Also for the boundary shared by two cells, we define

\[
\{ f \} = \frac{1}{2} (f^{ext(E)} + f^{int(E)}),
\]

\[
[f] = f^{int(E)} - f^{ext(E)}.
\]

For the side on the boundary of \( \Omega \), we make a special definition that on \( \Sigma_h \), \( f^{int(E)} = \{ f \} = [f] \) and \( f^{ext(E)} = 0 \), while on \( \Sigma \), \( f^{int(E)} = f^{ext(E)} = \{ f \} \) and \( [f] = 0 \). The definitions are reasonable, since the system (1.8) for \( u \) gives Dirichlet-type condition on \( \Sigma_h \) and Neumann-type condition on \( \Sigma \).

The numerical scheme is based on the weak formulation of the system (1.8). Taking test functions \( v_h \in X_h^k \) and \( q_h \in M_h^{k-1} \). We multiply \( Jv_h \) and \( Jq_h \) respectively on both sides of the Navier-Stokes equations and integrate over \( E \in \mathcal{E}_h \) to obtain

\[
\int_E J \partial_t u \cdot v_h - \int_E J (\partial_t \eta \bar{K} \partial_2 u) \cdot v_h
+ \int_E J (u \cdot \nabla_{\text{sd}} u) \cdot v_h - \nu \int_E J \Delta_{\text{sd}} u \cdot v_h + \int_E J \nabla_{\text{sd}} p \cdot v_h = \int_E J (S \cdot v_h),
\]

\[
\int_E J (\nabla_{\text{sd}} \cdot u) q_h = 0.
\]

Considering the transport equation, we make an addition and subtraction to modify two terms in above formulation as

\[
\left( \int_E J \partial_t u \cdot v_h - \int_E J (\partial_t \eta \bar{K} \partial_2 u) \cdot v_h + \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_1 \eta (u \cdot v_h) \right)
+ \left( \int_E J (u \cdot \nabla_{\text{sd}} u) \cdot v_h - \frac{1}{2} \int_{\partial E \cap \Sigma} (u \cdot \mathcal{N}) (u \cdot v_h) \right) - \nu \int_E J \Delta_{\text{sd}} u \cdot v_h + \int_E J \nabla_{\text{sd}} p \cdot v_h = \int_E J (S \cdot v_h),
\]

\[
\int_E J (\nabla_{\text{sd}} \cdot u) q_h = 0,
\]

where \( \partial E \) denotes the boundary of the triangular element \( E \). This trick shows its power in the stability proof.

Multiplying a test functions \( \phi_h \in S_h^k \) on both sides of the transport equation and integrating over \( F \in \mathcal{F}_h \), we obtain

\[
\int_F \partial_t \eta \cdot \phi_h + \int_F \bar{u}_1 \partial_1 \eta \cdot \phi_h - \int_F \bar{u}_2 \cdot \phi_h = 0,
\]

where \( \bar{u}_1 \) and \( \bar{u}_2 \) denote the traces of \( u_1 \) and \( u_2 \) on \( \Sigma \).

The weak formulations of the system (1.8) are based on the integration by parts of (2.11), (2.12) and (2.13). In the following, we define and analyze the discrete solution term by term.
2.2. Transport Discretization. We first study the transport equation. The discretization can be divided into two steps:

Step 1: Calculation of $\eta_h$:

An integration by parts in (2.13) naturally implies for any test function $\phi_h \in S_h^k$,

$$
(2.14) \int_F \partial_t \eta \cdot \phi_h - \int_F \partial_t \bar{u}_1 \eta \cdot \phi_h - \int_F \bar{u}_1 \eta \cdot \partial_t \phi_h + \bar{u}_1 \eta \cdot \phi^+_h |_{F+1/2} - \bar{u}_1 \eta \cdot \phi^-_h |_{F-1/2} - \int_F \bar{u}_2 \cdot \phi_h = 0,
$$

where $F-1/2$ and $F+1/2$ denote the left and right boundary point of $F$. Hence, given $u_h = ((u_1)_h, (u_2)_h) \in X_h^k$, we define $\eta_h$ as the solution of the following

$$
(2.15) \kappa_F(\eta_h, u_h, \phi_h) = \int_F \partial_t \eta_h \cdot \phi_h - \int_F (\bar{u}_2)_h \cdot \phi_h - \int_F \partial_t (\bar{u}_1)_h \eta_h \cdot \phi_h - \int_F (\bar{u}_1)_h \eta_h \cdot \partial_t \phi_h
$$

$$
+ (\bar{u}_1)_h \bar{\eta}_h \cdot \phi^+_h |_{F+1/2} - (\bar{u}_1)_h \bar{\eta}_h \cdot \phi^-_h |_{F-1/2} + \frac{1}{2} \left( \eta^-_h \phi^-_h [u_h]|_{F+1/2} - \eta^+_h \phi^+_h [u_h]|_{F-1/2} \right) = 0,
$$

for any test function $\phi_h \in S_h^k$. Here, for each boundary point, “−” means the value read from the left and “+” means the value read from the right. Also, $(\bar{u}_1)_h \bar{\eta}_h$ denotes the numerical flux on the boundary. Note

$$
\frac{1}{2} \left( \eta^-_h \phi^-_h [u_h]|_{F+1/2} - \eta^+_h \phi^+_h [u_h]|_{F-1/2} \right)
$$

is an extra penalty term, which helps to show the stability. It is easy to see this term vanishes in the continuous case. Hence, we can directly verify our scheme is consistent. We always take

$$
(2.16) (\bar{u}_1)_h = \{(\bar{u}_1)_h\},
$$

and $\bar{\eta}_h$ should be determined from the sign of $\{(\bar{u}_1)_h\}$ at the boundary following the upwinding rule as

$$
\bar{\eta}_h = \begin{cases} 
\eta^-_h & \text{if } \{(\bar{u}_1)_h\} \geq 0, \\
\eta^+_h & \text{if } \{(\bar{u}_1)_h\} < 0.
\end{cases}
$$

Similarly, $(\bar{u}_1)_h$ and $(\bar{u}_2)_h$ are the traces of $(u_1)_h$ and $(u_2)_h$ on $\Sigma$.

Step 2: Calculation of $\chi_h$:

Since we have the relation $\chi = \partial_t \eta$, it naturally leads to the weak formulation that for any test function $\psi_h$, the following holds

$$
(2.17) \int_F \chi \cdot \psi_h + \int_F \eta \cdot \partial_t \psi_h - \eta \cdot \psi^-_h |_{F+1/2} + \eta \cdot \psi^+_h |_{F-1/2} = 0.
$$

We may simplify take $\chi_h \in S_h^k$ as the solution of the following

$$
(2.18) \chi_F(\chi_h, \psi_h) = \int_F \chi_h \cdot \psi_h + \int_F \eta_h \cdot \partial_t \psi_h - \eta^-_h \cdot \psi^-_h |_{F+1/2} + \eta^+_h \cdot \psi^+_h |_{F-1/2} = 0,
$$

for any test function $\psi_h \in S_h^k$.

In all the applications below, we use $\chi_h$ to discretize $\partial_t \eta$, and $\eta_h$ to discretize $\eta$. Hence, we denote $\mathcal{M}_h$ for the vector $(-\chi_h, 1)$. Also, $A_h$, $J_h$, $K_h$ and $\mathcal{A}_h$ can be defined in the same convention.

2.3. Temporal Discretization. We consider the temporal terms, i.e.

$$
(2.19) \left( \int_E J \partial_t u \cdot v_h - \int_E J (\partial_t \eta \bar{K} \partial_2 u) \cdot v_h + \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta (u \cdot v_h) \right).
$$
In the continuous case, a direct integration by parts reveals

\[ \int_E J \partial_t u \cdot v_h = \int_E \partial_t (J u) \cdot v_h - \int_E \partial_t J (u \cdot v_h), \]

\[ - \int_E J (\partial_t \tilde{b} \tilde{K} \partial_2 u) \cdot v_h = - \int_E \partial_t \tilde{b} \partial_2 u \cdot v_h \]

\[ = \int_E \partial_t \tilde{b} u \cdot \partial_2 v_h + \int_E \partial_2 (\partial_t \tilde{b})(u \cdot v_h) - \int_{\partial E} \partial_t \tilde{b} (u \cdot v_h) n_2. \]

Here \( n_E = (n_1, n_2) \) denotes the outward normal vector on \( \partial E \). Since \( \partial_t J = \partial_2 (\partial_t \tilde{b}) \), we can simplify above terms into

\[ \int_E J \partial_t u \cdot v_h - \int_E J (\partial_t \tilde{b} \tilde{K} \partial_2 u) \cdot v_h = \int_E \partial_t (J u) \cdot v_h + \int_E \partial_t \tilde{b} u \cdot \partial_2 v_h - \int_{\partial E} \partial_t \tilde{b} (u \cdot v_h) n_2. \]

Hence, we may define the discretization of the temporal term as \( \zeta_E(\eta, v_h) \) which satisfies

\[ \zeta_E(\eta, u_h, v_h) = \int_E \partial_t (J_h u_h) \cdot v_h + \int_E \partial_t \eta_h \tilde{b} u_h \cdot \partial_2 v_h - \int_{\partial E} \partial_t \eta_h \tilde{b} (u_h \cdot v_h^{\text{int}}) n_2 + \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h (u_h \cdot v_h). \]

The last term is the newly-added penalty term in the formulation (2.11). The argument \( \eta \) in \( \zeta_E(\eta, u_h, v_h) \) denotes \( \eta \) and its derivatives. It is easy to show this discretization is consistent. Since in our bulk mesh, there are only two effective boundary integrals for each triangle, we may call them the upper boundary \( \partial E + 1/2 \) and lower boundary \( \partial E - 1/2 \) without ambiguity. If we denote \( f(u_h) = -\partial_t \eta_h \tilde{b} u_h \), then above scheme is actually

\[ \zeta_E(\eta, u_h, v_h) = \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h (u_h \cdot v_h) \]

\[ = \int_E \partial_t (J_h u_h) \cdot v_h - \int_E f(u_h) \cdot \partial_2 u_h + \int_{\partial E + 1/2} f(u_h) \cdot v_h^+ - \int_{\partial E - 1/2} f(u_h) \cdot v_h^- \]

Since our flux only appears in the vertical direction, for simplicity, we can still use the notation “+” to denote the value read from the up direction and “−” denote that from the down direction. We utilize the upwinding flux in this scheme, i.e.

\[ f(u_h) = f(\tilde{u}_h) = \begin{cases} f(u_h^+) & \text{if } \partial_t \eta_h \tilde{b} \geq 0, \\ f(u_h^-) & \text{if } \partial_t \eta_h \tilde{b} < 0. \end{cases} \]

Lemma 2.2. If we take \( v_h = u_h \), the discretized temporal term satisfies

\[ \sum_{E \in \mathcal{E}_h} \zeta_E(\eta, u_h, u_h) \geq \frac{1}{2} \partial_t \int_{\Omega} J_h |u_h|^2. \]

