Energy Stable and Structure-Preserving Algorithms for the Stochastic Galerkin System of 2D Shallow Water Equations

Yekaterina Epshteyn[1], Akil Narayan[1,2], and Yinqian Yu[1]

[1]Department of Mathematics, University of Utah
[2]Scientific Computing and Imaging Institute, University of Utah

December 20, 2024

Abstract

Shallow water equations (SWE) are fundamental nonlinear hyperbolic PDE-based models in fluid dynamics that are essential for studying a wide range of geophysical and engineering phenomena. Therefore, stable and accurate numerical methods for SWE are needed. Although some algorithms are well studied for deterministic SWE, more effort should be devoted to handling the SWE with uncertainty. In this paper, we incorporate uncertainty through a stochastic Galerkin (SG) framework, and building on an existing hyperbolicity-preserving SG formulation for 2D SWE, we construct the corresponding entropy flux pair, and develop structure-preserving, well-balanced, second-order energy conservative and energy stable finite volume schemes for the SG formulation of the two-dimensional shallow water system. We demonstrate the efficacy, applicability, and robustness of these structure-preserving algorithms through several challenging numerical experiments.

Keywords: stochastic shallow water equations, stochastic Galerkin, entropy flux pair, structure-preserving algorithms, energy conservative schemes, energy stable schemes

Contents

1 Introduction **2**
1.1 Contributions of this paper 3

2 Stochastic Galerkin formulation of the two-dimensional shallow water equations **4**
2.1 Polynomial chaos expansion 4
2.2 Hyperbolic-preserving stochastic Galerkin formulation for 2D shallow water equation . . 5

3 An entropy flux pair for the 2D SG SWE **6**
3.1 Entropy flux pairs for the deterministic SWE 6
3.2 An entropy flux pair for the 2D SG SWE 7

4 Well-balanced energy conservative and energy stable schemes for the 2D SG SWE **8**
4.1 Energy conservative schemes 9
4.2 An energy conservative scheme for the 2D SG SWE (EC) 11
4.3 A first-order energy stable scheme (ES1) 12
4.4 A second-order energy stable scheme (ES2) 14

5 Algorithmic details and pseudocode **16**
5.1 Desingularization 16
5.2 Hyperbolicity-preserving criterion 17
5.3 Adaptive time step size 17
5.4 Numerical imposition of boundary conditions 17
5.5 Measuring energy change through augmented energy 18

6 Numerical experiments **19**
6.1 Accuracy test for the energy conservative and the energy stable schemes 21
6.2 Gaussian-shape hump with stochastic bottom 22
6.3 Gaussian-shape hump with stochastic initial water surface 23
6.4 A submerged flat plateau 23
6.5 Perturbation to lake-at-rest 27
6.6 Gaussian-shape hump with two-dimensional random variable 33

7 Conclusion **35**

A Proof of theorem 2 **38**
A.1 Proof of lemma 1 38
A.2 Proof of lemma 2 38
A.3 Proof of lemma 3 39

B Proofs for section 4.2 **40**
B.1 Proof of lemma 5 40
B.2 Proof of lemma 6 41
B.3 Proof of lemma 7 42

1 Introduction

The Saint-Venant system of shallow water equations is widely used in mathematical and physical modeling of geophysical fluid flows, where the horizontal length scale is much greater than the vertical length scale, such as in rivers, lakes, and coastal regions [14]. The shallow water equations can capture the essential dynamics of wave propagation, currents, and other hydrodynamic behaviors under the assumption of a shallow fluid layer, making them suitable for modeling a variety of practical problems involving water flows. Despite the wide applicability of shallow water systems, shallow water models with random component are even more relevant for practical real-world situations since precise knowledge of an environment or operating conditions is frequently absent. Such random component/uncertainty can improve the predictive capabilities of the models, motivating the development of advanced methods to incorporate and manage the uncertainty in the governing shallow water systems. In this context, we will consider the *parameterized* stochastic shallow water equations (SWE), with the parameter as a random variable modeling the uncertainty.

In order to discretize the parameterized SWE, we consider the polynomial chaos expansion (PCE), which is a technique for modeling uncertainty in parameterized systems. Initially introduced by Wiener [55] for Gaussian processes using Hermite polynomials, the method has since been extended to more general orthogonal polynomial bases to handle random inputs with various distributions [26, 36, 54, 64]. The PCE approach effectively encodes uncertainty in parameterized partial differential equations (PDEs), resulting in a system of equations that can be solved using two major classes of methods: *intrusive* methods and *non-intrusive* methods. Based on sampling [42, 43, 63], non-intrusive methods construct the polynomial solution to the parameterized system by collecting an ensemble of solutions to the deterministic form at a collection of fixed values of the parameters corresponding to random variables. This approach can utilize existing and trusted legacy solvers for the deterministic SWE, for example, see, e.g. [67, 32, 47, 9, 59, 60, 61, 33, 4, 18, 31, 40, 58, 57, 66]. However, the solutions

generated by the non-intrusive approaches may be less accurate than those from intrusive methods. Moreover, ensuring associated structure-preserving properties, including the entropy conditions, can be challenging or impossible.

In the context of PCE methods, intrusive methods typically refer to the stochastic Galerkin (SG) approach, where the truncated PCE [55, 64] representations of the state variables are projected using a Galerkin method in the stochastic space. This process results in a new system of deterministic PDEs, with the variables being the truncated PCE coefficients. To solve the new system, a substantial rewrite of legacy codes and PDEs solvers is required, which is one of the drawbacks of intrusive methods. However, the SG method is generally expected to be more accurate than some alternative non-intrusive methods. Since the SG method is based on projection, it leads to near-optimal accuracy in the L^2 sense for static problems [1, 35]. Additionally, the SG approach provides opportunities to rigorously prove the desired properties of the SG system and the subsequent numerical schemes. Therefore, SG methods have been effectively employed to model uncertainty in diffusion equations [65, 17], kinetic equations [28, 49], and conservation and balance laws with symmetric Jacobian matrices [53].

However, for a general hyperbolic system of conservation and balance laws, such as the SWE system, the associated SG system may not be hyperbolic [16, 24, 29]. Thus, the SG formulation may result in a system of partial differential equations that belongs to a different class from the original deterministic system, potentially yielding unphysical solutions and compromising the robustness of subsequent numerical schemes. Numerous recent efforts have explored numerical methods for the SG formulation of various types of hyperbolic conservation laws. Some advances include SG-type analysis and algorithms for scalar conservation laws [68] including well-balanced methods [30], Haar wavelet-based approaches [25], hyperbolicity preservation through a non-equivalent Roe variables formulation [23], filtering strategies [34], limiter-type methods [48], hyperbolicity formulations for linear problems [46] or using linearization techniques [56], operator splitting [7, 8], non-conservative formulations of SG SWE systems [5], and entropic variable representations [44, 45].

In this paper, we build upon the recent work on a hyperbolicity-preserving SG formulation for two-dimensional SWE [12]. Despite the hyperbolicity-preserving property, if the corresponding source term vanishes, the SG SWE is a nonlinear hyperbolic conservation and balance law and hence it inherits standard challenges in developing numerical methods for such a model. For example, solutions can develop shock discontinuities in finite time with generic initial conditions, potentially producing non-unique weak solutions. Therefore, an additional entropy condition should be imposed, either implicitly or explicitly, to identify the desired physical solution. Note that implementing implicit time-integration solvers is challenging because of the nonlinearity involved in the system [10, 37, 39]. In addition, depending on its structure, the SWE system is supposed to satisfy the *well-balanced* property [3], ensuring that numerical schemes accurately capture the steady-state solution of the PDE. Moreover, uncertainty in the stochastic model complicates all challenges when considering the SG SWE, making them more ambiguous and difficult to address.

1.1 Contributions of this paper

In this paper, we extend the ideas developed in [13] to stochastic shallow water equations in two-dimensional physical space. The main contributions of this paper are as follows:

- We derive an entropy flux pair for the hyperbolicity-preserving, positivity-preserving stochastic Galerkin (SG) formulation of two-spatial-dimensional shallow water equations (SWE) from [12]. Entropy-entropy flux pairs are the theoretical starting point towards the entropy admissibility criteria to resolve non-uniqueness of weak solutions.

- Using the entropy-entropy flux pair, we devise second-order energy conservative, and first- and second-order energy stable finite volume schemes for the 2D SG SWE, all of which are also well-balanced. The designed energy conservative and energy stable schemes are stochastic extensions of the schemes developed in [21, 22] and are also the two-dimensional extensions of schemes developed in [13] for the stochastic SWE in 1D physical space.