Proof. We can directly compute

\[ \zeta_E(\eta, u_h, u_h) - \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h |u_h|^2 \]

\[ = \int_E \partial_t (J_h u_h) \cdot u_h - \int_E f(u_h) \cdot \partial_2 u_h + \int_{\partial E + 1/2} f(u_h) \cdot u_h^- - \int_{\partial E - 1/2} f(u_h) \cdot u_h^+ \]

\[ = \frac{1}{2} \partial_t \int_E J_h |u_h|^2 + \frac{1}{2} \int_E \partial_t J_h |u_h|^2 - \frac{1}{2} \int_E \partial_2 (\partial_t \eta_h \tilde{b}) |u_h|^2 \]

\[ - \frac{1}{2} \int_{\partial E + 1/2} f(u_h^-) \cdot u_h^- + \frac{1}{2} \int_{\partial E - 1/2} f(u_h^+) \cdot u_h^- + \int_{\partial E + 1/2} f(u_h^+) \cdot u_h^+ - \int_{\partial E - 1/2} f(u_h^-) \cdot u_h^+ \]

\[ = \frac{1}{2} \partial_t \int_E J_h |u_h|^2 \]

\[ - \frac{1}{2} \int_{\partial E + 1/2} f(u_h^-) \cdot u_h^- + \frac{1}{2} \int_{\partial E - 1/2} f(u_h^+) \cdot u_h^- + \int_{\partial E + 1/2} f(u_h^+) \cdot u_h^+ - \int_{\partial E - 1/2} f(u_h^-) \cdot u_h^+ \]

\[ = \frac{1}{2} \partial_t \int_E J_h |u_h|^2 + \hat{F}_{\partial E + 1/2} - \hat{F}_{\partial E - 1/2} + \Theta_E, \]
where

\begin{align}
\hat{F}_{\partial E+1/2} &= -\frac{1}{2} \int_{\partial E+1/2} f(u_h^-) \cdot u_h^- + \int_{\partial E+1/2} \hat{f}(u_h) \cdot u_h^- , \\
\hat{F}_{\partial E-1/2} &= -\frac{1}{2} \int_{\partial E-1/2} f(u_h^-) \cdot u_h^- + \int_{\partial E-1/2} \hat{f}(u_h) \cdot u_h^- ,
\end{align}

and

\begin{equation}
\Theta_E = \frac{1}{2} \int_{\partial E-1/2} \left( f(u_h^+ \cdot u_h^+ - f(u_h^-) \cdot u_h^- \right) + \int_{\partial E-1/2} \left( \hat{f}(u_h) \cdot u_h^- - f(u_h^-) \cdot u_h^- \right).
\end{equation}

Based on the flux definition (2.26), we have the estimate

\begin{equation}
\zeta_E(\eta_h, u_h, u) = \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h |u_h|^2 \geq \frac{1}{2} \partial_t \int_E J_h |u_h|^2 + \hat{F}_{\partial E+1/2} - \hat{F}_{\partial E-1/2}.
\end{equation}

When summing up over all $E \in \mathcal{E}_h$, we can easily see when $\partial E \subset \Omega$, all the terms involving $\hat{F}$ are canceled out. Therefore, only the terms on $\partial \Omega$ remain, i.e.

\begin{equation}
\sum_{E \in \mathcal{E}_h} \zeta_E(\eta_h, u_h, u_h) - \frac{1}{2} \int_{\partial \Omega} \partial_t \eta_h |u_h|^2 \geq \frac{1}{2} \partial_t \int_E J_h |u_h|^2 + \sum_{\partial E \cap \Sigma \neq \emptyset} \hat{F}_{\partial E+1/2} - \sum_{\partial E \cap \Sigma \neq \emptyset} \hat{F}_{\partial E-1/2}.
\end{equation}

On $\Sigma$, we have $u_h^- = 0$, so $-\hat{F}_{\partial E-1/2} = 0$, i.e.

\begin{equation}
\sum_{\partial E \cap \Sigma \neq \emptyset} \hat{F}_{\partial E+1/2} = 0.
\end{equation}

On $\Sigma$, we have $u_h^+ = u_h^- = u_h|\Sigma$ and we always take $\hat{u}_h = u_h$, so it implies

\begin{equation}
\hat{F}_{\partial E+1/2} = -\frac{1}{2} \int_{\partial E+1/2} f(u_h^-)u_h^- + \int_{\partial E+1/2} \hat{f}(u_h^-)u_h^- = -\frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h |u_h|^2 .
\end{equation}

Then summing up over $E \in \mathcal{E}_h$ gives a full integration over $\Sigma$, i.e.

\begin{equation}
\sum_{E \in \mathcal{E}_h} \hat{F}_{\partial E+1/2} = -\frac{1}{2} \int_{\Sigma} \partial_t \eta_h |u_h|^2.
\end{equation}

Therefore, combining (2.34), (2.35) and (2.37), we deduce

\begin{equation}
\sum_{E \in \mathcal{E}_h} \zeta_E(\eta_h, u_h, u_h) - \frac{1}{2} \int_{\Sigma} \partial_t \eta_h |u_h|^2 \geq \frac{1}{2} \partial_t \int_E J_h |u_h|^2 - \frac{1}{2} \int_{\Sigma} \partial_t \eta_h |u_h|^2 .
\end{equation}

Hence, our result easily follows.

\[ \square \]

2.4. Convection Discretization. We consider the convection term

\begin{equation}
\int_E J(u \cdot \nabla u) \cdot v_h - \frac{1}{2} \int_F (u \cdot N) (u \cdot v_h).
\end{equation}

This is the key nonlinear term in the Navier-Stokes equations. Inspired by the idea in [7], we introduce a discretized multi-linear form. We define the discretization of the convection term as $\gamma_E(\eta_h, u_h, v_h, v)$ which satisfies

\begin{equation}
\gamma_E(\eta_h, u_h, v_h, v) = -\frac{1}{2} \int_{\partial E \cap \Sigma} (u_h \cdot N_h)(u_h \cdot v_h) + \int_E J_h(u_h \cdot \nabla \phi_h) \cdot v_h
\end{equation}

\[ + \frac{1}{2} \int_E J_h(\nabla \phi_h \cdot u_h)(u_h \cdot v_h) - \frac{1}{2} \int_{\partial E \cap \partial \Omega} [u_h \cdot J_h \phi_h] \cdot n_E \frac{u^{\text{int}(E)} \cdot v^{\text{int}(E)}}{2}
\]

\[ + \int_{\partial E^-} (u_h \cdot J_h \phi_h) \cdot n_E (u_h^{\text{int}(E)} - u_h^{\text{ext}(E)}) \cdot v_h^{\text{int}(E)},\]
where
\[ \partial E^- = \{ x \in \partial E : \{ u_h \cdot J_h A_h \} \cdot n_E < 0 \}. \]
The first term is the newly added penalty term in the formulation (2.11). Since
\[ J_h A_h = \begin{pmatrix} J_h & -A_h \\ 0 & 1 \end{pmatrix}, \]
we only need one \( \eta_h \) argument in \( \gamma_E(u_h, \eta_h, u_h, v_h) \).

In the continuous case, when summing up over all \( E \in \mathcal{E}_h \), we can see
\[ \sum_{E \in \mathcal{E}_h} \gamma_E(u, \eta, u, v_h) = \int_{\Omega} J(u \cdot \nabla u) \cdot v_h + \frac{1}{2} \int_{\Omega} J(\nabla u \cdot u)(u \cdot v_h) - \frac{1}{2} \int_{\Sigma} (u \cdot N)(u \cdot v_h), \]
where all the other boundary terms vanish. Hence, considering the continuous \( \mathcal{A} \)-divergence-free condition for \( u \), this discretization is consistent.

**Lemma 2.3.** If we take \( v_h = u_h \), the discretized convection term satisfies
\[ \sum_{E \in \mathcal{E}_h} \gamma_E(u_h, \eta_h, u_h, u_h) \geq 0. \]

**Proof.** We divide the proof into several steps:

**Step 1: Direct integration by parts.**
We plug the test function \( v_h = u_h \) into (2.40) to obtain
\[ \gamma_E(u_h, \eta_h, u_h, u_h) + \frac{1}{2} \int_{E} (u_h \cdot \mathcal{A}_h) |u_h|^2 = \int_{E} J_h (u_h \cdot \nabla A_h u_h) \cdot u_h + \frac{1}{2} \int_{E} J_h (\nabla A_h \cdot u_h)(u_h \cdot u_h) \]
\[ - \frac{1}{2} \int_{\partial E \setminus \partial T} [u_h \cdot J_h A_h] \cdot n_E \frac{u_h^{int(E)} \cdot u_h^{int(E)}}{2} + \int_{\partial E^-} |\{ u_h \cdot J_h A_h \} \cdot n_E | (u_h^{int(E)} - u_h^{ext(E)}) \cdot u_h^{int(E)}. \]
A direct integration by parts yields
\[ \int_{E} J_h (u_h \cdot \nabla A_h u_h) \cdot u_h = - \int_{E} J_h (u_h \cdot \nabla A_h u_h) \cdot u_h - \int_{E} J_h (\nabla A_h \cdot u_h)(u_h \cdot u_h) \]
\[ + \int_{\partial E} (u_h^{int(E)} \cdot J_h^{int(E)} A_h^{int(E)} \cdot n_E)(u_h^{int(E)} \cdot u_h^{int(E)}). \]
Therefore, (2.45) can be simplified as
\[ \gamma_E(u_h, \eta_h, u_h, u_h) + \frac{1}{2} \int_{\partial E \setminus \Sigma} (u_h \cdot \mathcal{A}_h) |u_h|^2 \]
\[ = -(\int_{E} J_h (u_h \cdot \nabla A_h u_h) \cdot u_h + \frac{1}{2} \int_{E} J_h (\nabla A_h \cdot u_h)(u_h \cdot u_h) \]
\[ - \frac{1}{2} \int_{\partial E \setminus \partial T} [u_h \cdot J_h A_h] \cdot n_E \frac{u_h^{int(E)} \cdot u_h^{int(E)}}{2} \]
\[ + \int_{\partial E^-} |\{ u_h \cdot J_h A_h \} \cdot n_E | (u_h^{int(E)} - u_h^{ext(E)}) \cdot u_h^{int(E)} \]
\[ + \int_{\partial E} (u_h^{int(E)} \cdot J_h^{int(E)} A_h^{int(E)} \cdot n_E)(u_h^{int(E)} \cdot u_h^{int(E)}). \]
where
\[ I = -(\int_{E} J_h (u_h \cdot \nabla A_h u_h) \cdot u_h + \frac{1}{2} \int_{E} J_h (\nabla A_h \cdot u_h)(u_h \cdot u_h), \]
and II, III and IV can be understood respectively.

\[ (2.48) \]
Step 2: Estimates of $II$ and $IV$.
In $IV$, for $e \in \partial E \setminus \partial \Omega$, we have the decomposition

$$\tag{2.49} u_h^{\text{int}(E)} \cdot J_h^{\text{int}(E)} \cdot \mathcal{A}_h^{\text{int}(E)} = \{ u_h \cdot J_h \mathcal{A}_h \} + \frac{1}{2} \left( u_h^{\text{int}(E)} \cdot J_h^{\text{int}(E)} \mathcal{A}_h^{\text{int}(E)} - u_h^{\text{ext}(E)} \cdot J_h^{\text{ext}(E)} \mathcal{A}_h^{\text{ext}(E)} \right).$$

For $e \subset \partial E \setminus \partial \Omega$, we can sum up the second term on the right-hand side of (2.49) over $E$ and its neighboring cells, which have the opposite outward normal vectors, to show

$$\tag{2.50} \sum_{E \in \mathcal{E}_h} IV = \sum_{e \in \mathcal{E} \setminus \partial \Omega} \frac{1}{2} \int_e \left( u_h^{\text{int}(E)} \cdot J_h^{\text{int}(E)} \mathcal{A}_h^{\text{int}(E)} - u_h^{\text{ext}(E)} \cdot J_h^{\text{ext}(E)} \mathcal{A}_h^{\text{ext}(E)} \cdot n_E (u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) \right) + \sum_{e \in \mathcal{E} \setminus \partial \Omega} \int_e \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \}.
$$

where $n_e$ denotes the normal vector read from the same reference cell as $[u_h \cdot J_h \mathcal{A}_h]$. Note for $e \subset \partial E \cap \partial \Omega$, $u_h^{\text{int}(E)} \cdot J_h^{\text{int}(E)} \mathcal{A}_h^{\text{int}(E)} = \{ u_h \cdot J_h \mathcal{A}_h \}$. Therefore, we obtain

$$\tag{2.51} \sum_{E \in \mathcal{E}_h} IV = \sum_{e \in \mathcal{E} \setminus \partial \Omega} \frac{1}{2} \int_e \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \}.$$

Also, in $II$, for $e \subset \partial E \setminus \partial \Omega$, we can sum up over $E$ and its neighboring cells to achieve

$$\tag{2.52} -\frac{1}{2} \int_e \int_{\partial E} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \} = -\frac{1}{2} \int_e \int_{\partial E} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \},$$

which further leads to

$$\tag{2.53} \sum_{E \in \mathcal{E}_h} II = -\sum_{e \in \mathcal{E} \setminus \partial \Omega} \frac{1}{2} \int_e \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \}.$$

Step 3: Further estimates in (2.47).
Hence, summing over all $E \in \mathcal{E}_h$ in (2.47) and combining (2.51) and (2.53), we deduce

$$\tag{2.54} \sum_{E \in \mathcal{E}_h} \gamma_E (u_h, \eta_h, u_h, u_h) + \frac{1}{2} \int_{\partial \Omega} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot u_h \leq \sum_{e \in \mathcal{E} \setminus \partial \Omega} \frac{1}{2} \int_e \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \}$$

where $\gamma_E (u_h, \eta_h, u_h, u_h)$ is defined as

$$\gamma_E (u_h, \eta_h, u_h, u_h) = \int_{\partial E} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_e \{ u_h \cdot u_h \}. $$

We can further estimate the last two terms in (2.54) as follows:

$$\tag{2.55} \sum_{E \in \mathcal{E}_h} \left[ \int_{\partial E^-} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_E (u_h^{\text{int}(E)} - u_h^{\text{ext}(E)}) \cdot u_h^{\text{int}(E)} \right]
+ \int_{\partial E} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_E (u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)})$$

$$+ \sum_{E \in \mathcal{E}_h} \left[ \int_{\partial E^+ \cap \partial \Omega} \{ u_h \cdot J_h \mathcal{A}_h \} \cdot n_E (u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) \right]
+ \sum_{E \in \mathcal{E}_h} \int_{\partial E^+ \cap \partial \Omega} \left[ u_h \cdot J_h \mathcal{A}_h \right] \cdot n_E (u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}).$$
Then combining (2.54) and (2.55), we have the complete form

\[
\begin{align*}
(2.56) \quad & \sum_{E \in \mathcal{E}_h} \gamma_E(u_h, \eta_h, u_h, u_h) + \frac{1}{2} \int_{\Sigma} (u_h \cdot \mathcal{M}_h) |u_h|^2 \\
& = - \left( \int_{\Omega} J_h(u_h \cdot \nabla \phi_f u_h) \cdot u_h + \frac{1}{2} \int_{\Omega} J_h(\nabla \phi_f u_h)(u_h \cdot u_h) \\
& \quad - \sum_{e \in \partial \mathcal{E}_h} \int_{e} \frac{1}{2} \nabla \phi_f \cdot (u_h \cdot J_h \mathcal{S}_h) \cdot n_E(u_h \cdot u_h) \right) + \sum_{E \in \mathcal{E}_h} \int_{\partial E^-} \left| \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E \right| u_h^{\text{ext}(E)} \cdot (u_h^{\text{int}(E)} - u_h^{\text{int}(E)}) \\
& \quad + \sum_{E \in \mathcal{E}_h} \int_{\partial E^+ \cap \partial \Omega} \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E(u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) + \sum_{E \in \mathcal{E}_h} \int_{\partial E^- \cap \Sigma} \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E(u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) \].
\end{align*}
\]

Step 4: Synthesis.
Summing up (2.45) over \( E \in \mathcal{E}_h \) and adding it to (2.56) imply

\[
\begin{align*}
(2.57) \quad & \sum_{E \in \mathcal{E}_h} \gamma_E(u_h, \eta_h, u_h, u_h) + \frac{1}{2} \int_{\Sigma} (u_h \cdot \mathcal{M}_h) |u_h|^2 \\
& = \frac{1}{2} \sum_{E \in \mathcal{E}_h} \int_{\partial E^-} \left| \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E \right| u_h^{\text{ext}(E)} - u_h^{\text{int}(E)} \right|^2 \right) + \frac{1}{2} \sum_{E \in \mathcal{E}_h} \int_{\partial E^+ \cap \partial \Omega} \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E(u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) \\
& \quad + \sum_{E \in \mathcal{E}_h} \int_{\partial E^- \cap \Sigma} \{ u_h \cdot J_h \mathcal{S}_h \} \cdot n_E(u_h^{\text{int}(E)} \cdot u_h^{\text{int}(E)}) \].
\end{align*}
\]

where the last equality can be directly verified by the definitions of \( J_h, \mathcal{S}_h \) and \( \mathcal{M}_h \). Then our result naturally follows. \( \square \)

2.5. Diffusion Discretization. In order to show the coercivity of the diffusion term, we need the discrete form of Korn’s inequality. The proof here is based on [19].

**Lemma 2.4.** Assume \( f_h \in X_h^k \) satisfies for \( e \subset \partial \mathcal{E}_h \cap \Sigma_h \), it holds that \( f_h^{\text{ext}}|_e = 0 \). Then for sufficiently large \( \sigma_0 > 0 \), we have

\[
(2.58) \quad \sum_{E \in \mathcal{E}_h} \int_E |Df_h|^2 + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e |f_h|^2 \geq \|f_h\|^2_{X_h} \geq \|f_h\|^2_{H_h}.
\]

**Proof.** The second inequality has been shown in [7], so we turn to the first one. It is easy to see the key part is to show the derivatives of \( f_h \) can be controlled. Hence, we only need to show

\[
(2.59) \quad \sum_{E \in \mathcal{E}_h} \int_E |Df_h|^2 + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e |f_h|^2 \geq \|f_h\|^2_{H_h}.
\]

If this is not true, then we can construct a sequence \( f_h^n \in X_h^k \) satisfying

\[
(2.60) \quad \|f_h^n\|_{H_h} = 1,
\]

and

\[
(2.61) \quad \sum_{E \in \mathcal{E}_h} \|Df_h^n\|_{L^2(E)} + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e |f_h^n|^2 \leq \frac{1}{n}.
\]

To abuse the notations, we can extract the weakly convergent subsequence

\[
(2.62) \quad f_h^n \rightharpoonup f_h \text{ in } H^1(E) \text{ for } \forall E \in \mathcal{E}_h.
\]
By the compact embedding theorem in each cell $E$ and the weak lower semi-continuity of $H^1(E)$ norm, we have the strongly convergent subsequence

\[(2.63) \quad f^n_h \to f_h \text{ in } L^2(\Omega).\]

By (2.61), we also have

\[(2.64) \quad \sum_{E \in \mathcal{E}_h} \|\nabla f^n_h\|_{L^2(\Omega)} \to 0.\]

Notice the fact that in each cell $E$, we still have the continuous Korn’s inequality

\[(2.65) \quad \|f^n_h - f^n_h\|_{H^1(E)} \lesssim \|\nabla f^n_h - \nabla f^n_h\|_{L^2(E)} + \|f^n_h - f^n_h\|_{L^2(E)} \quad \text{for } \forall m, n \in \mathbb{N}.\]

Hence, combining all above, we know $f^n_h$ is a Cauchy’s sequence under the norm $H^1(E)$, which means it is a Cauchy’s sequence under the norm $H_h$. Then

\[(2.66) \quad f^n_h \to f_h \text{ in } H_h.\]

Thus this means

\[(2.67) \quad \|f_h\|_{H_h} = 1,\]

which further implies

\[(2.68) \quad \nabla f_h = 0 \text{ in } \forall E \in \mathcal{E}_h,\]

and

\[(2.69) \quad \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e |f_h|^2 = 0.\]

This means

\[(2.70) \quad f_h = (a, b) + c(-x_1, x_2),\]

for some constant $a, b, c$ and $f_h$ is continuous in $\bar{\Omega}$. Certainly, the zero bottom implies $f_h = 0$, which contradicts $\|f_h\|_{H_h} = 1$. Therefore, our hypothesis is invalid and (2.59) holds.

\[\square\]

**Lemma 2.5.** Assume $f_h$ satisfies for $e \subset \partial \mathcal{E}_h \cap \Sigma_b$, it holds that $f_h^{ext}|_e = 0$. Also, $\eta_h$ satisfies $\eta_h + 1 \geq \delta > 0$ and

\[(2.71) \quad Q = \sup_{F \in \mathcal{F}_h} \|\eta_h\|_{W^{1,\infty}(F)} < \infty.\]

Then for sufficiently large $\sigma_0 > 0$, we have

\[(2.72) \quad \sum_{E \in \mathcal{E}_h} \int_E J_h \|\nabla \phi_h f_h\|^2 + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e |f_h|^2 \gtrsim C(Q) \|f_h\|^2_{H_h} \gtrsim C(Q) \|f_h\|^2_{H_h}.\]