- We present several challenging numerical experiments to evaluate the performance of our schemes, including accuracy, well-balanced property, energy decay, and numerical robustness.

An outline of this paper is as follows: In Section 2, we present preliminaries of polynomial chaos expansion (PCE) and the SG formulation of two-dimensional SWE system from [12]. In Section 3, we introduce entropy flux pairs for the deterministic SWE discussed in [21] and construct an entropy flux pair for the SG formulation of the two-dimensional SWE system. In Section 4, we develop energy conservative and energy stable schemes, accompanied by theoretical proofs and algorithmic details. In Section 6, we present several challenging numerical examples to demonstrate the performance of our schemes and verify their theoretical properties. In Section 7 we provide a brief summary of the main results of this paper and some potential topics for future research. Finally, we include some technical details for several proofs of lemmas and theorems in appendices A and B.

2 Stochastic Galerkin formulation of the two-dimensional shallow water equations

In this section, we review the stochastic Galerkin (SG) formulation of the two-dimensional (2D) shallow water equations (SWE), which possess a hyperbolicity-preserving property, as developed in [12]. We begin with the deterministic form of the 2D SWE:

$$U_t + F(U)_x + G(U)_y = S(U), \qquad U = (h, q^x, q^y)^\top, \tag{2.1}$$

where F and G denote the fluxes in the x- and y-directions, respectively, and S represents the source term,

$$F(U) = \begin{pmatrix} q^x \\ \frac{(q^x)^2}{h} + \frac{gh^2}{2} \\ \frac{q^x q^y}{h} \end{pmatrix}, \qquad G(U) = \begin{pmatrix} q^y \\ \frac{q^x q^y}{h} \\ \frac{(q^y)^2}{h} + \frac{gh^2}{2} \end{pmatrix}, \qquad S(U) = \begin{pmatrix} 0 \\ -gh\frac{\partial B}{\partial x} \\ -gh\frac{\partial B}{\partial y} \end{pmatrix}, \tag{2.2}$$

where $U = U(x, y, t)$ is the vector of conservative variables, $h = h(x, y, t)$ is the water height, $q^x(x, y, t)$ and $q^y(x, y, t)$ are the discharges in the x- and y-directions, respectively, and $B(x, y)$ is the time-independent bottom topography. For the *stochastic* shallow water equations, we consider introduction of a random field ξ, which could result from uncertainty or ignorance of the inputs, for example, bottom topography and initial data. We begin with some preliminaries for the model formulation of the stochastic Galerkin approach to the stochastic shallow water equations.

2.1 Polynomial chaos expansion

In this subsection, we provide a brief review of polynomial chaos expansion (PCE). For further details, see [15, 50, 62]. Let $\xi \in \mathbb{R}^d$ be a d-dimensional random variable with a Lebesgue density function ρ. Define the L^2-integrable function space associated with ρ as follows:

$$L^2_\rho(\mathbb{R}^d) \coloneqq \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \left(\int_{\mathbb{R}^d} f^2(s)\rho(s)ds \right)^{1/2} < +\infty \right\}. \tag{2.3}$$

Assuming that ρ has finite polynomial moments of all orders, there exists a d-variate orthonormal polynomial basis $\{\phi_k\}_{k=1}^\infty$ such that,

$$\mathbb{E}[\phi_k(\xi)\phi_l(\xi)] = \langle \phi_k, \phi_l \rangle_\rho \coloneqq \int_{\mathbb{R}^d} \phi_k(s)\phi_l(s)\rho(s)ds = \delta_{k,l}, \quad \forall k, l \in \mathbb{N}, \quad \phi_1(\xi) \equiv 1, \tag{2.4}$$

where $\delta_{k,l}$ is the Kronecker delta. Further, under mild conditions [19], then these basis functions span L^2_ρ: For any $z \in L^2_\rho$, then

$$z(x, y, t, \xi) \overset{L^2_\rho}{=} \sum_{k=1}^{\infty} \widehat{z}_k(x, y, t)\phi_k(\xi), \qquad \widehat{z}_k = \langle z, \phi_k \rangle_\rho \tag{2.5}$$

where x, y, and t are deterministic spatial and temporal variables, and $\widehat{z}_k(x,y,t)$ are the deterministic Fourier-type coefficients corresponding to the orthonormal basis $\{\phi_k\}_{k=1}^{\infty}$. Numerical computations require finite truncations of these expansions. Let $P = \text{span}\{\phi_k, k = 1,2,\ldots,K\}$ be a K-dimensional polynomial subspace of L^2_ρ. We then define the K-term PCE approximation of a random field z on this subspace as follows:

$$\Pi_P[z](x,y,t,\xi) \coloneqq \sum_{k=1}^{K} \widehat{z}_k(x,y,t)\phi_k(\xi). \tag{2.6}$$