**Proof.** Note that in each cell $E$, the free surface $\eta_h$ is smooth. Hence, we can change it back by the transform $\Phi^{-1}_h$ to a curved cell $\Phi^{-1}_h E$ which satisfies

\[(2.73) \quad \int_E J_h(x) \|\nabla \phi_h(x) f_h(x)\|^2 \, dx = \int_{\Phi^{-1}_h E} \|\nabla f_h(y)\|^2 \, dy.\]

Hence, in this curved cell, we have the Korn’s inequality

\[(2.74) \quad \|f_h\|_{H^1(\Phi^{-1}_h E)} \lesssim \|\nabla f_h\|_{L^2(\Phi^{-1}_h E)} + \|f_h\|_{L^2(\Phi^{-1}_h E)}.\]

Since the transform $\Phi^{-1}_h$ is a diffeomorphism between $E$ and $\Phi^{-1}_h (E)$, then the $H^1$ norms in these two spaces are comparable. Hence, we have

\[(2.75) \quad \|f_h\|_{H^1(E)} \lesssim C(Q) \left( \int_E J_h \|\nabla \phi_h f_h\|^2 + \|f_h\|_{L^2(E)} \right).\]

Then a similar proof as that of Lemma 2.4 naturally yields the desired result.  \[\square\]
Lemma 2.6. Assume $f_h$ satisfies for $e \subset \partial \mathcal{E}_h \cap \Sigma_h$, it holds that $f_h^{ext}|_e = 0$. Also, $\eta_h$ satisfies $\eta_h + 1 \geq \delta > 0$ and

$$Q = \sup_{F \in \mathcal{F}_h} \| \eta_h \|_{W^{1,\infty}(F)} < \infty.$$  

Then there exists a sufficiently small constant $\delta' > 0$ such that for $\eta'$ satisfying

$$\sup_{F \in \mathcal{F}_h} \| \eta'_h - \eta_h \|_{W^{1,\infty}(F)} \leq \delta',$$

and for sufficiently large $\sigma'_0 > 0$, we have

$$\sum_{E \in \mathcal{T}_h} \int_E J'_h \| \nabla \phi_h \|^2 + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma'_0}{h} \int_e [f_h]^2 \geq C(Q) \| f_h \|^2_{\mathcal{X}_h} \geq C(Q) \| f_h \|^2_{H_h}.$$  

Proof. By Lemma 2.5, we know

$$\sum_{E \in \mathcal{T}_h} \int_E J_h \| \nabla \phi_h \|^2 \geq \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e [f_h]^2 \geq C(Q) \| f_h \|^2_{\mathcal{X}_h} \geq C(Q) \| f_h \|^2_{H_h}.$$  

We can rewrite our formula in a perturbed form as

$$\sum_{E \in \mathcal{T}_h} \int_E J'_h \| \nabla \phi_h \|^2 - \sum_{E \in \mathcal{T}_h} \int_E J_h \| \nabla \phi_h \|^2 \leq \sum_{E \in \mathcal{T}_h} \int_E (J'_h - J_h) \| \nabla \phi_h \|^2 \leq \delta' \| f_h \|^2_{H_h}.$$  

Naturally, when taking $\delta'$ sufficiently small, we can absorb the perturbation into the principle part. Then our result easily follows.  

For the diffusion term

$$-\int_E J \nabla \phi \cdot \nu_h,$$

we discretize it based on the Symmetric Internal Penalty Galerkin Method (SIPG) or the Nonsymmetric Internal Penalty Galerkin Method (NIPG). Here, we follow the idea in [7]. We define the discretization of the diffusion term as $\alpha_E(\eta_h, u_h, \eta_h, v_h)$ which satisfies,

$$\alpha_E(\eta_h, u_h, \eta_h, v_h) = \frac{1}{2} \int_E J_h \nabla \phi_h \cdot \nabla \phi_h - \int_{\partial \mathcal{E}_h \setminus \Sigma} \{ \nabla \phi_h \cdot J_h \phi_h \} \cdot n_E \cdot v_h^{int(E)}$$

$$\pm \frac{1}{2} \int_{\partial \mathcal{E}_h \setminus \Sigma} \{ \nabla \phi_h \cdot J_h \phi_h \} \cdot n_E \cdot [u_h] + \frac{\sigma}{h} \int_{\partial \mathcal{E}_h \setminus \Sigma} [u_h] \cdot v_h^{int(E)}.$$  

For the third term, if we take $+$, it is NIPG and if we take $-$, it is SIPG. The penalty constant $\sigma$ can always be taken as 1 for NIPG and a sufficiently large number for SIPG, which is specified in the following. Note that we need two $\eta_h$ arguments in $\alpha_E(\eta_h, u_h, \eta_h, v_h)$ to denote the dependence of $\phi_h$.

A direct integration by parts in the continuous case implies

$$-\sum_{E \in \mathcal{T}_h} \int_E J \nabla \phi \cdot \nu_h = \frac{1}{2} \int_{\Omega} J \nabla \phi : \nabla \phi - \sum_{e \in \partial \mathcal{E}_h} \int_e \{ \nabla \phi \cdot J \phi \cdot n_E \} \cdot [v_h] - \int_{\Sigma} (\nabla \phi \cdot \nu_h)^{int(E)}.$$  

Also in the continuous case, above bilinear form reduces to

$$\sum_{E \in \mathcal{T}_h} \alpha_E(u, v, u, v) = \frac{1}{2} \int_{\Omega} J \nabla \phi : \nabla \phi - \sum_{e \in \partial \mathcal{E}_h} \int_e \{ \nabla \phi \cdot J \phi \cdot n_E \} \cdot [v_h].$$  

We can notice the difference is the extra physical boundary term $-\int_{\Sigma} (\nabla \phi \cdot \nu_h)^{int(E)}$, which contributes to the boundary condition on $\Sigma$. This is separately added to the scheme as a forcing term later combined with the pressure contribution. Hence, our scheme is consistent.
Lemma 2.7. Assume \( \eta_h \) satisfies \( \eta_h + 1 \geq \delta > 0 \). If we take \( v_h = u_h \) and choose the penalty constant \( \sigma \) properly, the discretized diffusion term with NIPG satisfies

\[
\sum_{E \in \mathcal{E}_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) \geq \int_\Omega J_h [\mathcal{D}_{\sigma h} u_h]^2.
\]

If we further assume \( \eta_h \) satisfies

\[
Q = \sup_{P \in \mathcal{P}_h} \| \eta_h \|_{W^{1, \infty}(P)} < \infty,
\]

then the discretized diffusion term with SIPG or NIPG satisfies

\[
\sum_{E \in \mathcal{E}_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) \geq C(Q) \| u_h \|_{X_h}^2.
\]

Proof. We can directly compute

\[
\sum_{E \in \mathcal{E}_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) = \frac{1}{2} \sum_{E \in \mathcal{E}_h} \int_E J_h [\mathcal{D}_{\sigma h} u_h : \mathcal{D}_{\sigma h} v_h - \sum_{e \in \partial \mathcal{E}_h} \int_e \{ \mathcal{D}_{\sigma h} u_h J_h A \cdot n_E \} \cdot [v_h]
\]

\[
\pm \sum_{e \in \partial \mathcal{E}_h} \int_e \{ \mathcal{D}_{\sigma h} v_h J_h A \cdot n_E \} \cdot [u_h] + \frac{\sigma}{h} \sum_{e \in \partial \mathcal{E}_h} [u_h][v_h].
\]

For NIPG, when \( v_h = u_h \), (2.88) reduces to

\[
\sum_{E \in \mathcal{E}_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) = \frac{1}{2} \int_\Omega J_h [\mathcal{D}_{\sigma h} u_h]^2 + \frac{\sigma}{h} \sum_{e \in \partial \mathcal{E}_h} [u_h]^2.
\]

Hence, we may simply take the penalty \( \sigma = 1 \) and by the discrete Korn’s inequality in Lemma 2.5, our result naturally follows.

For SIPG, when \( v_h = u_h \), (2.88) reduces to

\[
\sum_{E \in \mathcal{E}_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) = \frac{1}{2} \int_\Omega J_h [\mathcal{D}_{\sigma h} u_h]^2 - 2 \sum_{e \in \partial \mathcal{E}_h} \int_e \{ \mathcal{D}_{\sigma h} u_h J_h A \cdot n_E \} \cdot [u_h] + \frac{\sigma}{h} \sum_{e \in \partial \mathcal{E}_h} [u_h]^2.
\]

By the discrete Korn’s inequality in Lemma 2.5, we have

\[
\sum_{E \in \mathcal{E}_h} \int_E J_h [\mathcal{D}_{\sigma h} u_h]^2 + \sum_{e \in \partial \mathcal{E}_h} \frac{\sigma_0}{h} \int_e [u_h]^2 \geq C(Q) \| u_h \|_{X_h}^2 \geq C(Q) \| u_h \|_{H_h}^2.
\]

Then we utilize Hölder’s inequality, the trace theorem and Cauchy’s inequality to estimate

\[
\sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \int_e \{ \mathcal{D}_{\sigma h} u_h J_h A \cdot n_E \} \cdot [u_h] \leq C(Q) \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \| \nabla u_h \|_{L^2(e)} \| [u_h] \|_{L^2(e)}
\]

\[
\leq C(Q) \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \| u_h \|_{H^1(e)} \| [u_h] \|_{L^2(e)}
\]

\[
\leq CC(Q) \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} h \| u_h \|_{H^1(e)}^2 + \frac{C(Q)}{4C} \frac{1}{h} \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \| [u_h] \|_{L^2(e)}^2
\]

\[
\leq CC(Q) \sum_{e \in \partial \mathcal{E}_h} h \| u_h \|_{H^1/2(E)}^2 + \frac{C(Q)}{4C} \frac{1}{h} \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \| [u_h] \|_{L^2(e)}^2
\]

\[
\leq CC(Q) \| u_h \|^2_{H_h} + \frac{C(Q)}{4C} \frac{1}{h} \sum_{e \in \partial \mathcal{E}_h \setminus \partial \Omega} \| [u_h] \|_{L^2(e)}^2.
\]

The last inequality is valid since \( u_h \in P^1_h \). Then we take \( C \) sufficiently small, and \( \sigma \geq \sigma + C(Q)/(4C) \) to absorb (2.92) into (2.91). Hence, our result easily follows. \( \square \)
2.6. Pressure Discretization. For the pressure term
\[
\int_{E} J \nabla_{\mathcal{A}} p \cdot v_{h},
\]
a direct integration by parts reveals the following equality in the continuous case
\[
\sum_{E \in \mathcal{E}_{h}} \int_{E} J \nabla_{\mathcal{A}} p \cdot v_{h} = - \int_{\Omega} J p \nabla_{\mathcal{A}} \cdot v_{h} + \sum_{e \in \partial \mathcal{E}_{h} \setminus \Sigma} \int_{e} [v_{h}] \cdot (p J \mathcal{A}) \cdot n_{e} + \int_{\Sigma} p N \cdot v_{h}^{\text{int}(E)}. \tag{2.94}
\]
Inspired by [7], we define the discretization of the pressure term as \( \beta_{E}(p_{h}, \eta_{h}, v_{h}) \) which satisfies,
\[
\beta_{E}(p_{h}, \eta_{h}, v_{h}) = - \int_{E} J_{h} p_{h} \nabla_{\mathcal{A}h} \cdot v_{h} + \int_{\partial \mathcal{E}_{h} \setminus \Sigma} [v_{h}] \cdot (p J \mathcal{A}h) \cdot n_{e}. \tag{2.95}
\]
It is easy to see in the continuous case, this discretization reduces to
\[
\sum_{E \in \mathcal{E}_{h}} \beta_{E}(p, \eta, v_{h}) = - \int_{\Omega} J p \nabla_{\mathcal{A}} \cdot v_{h} + \sum_{e \in \partial \mathcal{E}_{h} \setminus \Sigma} [v_{h}] \cdot (p J \mathcal{A}) \cdot n_{e}. \tag{2.96}
\]
We can notice the difference is the extra physical boundary term \( \int_{\Sigma} p N \cdot v^{\text{int}(E)} \), which contributes to the boundary condition on \( \Sigma \). This is separately added to the scheme as a forcing term later combined with the diffusion contribution. Hence, our scheme is consistent.

Lemma 2.8. When we take \( v_{h} = u_{h} \), the discretized pressure term satisfies
\[
\sum_{E \in \mathcal{E}_{h}} \beta_{E}(p_{h}, \eta_{h}, u_{h}) = - \int_{\Omega} J_{h} p_{h} \nabla_{\mathcal{A}h} \cdot u_{h} + \sum_{e \in \partial \mathcal{E}_{h} \setminus \Sigma} [u_{h}] \cdot (p J \mathcal{A}h) \cdot n_{e}, \tag{2.97}
\]
where \( n_{e} \) denotes the outward normal vector read from the same direction as \( [u_{h}] \).

Proof. This is a direct corollary of the definition, so we omit the proof here. \( \square \)

2.7. Forcing Discretization. Considering the physical boundary condition on \( \Sigma \) of the system (1.8), we need to take the upper boundary condition as a forcing term, i.e.
\[
(pI - \mathcal{D}_{\mathcal{A}} u)N = \eta \mathcal{N} \quad \text{on} \quad \Sigma. \tag{2.98}
\]
Multiplying the test function \( v_{h} \) and integrating over \( F \) yield
\[
\int_{F} \eta \mathcal{N} \cdot v_{h} = \int_{F} \eta \cdot (\bar{v}_{2})_{h} - \int_{F} \eta \partial_{1} \eta \cdot (\bar{v}_{1})_{h}. \tag{2.99}
\]
We define the discretization of the forcing term as \( \mu_{E}(\eta_{h}, \eta_{h}, v_{h}) \) which satisfies,
\[
\mu_{E}(\eta_{h}, \eta_{h}, v_{h}) = \int_{\partial \mathcal{E} \cap \Sigma} \eta_{h} \cdot (\bar{v}_{2})_{h} - \int_{\partial \mathcal{E} \cap \Sigma} \eta_{h} \partial_{1} \eta \cdot (\bar{v}_{1})_{h} + \frac{1}{4} [\eta_{h}^{2}] \cdot \bar{v}_{h}^{-}\cdot \bar{v}_{h}^{+} |_{\partial \mathcal{E} \cap \Sigma+1/2} - \frac{1}{4} [\eta_{h}^{2}] \cdot \bar{v}_{h}^{+} |_{\partial \mathcal{E} \cap \Sigma-1/2}, \tag{2.100}
\]
where \( \partial \mathcal{E} \cap \Sigma + 1/2 \) and \( \partial \mathcal{E} \cap \Sigma - 1/2 \) denote the left and right boundary point of \( \partial \mathcal{E} \cap \Sigma \) respectively, i.e. similar to \( F - 1/2 \) and \( F + 1/2 \). Note that forcing term is nontrivial only for the top cells. For all the others, \( \mu_{E}(\eta_{h}, \eta_{h}, v_{h}) = 0 \). We add two penalty terms in order to show the stability. In the continuous case, above formulation reduces to
\[
\sum_{E \in \mathcal{E}_{h}} \mu_{E}(\eta, \eta, v_{h}) = \int_{\Sigma} \eta \cdot (\bar{v}_{2})_{h} - \int_{\Sigma} \eta \partial_{1} \eta \cdot (\bar{v}_{1})_{h}, \tag{2.101}
\]
and the penalty terms vanish. Hence, this discretization is consistent.

Lemma 2.9. When we take \( v_{h} = u_{h} \), the discretized forcing term satisfies
\[
\sum_{E \in \mathcal{E}_{h}} \mu_{E}(\eta_{h}, \eta_{h}, u_{h}) \geq \frac{1}{2} \int_{\Sigma} \partial_{1} |\eta_{h}|^{2}. \tag{2.102}
\]
Proof. We may sum up over \( E \in \mathcal{E}_h \) to obtain

\[
\sum_{E \in \mathcal{E}_h} \mu_E (\eta_h, \eta_h, v_h) = \int_\Sigma \eta_h \cdot (\bar{v}_2)_h + \int_\Sigma \eta_h \partial_1 \eta_h \cdot (\bar{v}_1)_h + \frac{1}{2} \sum_{F \in \mathcal{F}_h} \left( \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} - \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} \right). \tag{2.103}
\]

Taking \( \phi_h = \eta_h \) in the transport discretization (2.15) and integrating by parts imply

\[
\frac{1}{2} \int_F \partial_t |\eta_h| \, |\eta_h| - \int_F (\bar{u}_2)_h \cdot \eta_h + \int_F (\bar{u}_1)_h \cdot \eta_h \partial_1 \eta_h = -\{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} + (\bar{u}_1)_h \hat{\eta}_h \cdot \eta_h^-|_{F+1/2} + \frac{1}{2} \left( (\eta_h^-)^2 u_h|_{F+1/2} - (\eta_h^+)^2 u_h|_{F-1/2} \right) = 0,
\]

which further leads to

\[
\frac{1}{2} \int_F \partial_t |\eta_h| \, |\eta_h| - \int_F (\bar{u}_2)_h \cdot \eta_h + \int_F (\bar{u}_1)_h \cdot \eta_h \partial_1 \eta_h = -\{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} + (\bar{u}_1)_h \hat{\eta}_h \cdot \eta_h^-|_{F+1/2} - (\bar{u}_1)_h \hat{\eta}_h \cdot \eta_h^+|_{F-1/2} = 0.
\]

Summing over \( F \in \mathcal{F}_h \) implies

\[
\frac{1}{2} \int_\Sigma \partial_t |\eta_h| \, |\eta_h| - \int_\Sigma (\bar{u}_2)_h \cdot \eta_h + \int_\Sigma (\bar{u}_1)_h \cdot \eta_h \partial_1 \eta_h = \sum_{F \in \mathcal{F}_h} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} - \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F+1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^+|_{F-1/2} = 0.
\]

Since we always take \( \bar{u}_1 = \{(\bar{u}_1)_h\} \), combining (2.103) and (2.106) yields

\[
\sum_{E \in \mathcal{E}_h} \mu_E (\eta_h, \eta_h, v_h) = \frac{1}{2} \int_\Sigma \partial_t |\eta_h| \, |\eta_h| + \sum_{F \in \mathcal{F}_h} \left( -\frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F+1/2} - \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^+|_{F-1/2} \right).
\]

We can decompose the boundary terms in (2.107) to obtain

\[
-\frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} = \Psi_{F+1/2} - \Psi_{F-1/2} + \Theta_F,
\]

where

\[
\Psi_{F+1/2} = -\frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F+1/2},
\]

\[
\Psi_{F-1/2} = -\frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F-1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F-1/2},
\]

and

\[
\Theta_F = \frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} - \frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F-1/2} = \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^+|_{F-1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F-1/2}.
\]

By the flux definition (2.17), we can derive

\[
\Theta_F = \{(\bar{u}_1)_h\} (\eta_h^+ - \eta_h^-) \left( \frac{\eta_h^+ + \eta_h^-}{2} - \hat{\eta}_h \right) \geq 0.
\]

When we sum up (2.108) over \( F \in \mathcal{F}_h \), all \( \Psi_{F+1/2} \) are canceled out due to the periodicity. Then we have

\[
\sum_{F \in \mathcal{F}_h} \left( -\frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^-)^2|_{F+1/2} + \frac{1}{2} \{(\bar{u}_1)_h\} (\eta_h^+)^2|_{F-1/2} + \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^-|_{F+1/2} - \{(\bar{u}_1)_h\} \hat{\eta}_h \cdot \eta_h^+|_{F-1/2} \right) \geq 0.
\]
Hence, combining (2.107) and (2.113), we can obtain the desired result.

2.8. Source Discretization. We define the discretization of the source term as \( \omega_E(\eta_h, v_h) \) which satisfies,

\[
\omega_E(\eta_h, v_h) = \int_E J_h S \cdot v_h.
\]

Naturally, this discretization is consistent for given \( S \).

2.9. Divergence Discretization. Inspired by [7], we define the discretization of the divergence term as \( \rho_E(u_h, \eta_h, q_h) \) which satisfies

\[
\rho_E(u_h, \eta_h, q_h) = - \int_E J_h q_h \nabla \cdot u_h + \frac{1}{2} \int_{\partial E\setminus \Sigma} [\eta_h] \cdot \eta_h^{\text{int}(E)} \chi_h^{\text{int}(E)} \cdot n_E.
\]

In the continuous case, above discretization reduces to

\[
\sum_{E \in \mathcal{E}_h} \rho_E(u, \eta, q_h) = - \int_{\Omega} (\nabla \cdot u) q_h.
\]

Hence, this discretization is consistent.

Lemma 2.10. When we take \( q_h = p_h \), the discretized divergence term satisfies

\[
\sum_{E \in \mathcal{E}_h} \rho_E(u_h, \eta_h, p_h) = - \int_{\Omega} (\nabla \cdot u) q_h.
\]

Proof. This is a natural corollary of the definition, so we omit the proof here.

2.10. Summary. We can summarize all above to define the complete semi-discrete scheme as follows:

\[
\begin{align*}
\kappa_F(\eta_h, u_h, \phi_h) &= 0, \\
\lambda_F(\chi_h, \psi_h) &= 0, \\
\zeta_E(\eta_h, u_h, v_h) + \gamma_E(u_h, \eta_h, u_h, v_h) + \nu \omega_E(\eta_h, u_h, \eta_h, v_h) + \beta_E(p_h, \eta_h, \eta_h) + g \mu E(\eta_h, \eta_h, v_h) &= \omega_E(\eta_h, v_h), \\
\rho_E(u_h, \eta_h, q_h) &= 0,
\end{align*}
\]

for any test functions \( \phi_h \in \mathcal{S}_h^k, \psi_h \in \mathcal{S}_h^k, v_h \in X^k_h \) and \( q_h \in M_h^{k-1} \). This forms a complete differential-algebraic system for \( \eta_h(t) \in \mathcal{S}_h^k, \chi_h(t) \in \mathcal{S}_h^k, u_h(t) \in X^k_h \) and \( p_h(t) \in M_h^{k-1} \).

3. Stability Analysis

Condition 3.1. The free surface \( \eta_h(t) \) satisfies \( 1 + \eta_h(t) \geq \delta > 0 \) for some \( \delta > 0 \) independent of \( h \) and \( t \).

Condition 3.2. The free surface \( \eta_h(t) \) satisfies \( \sup_{F \in \mathcal{F}_h} \| \eta_h(t) \|_{W^{1,\infty}(F)} \leq Q \) for some \( Q > 0 \) independent of \( h \) and \( t \).

Theorem 3.3. If \( S = 0 \), suppose Condition 3.1 is valid in \( t \in [0, T] \) for some \( T > 0 \). Then there exists a unique numerical solution triple \( (u_h, p_h, \eta_h) \in X^k_h \times M_h^{k-1} \times \mathcal{S}_h^k \) to the scheme (2.118) with NIPG, which satisfies the estimate

\[
\begin{align*}
\left\| \sqrt{J_h(t)} u_h(t) \right\|_{L^2(\Omega)}^2 + g \| \eta_h(t) \|_{L^2(\Sigma)}^2 + \nu \int_0^t \int_{\Omega} J_h(s) |D \chi_{h,s}(s) u_h(s)|^2 \, ds & \leq \left\| \sqrt{J_h(0)} u_h(0) \right\|_{L^2(\Omega)}^2 + g \| \eta_h(0) \|_{L^2(\Sigma)}^2,
\end{align*}
\]

for any \( t \in [0, T] \).

For general \( S \), suppose Condition 3.1 and Condition 3.2 are valid in \( t \in [0, T] \) for some \( T > 0 \). Then there exists a unique numerical solution triple \( (u_h, p_h, \eta_h) \in X^k_h \times M_h^{k-1} \times \mathcal{S}_h^k \) to the scheme (2.118) with
SIPG or NIPG, which satisfies the estimate

\[
\left\| \sqrt{J_h(t)} u_h(t) \right\|_{L^2(\Omega)}^2 + \left\| \eta_h(t) \right\|_{L^2(\Sigma)}^2 + \int_0^t \left\| u_h(s) \right\|_{X_h}^2 \, ds \lesssim C(Q) \left( \left\| \sqrt{J_h(0)} u_h(0) \right\|_{L^2(\Omega)}^2 + \left\| \eta_h(0) \right\|_{L^2(\Sigma)}^2 + \int_0^t \int_\Omega \left| S(r) \right|^2 \, dr \right),
\]

for any \( t \in [0, T] \).

**Proof.** The existence and uniqueness follow from a standard argument for the differential-algebraic equations, so we omit it here and focus on the energy estimate. In the system (2.118), we take the test function \( v_h = u_h \) and \( q_h = p_h \), and sum up over \( E \in E_h \). Then it yields

\[
\sum_{E \in E_h} \zeta_E(\eta_h, u_h, u_h) + \sum_{E \in E_h} \gamma_E(u_h, \eta_h, u_h) + \nu \sum_{E \in E_h} \alpha_E(\eta_h, u_h, \eta_h, u_h) + \sum_{E \in E_h} \beta_E(p_h, \eta_h, u_h) + g \sum_{E \in E_h} \mu_E(\eta_h, \eta_h, u_h) = \sum_{E \in E_h} \omega_E(\eta_h, u_h),
\]

(3.4)  

Then by Lemma 2.2, Lemma 2.3, Lemma 2.7 and Lemma 2.9, we have

\[
\sum_{E \in E_h} \beta_E(p_h, \eta_h, u_h) = \sum_{E \in E_h} \rho_E(u_h, \eta_h, p_h) = 0.
\]

Hence, we can simplify (3.3) into

\[
\frac{1}{2} \frac{\partial t}{\partial t} \int_\Omega J_h |u_h|^2 + \frac{g}{2} \frac{\partial t}{\partial t} \int_\Sigma |\eta_h|^2 + CC(Q) \left\| u_h \right\|_{X_h}^2 \leq \sum_{E \in E_h} \omega_E(u_h).
\]

(3.7)

Note Condition 3.1 and Condition 3.2 guarantee the existence of \( C(Q) > 0 \) which is independent of \( t \). An application of Cauchy’s inequality implies

\[
\sum_{E \in E_h} \omega_E(u_h) \leq C' \int_\Omega J_h |u_h|^2 + \frac{1}{4C'} \int_\Omega |S|^2 \leq \frac{Q^2C'}{4} \left\| u_h \right\|_{X_h}^2 + \frac{1}{4C'} \int_\Omega |S|^2.
\]

(3.8)

In (3.8), when taking \( C' \) sufficiently small, we can always absorb it into \( CC(Q) \left\| u_h \right\|_{X_h}^2 \) in (3.7). Hence, we have

\[
\frac{1}{2} \frac{\partial t}{\partial t} \int_\Omega J_h |u_h|^2 + \frac{g}{2} \frac{\partial t}{\partial t} \int_\Sigma |\eta_h|^2 + \frac{CC(Q)}{2} \left\| u_h \right\|_{X_h}^2 \leq C(Q) \int_\Omega |S|^2.
\]

Then integrating over \([0, t]\) leads to the desired result. The \( S = 0 \) case is easily derived from the general case without discussion of the source term. \( \square \)

**Remark 3.4.** Condition 3.1 and Condition 3.2 are not always satisfied a priori. In [14], [1] and [15], this type of assumptions were also introduced in the numerical analysis.

4. Discussion on the Error Analysis

For the continuous solution \((u, \eta)\), the analysis in [20] reveals in the Navier-Stokes equations of (1.8), we need \( H^1 \) norm of \( \eta(t) \) to bound \( L^2 \) norm of \( u(t) \). However, the result in [6] implies in the transport equation of (1.8), we need \( H^2 \) norm of \( u(t) \) to control \( H^1 \) norm of \( \eta(t) \). This type of inconsistent coupling cannot be improved even if we go to higher order derivatives. Hence, this implies the coupled system in (1.8) is not closed in the usual Sobolev norms, which means we cannot expect to obtain the error estimates for the whole system (1.8).
In the following, we mainly analyze the error in the Navier-Stokes equations provided we have the error estimates of the free surface.

**Condition 4.1.** The free surface \( \eta_h \) satisfies the error estimates

\[
\begin{align*}
\| \eta_h - \eta \|_{L^2} &\lesssim h^{k+1}, \\
\| \eta_h - \eta \|_{L^\infty} &\lesssim h^k, \\
\| \chi_h - \chi \|_{L^2} &\lesssim h^k, \\
\| \chi_h - \chi \|_{L^\infty} &\lesssim h^{k-1}, \\
\| \partial_t (\eta_h - \eta) \|_{L^2} &\lesssim h^k, \\
\| \partial_t (\eta_h - \eta) \|_{L^\infty} &\lesssim h^{k-1}.
\end{align*}
\]

4.1. Velocity Error Analysis. We decompose the velocity error and the pressure error as follows:

\[
\begin{align*}
u_h - u &= (u_h - P)u + (P u - u) = \epsilon_u + \delta_u, \\
p_h - p &= (p_h - P)p + (P p - p) = \epsilon_p + \delta_p,
\end{align*}
\]

where \( P \) is some projection such that \( Pu \in X^k_h \) and \( Pp \in M^{k-1}_h \) achieving the optimal accuracy, i.e.

\[
\| \delta_u \|_{L^2} \lesssim h^{k+1}, \\
\| \delta_p \|_{L^2} \lesssim h^k.
\]

Hence, the key part of the error estimates is \( \epsilon_u \) and \( \epsilon_p \).

**Lemma 4.2.** Suppose Condition 3.1, Condition 3.2 and Conditions 4.1 are valid. Assume the exact solution satisfies \( u \in C^2(\Omega), p \in C^1(\Omega) \) and \( \eta \in C^1(\Sigma) \). Then for \( h \) sufficiently small, the numerical solution to the scheme (2.118) satisfies the estimate

\[
\partial_t \int_{\Omega} J_h |\epsilon_u|^2 + \| \epsilon_u \|_{H^1_h}^2 \lesssim C(Q) \left( h^{k-1-r} \| \epsilon_u \|_{H^1_h} + h^{k-r} \| \epsilon_p \|_{L^2} + \| \epsilon_u \|_{L^2}^2 \right) \quad \text{for} \quad 1/2 < r < 1.
\]

**Proof.** The discretization of the Navier-Stokes equations is

\[
\begin{align*}
\zeta_E(\eta_h, u_h, v_h) + \gamma_E(u_h, \eta_h, u_h, v_h) + \nu \alpha_E(\eta_h, u_h, \eta_h, v_h) + \beta_E(p_h, \eta_h, v_h) + g \mu_E(\eta_h, \eta_h, v_h) &= \omega_E(\eta_h, v_h), \\
\rho_E(u_h, \eta_h, \eta_h) &= 0.
\end{align*}
\]

The consistency of the scheme (2.118) implies the exact solution \( (u, p, \eta) \) also satisfies this scheme, i.e.

\[
\begin{align*}
\zeta_E(\eta, u, v) + \gamma_E(u, \eta, u, v) + \nu \alpha_E(\eta, u, \eta, v) + \beta_E(p, \eta, v) + g \mu_E(\eta, \eta, v) &= \omega_E(\eta, v), \\
\rho_E(\eta, \eta, \eta) &= 0.
\end{align*}
\]

Therefore, taking the difference of above two sets of equations, we obtain the error equations

\[
\begin{align*}
(\zeta_E(\eta_h, u_h, v_h) - \zeta_E(\eta, u, v)) + (\gamma_E(u_h, \eta_h, u_h, v_h) - \gamma_E(u, \eta, u, v)) + \\
\nu(\alpha_E(\eta_h, u_h, \eta_h, v_h) - \alpha_E(\eta, u, \eta, v)) + \\
(\beta_E(p_h, \eta_h, v_h) - \beta_E(p, \eta, v)) + g(\mu_E(\eta_h, \eta_h, v_h) - \mu_E(\eta, \eta, v)) &= \omega_E(\eta_h, v_h) - \omega_E(\eta, v_h), \\
\rho_E(u_h, \eta_h, \eta_h) - \rho_E(u, \eta, \eta) &= 0.
\end{align*}
\]

Now we need to analyze each term in the error equations (4.16) and (4.17). We always decompose the difference of the bilinear forms into two parts: the energy part \( W_* \) and the remaining part \( R_*, \) such that

\[
\text{Difference} = W_* + R_*.
\]
where $W_\ast$ builds the main body of the error equations and is put in the left-hand side (LHS), and $R_\ast$ is moved to the right-hand side (RHS) and can be taken as the perturbed source term. The $\ast$ can be $\zeta$, $\gamma$ or $\alpha$, etc. For example, we can decompose the convection difference in this fashion, we can make the best use of the regularity of the exact solutions and the known error estimates in the free surface to avoid the estimates of $(4.24)$ as follows:

\[(4.19) \quad \gamma(u_h, \eta_h, u_h, v_h) = \gamma(u_h, \eta_h, u_h - u, v_h) + \gamma(u_h, \eta_h - \eta, u, v_h) + \gamma(u_h - u, \eta, u, v_h) + \gamma(u_h - u, \eta, u, v_h)
\]

\[(4.20) \quad W_{\gamma} = \gamma(u_h, \eta_h, u_h, v_h),
\]

\[(4.21) \quad R_{\gamma} = \gamma(u_h, \eta_h, \delta_u, \varepsilon_u) + \gamma(u_h - u, \eta, u, \varepsilon_u).
\]

$W_\ast$ can be bounded as in Theorem 3.3 and $R_\ast$ can be estimated based on the following rules:

**Rule 1:**
For the estimates of the product terms as

\[(4.22) \quad \int F_1 F_2 \ldots F_k,
\]

we apply Hölder’s inequality to obtain two terms in the $L^2$ norm and other terms in the $L^\infty$ norm. The priority of taking the $L^\infty$ norm is as follows:

1. $F_i$ only containing the exact solution $(u, p, \eta)$ has the highest priority.
2. $F_i$ containing the free surface error $\eta_h - \eta$, $\chi_h - \chi$ or $\partial_t (\eta_h - \eta)$, or the numerical solution $\eta_h$, $\chi_h$ or $\partial_t \eta_h$ has the second priority.
3. $F_i$ containing the projection error $u - P u$ or $p - \mathcal{P} p$ for some $\mathcal{P}$ has the third priority.
4. $F_i$ only containing the velocity error $u_h - P u$ or the pressure error $p_h - \mathcal{P} p$ for some projection $\mathcal{P}$, or the numerical solution $u_h$ or $p_h$, has the lowest priority.

In this fashion, we can make the best use of the regularity of the exact solutions and the known error estimates in the free surface to avoid the estimates of $(u_h, p_h)$ in undesired norms due to the embedding theorem.

**Rule 2:**
For the boundary term, we apply the trace theorem

\[(4.23) \quad \|F\|_{L^2(\partial E)} \lesssim \|F\|_{H^r(E)},
\]

for $r > 1/2$. This introduces more regularity in the estimates, so we should always try to avoid to use it directly and apply certain projection property to eliminate the boundary terms.

**Rule 3:**
Note the simple fact that for $r > 0$, we have

\[(4.24) \quad h^r \|f_h\|_{H^r(E)} \lesssim \|f_h\|_{L^2(E)},
\]

where $h$ is the mesh size and $f_h \in P^k_h$. This can bound the higher order Sobolev norm by the lower norm at the price of some order of $h$.

In the error equations (4.16) and (4.17), we always take the test functions $v_h = \varepsilon_u$ and $q_h = \varepsilon_p$. Since the estimates are standard, we omit the details here and just present the results. We can simplify the error equations (4.16) and (4.17) as

\[(4.25) \quad W_{\zeta} + W_{\gamma} + W_{\alpha} + W_{\beta} = -R_{\zeta} - R_{\gamma} - R_{\alpha} - R_{\beta} - R_{\mu} - R_{\nu},
\]

\[(4.26) \quad W_{\rho} = -R_{\rho}.
\]
Then we can utilize a similar argument as in the proof of Theorem 3.3 to bound the energy part as

\[ (4.27) \quad \frac{1}{2} \partial_t \int_{\Omega} J_h |\epsilon_u|^2 + C(Q) \|\epsilon_u\|_{H_h}^2 \lesssim R_\rho - R_\xi - R_\gamma - R_\alpha - R_\beta - R_\mu - R_\omega. \]

Utilizing the estimates about the remaining part, we can further obtain

\[ (4.28) \quad \frac{1}{2} \partial_t \int_{\Omega} J_h |\epsilon_u|^2 + C(Q) \|\epsilon_u\|_{H_h}^2 \lesssim h^{k-1-r} \|\epsilon_u\|_{H_h} + h^{k-r} \|\epsilon_p\|_{L^2} + h^{k-1}(1 + \|\epsilon_u\|_{L^2}) \|\epsilon_u\|_{H_h} + \|\epsilon_u\|_{L^2} \|\epsilon_u\|_{H_h} \quad \text{for } 1/2 < r < 1. \]

The stability shows both \( u_h \) and \( u \) are bounded in \( L^2 \), so \( \epsilon_u \) is also bounded in \( L^2 \). Then we get

\[ (4.29) \quad \frac{1}{2} \partial_t \int_{\Omega} J_h |\epsilon_u|^2 + C(Q) \|\epsilon_u\|_{H_h}^2 \lesssim h^{k-1-r} \|\epsilon_u\|_{H_h} + h^{k-r} \|\epsilon_p\|_{L^2} + \frac{1}{4C''} \|\epsilon_u\|_{L^2} + C'' \|\epsilon_u\|_{H_h} \quad \text{for } 1/2 < r < 1. \]

Applying Cauchy’s inequality to the last term of (4.29) implies

\[ (4.30) \quad \frac{1}{2} \partial_t \int_{\Omega} J_h |\epsilon_u|^2 + C(Q) \|\epsilon_u\|_{H_h}^2 \lesssim h^{k-1-r} \|\epsilon_u\|_{H_h} + h^{k-r} \|\epsilon_p\|_{L^2} + \frac{1}{4C''} \|\epsilon_u\|_{L^2} + C'' \|\epsilon_u\|_{H_h} \quad \text{for } 1/2 < r < 1. \]

Taking \( C'' \) sufficiently small, we can absorb it into the left-hand side to achieve

\[ (4.31) \quad \frac{1}{2} \partial_t \int_{\Omega} J_h |\epsilon_u|^2 + C(Q) \|\epsilon_u\|_{H_h}^2 \lesssim h^{k-1-r} \|\epsilon_u\|_{H_h} + h^{k-r} \|\epsilon_p\|_{L^2} + C'' \|\epsilon_u\|_{H_h} \quad \text{for } 1/2 < r < 1. \]

This is the desired result. 

\[ \square \]

4.2. Pressure Error Analysis. In order for further analysis, we first need to show the inf-sup condition of our pressure discretization.

**Lemma 4.3.** In the pressure discretization \( \beta(p_h, \eta_h, v_h) = \sum_{E \in \mathcal{E}_h} \beta_E(p_h, \eta_h, v_h) \), suppose Condition 3.1, Condition 3.2 and Conditions 4.1 are valid. Assume the exact solution satisfies \( \eta(t) \in C^1(\Sigma) \). Then for sufficiently small time \( T \) which is independent of \( h \), there exists a constant \( \Xi > 0 \) such that

\[ (4.32) \quad \inf_{p_h \in M_h} \sup_{v_h \in X_h} \frac{\beta(p_h, \eta_h, v_h)}{\|p_h\|_{L^2} \|\eta_h\|_{H_h}} \geq \Xi, \]

for \( t \in [0, T] \), where \( \Xi \) is independent of \( h \).

**Proof.** This result is similar to [8, Theorem 4.5], so we omit the details here. Note the fact that for \( J_0\omega_0 \) which is the value of \( J\omega \) for \( t = 0 \), and \( \tilde{J}_0\omega_0 \) the average of \( J_0\omega_0 \) in each cell \( E \) which is piecewise constants, we naturally have

\[ (4.33) \quad \|J_0\omega_0 - \tilde{J}_0\omega_0\|_{L^\infty} \lesssim h. \]

Hence, the weight functions in the differential operators is only a small perturbation, which can be absorbed into the main part. 

\[ \square \]

**Lemma 4.4.** Suppose Condition 3.1, Condition 3.2 and Conditions 4.1 are valid. Assume the exact solution satisfies \( u \in C^2(\Omega) \), \( p \in C^1(\Omega) \) and \( \eta \in C^1(\Sigma) \). Then for sufficiently small \( T \) and \( h \), the numerical solution to the scheme (2.118) satisfies the estimate

\[ (4.34) \quad \int_0^t \|\epsilon_p\|_{L^2} \lesssim C(Q) \left( h^{k-1-r} + \int_0^t \|\epsilon_u\|_{L^2} + \|\epsilon_u(t)\|_{L^2} + \frac{1}{h} \int_0^t \|\epsilon_u\|_{H_h} \right) \quad \text{for } 1/2 < r < 1, \]

within \( t \in [0, T] \).
Proof. We can take any test function \( v_h \in X_h^k \) in the error equation (4.16) and integrate over time \([0, t]\). Here we do not distinguish between the energy part and remaining part, but only concentrate on

\[
(4.35) \quad \int_0^t \beta(p_h, \eta_h, v_h).
\]

All the other terms can be moved into the right-hand side and estimated in terms of \( \|v_h\|_{H_h} \). Basically, the estimates are similar to the velocity error estimates in the proof of Lemma 4.2, so we omit the details here.

Based on the inf-sup condition in Lemma 4.3, we can deduce the desired result. \( \square \)

4.3. Error Analysis of the Fluid.

Theorem 4.5. Suppose Conditions 3.1, 3.2 and Conditions 4.1 are valid. Assume the exact solution satisfies \( u \in C^2(\Omega) \), \( p \in C^1(\Omega) \) and \( \eta \in C^1(\Sigma) \). Then for sufficiently small \( T \) and \( h \), the numerical solution to the scheme (2.118) satisfies the estimates

\[
(4.36) \quad \|u_h - u\|_{L^2} \lesssim C(Q) h^{k-1-r},
\]

\[
(4.37) \quad \left( \int_0^t \|u_h - u\|_{H_h}^2 \right)^{1/2} \lesssim C(Q) h^{k-1-r},
\]

\[
(4.38) \quad \int_0^t \|p_h - p\|_{L^2} \lesssim C(Q) h^{k-1-r} \text{ for } 1/2 < r < 1,
\]

in \( t \in [0, T] \).

Proof. In Lemma 4.2 and Lemma 4.4, we have shown

\[
(4.39) \quad \partial_t \int_\Omega |\epsilon_u|^2 + \|\epsilon_u\|_{H_h}^2 \lesssim C(Q) \left( h^{k-1-r} \|\epsilon_u\|_{H_h} + h^{k-r} \|\epsilon_p\|_{L^2} + \|\epsilon_u\|_{L^2} \right) \text{ for } 1/2 < r < 1,
\]

and

\[
(4.40) \quad \int_0^t \|\epsilon_p\|_{L^2} \lesssim C(Q) \left( h^{k-1-r} + \int_0^t \|\epsilon_u\|_{L^2} \right) \text{ for } 1/2 < r < 1.
\]

Integrating over \([0, t]\) for any \( t \in [0, T]\) in (4.39), we obtain

\[
(4.41) \quad \int_0^t J_h(t) |\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q) \left( h^{k+2} + h^{k-1-r} \int_0^t \|\epsilon_u\|_{H_h} + h^{k-r} \int_0^t \|\epsilon_p\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2} \right).
\]

Then we plug (4.40) into (4.41) to eliminate \( \epsilon_p \) and obtain

\[
(4.42) \quad \int_0^t J_h(t) |\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q) \left( h^{2k-2-2r} + h^{k-1-r} \int_0^t \|\epsilon_u\|_{H_h} + h^{k-r} \int_0^t \|\epsilon_u\|_{L^2} \right).
\]

When \( T \) is sufficiently small, we can absorb \( C(Q) \sqrt{t} \int_0^t \|\epsilon_u\|_{H_h}^2 \) into the left-hand side of (4.42) to achieve

\[
(4.43) \quad \int_0^t J_h(t) |\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q) \left( h^{2k-2-2r} + h^{k-r} \|\epsilon_u(t)\|_{L^2} \right).
\]

Since (4.43) holds for any \( t \in [0, T] \), we can take the maximum to get

\[
(4.44) \quad \max_{t \in [0, T]} \|\epsilon_u(t)\|_{L^2} \lesssim C(Q) \left( h^{2k-2-2r} + h^{k-r} \max_{t \in [0, T]} \|\epsilon_u(t)\|_{L^2} + T \max_{t \in [0, T]} \|\epsilon_u(t)\|_{L^2} \right).
\]
When $T$ and $h$ are sufficiently small, we can absorb $C(Q)(h^{k-r} + T)$ max$_{t \in [0,T]} \|\epsilon_u\|_{L^2}$ into the left-hand side of (4.44) to obtain

$$\max_{t \in [0,T]} \|\epsilon_u(t)\|_{L^2}^2 \lesssim C(Q)h^{2k-2r-2r}.$$  

Then this leads to

$$\max_{t \in [0,T]} \|\epsilon_u(t)\|_{L^2}^2 \lesssim C(Q)h^{k-1-r}.$$  

Hence, we plug (4.46) into (4.43) and (4.40) to achieve

$$\int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)h^{2k-2r},$$  

$$\int_0^t \|\epsilon_p\|_{L^2} \lesssim C(Q)h^{k-1-r}.$$  

Combining with the projection errors $\delta_u$ and $\delta_p$, we show the desired result. \hfill \Box 

5. Numerical Tests

We perform the accuracy tests for our numerical scheme. Our tests are based on a set of exact solutions

$$(5.1) \quad \eta = Z \sin(2\pi(x_1 - t)),$$

$$(5.2) \quad u_1 = Z(6x_2^2 + 15x_2^4 + 12x_2^3 + 3x_2^2) \sin(2\pi(x_1 - t)) + (6x_2^5 + 5x_2^4 + 1),$$

$$(5.3) \quad u_2 = \pi Z(4x_2^6 + 15x_2^4 + 21x_2^4 + 13x_2^3 + 3x_2^2) \sin(4\pi(x_1 - t)) + 2\pi Z(4x_2^6 + 7x_2^5 + 2x_2^4 - x_2^3) \cos(2\pi(x_1 - t)),$$

$$(5.4) \quad p = Z \sin(2\pi(x_1 - t)),$$

with the source term

$$(5.5) \quad S_1 = (u_1)_t - \eta_t(1 + x_2)K \partial_2 u_1 + \left(u_1(\partial_1 u_1 - AK \partial_2 u_2) + Ku_2 \partial_2 u_1\right)$$

$$- \nu \left(\partial_{11} u_1 + (1 + A^2)K^2 \partial_{22} u_1 - 2AK \partial_{12} u_1 + (AK^2 \partial_2 A - A \partial_1 K - K \partial_1 A) \partial_2 u_1\right)$$

$$+ (\partial_1 p - AK \partial_2 p),$$

$$(5.6) \quad S_2 = (u_2)_t - \eta_t(1 + x_2)K \partial_2 u_2 + \left(u_1(\partial_1 u_2 - AK \partial_2 u_2) + Ku_2 \partial_2 u_2\right)$$

$$- \nu \left(\partial_{11} u_2 + (1 + A^2)K^2 \partial_{22} u_2 - 2AK \partial_{12} u_2 + (AK^2 \partial_2 A - A \partial_1 K - K \partial_1 A) \partial_2 u_2\right)$$

$$+ K \partial_2 p,$$

where $0 < Z < 1$ can be any fixed constant, $A$ and $K$ are defined as in (1.7), and differential operators $\partial_i$, $\partial_j$ and $\partial_{ij}$ are defined in the usual sense. In the following test, we always take $Z = 0.1$, $\nu = 0.05$ and $T = 1/8$.

5.1. Accuracy Tests with SIPG. The following are the error tables for our numerical scheme with SIPG:

**Table 1. $L^2$ Error Table for $X_h - M_h^0 - S_h^1$ formulation with SIPG**

<table>
<thead>
<tr>
<th>$N$</th>
<th>Error</th>
<th>Order</th>
<th>$N$</th>
<th>Error</th>
<th>Order</th>
<th>$N$</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9.4028E-4</td>
<td></td>
<td>4</td>
<td>4.8233E-2</td>
<td></td>
<td>4</td>
<td>3.1028E-3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.5493E-4</td>
<td>1.8830</td>
<td>8</td>
<td>1.3863E-2</td>
<td>1.7988</td>
<td>8</td>
<td>1.6255E-3</td>
<td>0.9327</td>
</tr>
<tr>
<td>16</td>
<td>6.6612E-5</td>
<td>1.9362</td>
<td>16</td>
<td>3.6182E-2</td>
<td>1.9378</td>
<td>16</td>
<td>8.2709E-4</td>
<td>0.9748</td>
</tr>
<tr>
<td>32</td>
<td>1.7409E-5</td>
<td>1.9359</td>
<td>32</td>
<td>9.4871E-4</td>
<td>1.9312</td>
<td>32</td>
<td>4.1570E-4</td>
<td>0.9925</td>
</tr>
<tr>
<td>64</td>
<td>4.9397E-6</td>
<td>1.8173</td>
<td>64</td>
<td>2.4593E-4</td>
<td>1.9477</td>
<td>64</td>
<td>2.0818E-4</td>
<td>0.9977</td>
</tr>
</tbody>
</table>
Table 2. $L^2$ Error Table for $X_h^2 - M_h^1 - S_h^1$ formulation with SIPG

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.4335E-4</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1.6235E-5</td>
<td>3.1481</td>
</tr>
<tr>
<td>16</td>
<td>2.1495E-6</td>
<td>2.9213</td>
</tr>
<tr>
<td>32</td>
<td>2.961E-7</td>
<td>2.8428</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5.8590E-3</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>7.7408E-4</td>
<td>2.9201</td>
</tr>
<tr>
<td>16</td>
<td>9.8674E-5</td>
<td>2.9717</td>
</tr>
<tr>
<td>32</td>
<td>1.2586E-5</td>
<td>2.9709</td>
</tr>
</tbody>
</table>

5.2. Accuracy Tests with NIPG. The following are the error tables for our numerical scheme with NIPG:

Table 3. $L^2$ Error Table for $X_h^1 - M_h^0 - S_h^1$ formulation with NIPG

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9.3809E-4</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>2.5179E-4</td>
<td>1.8975</td>
</tr>
<tr>
<td>16</td>
<td>6.5861E-5</td>
<td>1.9347</td>
</tr>
<tr>
<td>32</td>
<td>1.7258E-5</td>
<td>1.9322</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4.5428E-2</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1.8218E-2</td>
<td>1.3182</td>
</tr>
<tr>
<td>16</td>
<td>6.7565E-3</td>
<td>1.4310</td>
</tr>
<tr>
<td>32</td>
<td>2.1758E-3</td>
<td>1.6347</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.2035E-3</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1.6481E-3</td>
<td>0.9588</td>
</tr>
<tr>
<td>16</td>
<td>8.4487E-4</td>
<td>0.9640</td>
</tr>
<tr>
<td>32</td>
<td>4.2453E-4</td>
<td>0.9929</td>
</tr>
<tr>
<td>64</td>
<td>2.1239E-4</td>
<td>0.9992</td>
</tr>
</tbody>
</table>

Table 4. $L^2$ Error Table for $X_h^2 - M_h^1 - S_h^2$ formulation with NIPG

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.4810E-4</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1.7895E-5</td>
<td>3.0489</td>
</tr>
<tr>
<td>16</td>
<td>3.3218E-5</td>
<td>2.4295</td>
</tr>
<tr>
<td>32</td>
<td>8.5392E-7</td>
<td>1.9598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Error</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9.2305E-4</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>2.8213E-4</td>
<td>1.7100</td>
</tr>
<tr>
<td>16</td>
<td>7.5701E-5</td>
<td>1.8980</td>
</tr>
<tr>
<td>32</td>
<td>2.0531E-5</td>
<td>1.8825</td>
</tr>
</tbody>
</table>

5.3. Discussion on the Numerical Tests. In above accuracy tests, we can obtain the optimal order of convergence when applying scheme (2.118) with SIPG, and the sub-optimal order with NIPG. However, in both cases, our numerical results are much better than the sub-optimal error estimates obtained in Theorem 4.5.

In order to show stability, we have introduced many penalty terms in the numerical scheme. However, more detailed numerical tests show that in our test cases, we can ignore some of them without destroying the stability and the convergence order. These ignorable penalty terms include

1. In the transport discretization (2.15), we can ignore

$$\frac{1}{2} \left( \eta_h^+ \phi_h^-[u_h]_{F+1/2} - \eta_h^- \phi_h^+[u_h]_{F-1/2} \right).$$

2. In the convection discretization (2.40), we can ignore

$$\frac{1}{2} \int_E J_h(\nabla u_h \cdot u_h)(u_h \cdot v_h).$$

3. In the diffusion discretization (2.82) when applying NIPG, we can ignore

$$\frac{\sigma}{h} \int_{\partial E} [u_h] \cdot v_h^{int(E)}.$$

4. In the forcing discretization (2.100), we can ignore

$$\frac{1}{4} [\eta_h^2] \cdot \bar{v}_h |_{\partial E \cap \Sigma+1/2} - \frac{1}{4} [\eta_h^2] \cdot \bar{v}_h |_{\partial E \cap \Sigma-1/2}.$$
6. Conclusions and Remarks

In this paper, we construct a stable numerical scheme to solve the system (1.8) with discontinuous Galerkin method, and discuss the error analysis in the fluid with certain assumptions.

Although we focus on the 2-D fluid throughout this paper, it is easy to see this scheme can be naturally extended to the 3-D case with periodic settings for two horizontal directions. The main restriction to our scheme is that we require the free surface to be a single-valued function of the horizontal variables, which is not always true in practice, especially when the topological structure of the free surface varies during the evolution. Our scheme is non-conservative. Hence, it might not give the qualitatively correct simulation if the exact solution possesses singularities. However, for smooth exact solutions, our scheme can give a quite good approximation.

References