The statistics of $\Pi_P[z]$ can be derived from its expansion coefficients. Specifically, the mean and variance of the random field z can be expressed in terms of these coefficients as follows:

$$\mathbb{E}[\Pi_P[z](x,y,t,\xi)] = \widehat{z}_1(x,y,t), \qquad \text{Var}[\Pi_P[z](x,y,t,\xi)] = \sum_{k=2}^{K} \widehat{z}_k^2(x,y,t). \tag{2.7}$$

Our numerical schemes involve specific manipulations of truncated expansion coefficients. With $\widehat{z} = (\widehat{z}_1,\ldots,\widehat{z}_k)^\top \in \mathbb{R}^K$ the size-K vector of truncated PCE coefficients of z, define the linear operator $\mathcal{P}:\mathbb{R}^K \to \mathbb{R}^{K\times K}$ as follows:

$$\mathcal{P}(\widehat{z}) \coloneqq \sum_{k=1}^{K} \widehat{z}_k \mathcal{M}_k, \qquad \mathcal{M}_k \in \mathbb{R}^{K\times K}, \qquad (\mathcal{M}_k)_{l,m} \coloneqq \langle \phi_k, \phi_l\phi_m\rangle_\rho. \tag{2.8}$$

Due to the symmetry and commutativity properties of the operator $\mathcal{P}(\cdot)$, the following identities hold,

$$\mathcal{P}(\widehat{z}) = (\mathcal{M}_1\widehat{z}, \mathcal{M}_2\widehat{z}, \ldots, \mathcal{M}_K\widehat{z}), \qquad \mathcal{P}(\widehat{a})\widehat{b} = \mathcal{P}(\widehat{b})\widehat{a}, \qquad \widehat{b}^\top\mathcal{P}(\widehat{a}) = \widehat{a}^\top\mathcal{P}(\widehat{b}), \tag{2.9}$$

The last two properties are proved in [13, Lemma 2.1], using the definition and symmetry of $\mathcal{P}$. A stochastic Galerkin (SG) formulation of a ξ-parameterized partial differential equation (PDE) assumes that the state variable lies in the polynomial space P and forms a scheme corresponding to projecting the PDE residual onto P. Note that the hyperbolicity of the straightforward SG formulation for nonlinear hyperbolic PDEs, such as the shallow water equations, is not automatically guaranteed. Therefore, special designs are required for the SG formulation of these nonlinear hyperbolic PDEs to preserve such an important property.

2.2 Hyperbolic-preserving stochastic Galerkin formulation for 2D shallow water equation

In this subsection, we review existing results on the hyperbolicity-preserving stochastic Galerkin formulation of the two-spatial-dimensional shallow water equations (2D SG SWE) [12]. We follow a standard Galerkin procedure in the stochastic space. It begins with reducing the problem to an alternative finite-dimensional form by replacing the solutions (h, q^x, q^y) by the ansatz,

$$\begin{aligned} h \simeq h_P &\coloneqq \sum_{k\in[K]} \widehat{h}_k(x,y,t)\phi_k(\xi), \\ q^x \simeq q^x_P &\coloneqq \sum_{k\in[K]} (\widehat{q^x})_k(x,y,t)\phi_k(\xi), \\ q^y \simeq q^y_P &\coloneqq \sum_{k\in[K]} (\widehat{q^y})_k(x,y,t)\phi_k(\xi), \end{aligned} \tag{2.10}$$

respectively, and the bottom B by $\Pi_P[B]$, where we use the notation $[K] \coloneqq \{1,2,\ldots,K\}$. With a special choice of how the Galerkin projection is applied to the nonlinear, non-polynomial terms $(q^x)^2/h, (q^y)^2/h, q^xq^y/h$ introduced in [12], a SG system of balance laws was derived, whose state variables are the coefficients in (2.10)

$$\widehat{U}_t + \widehat{F}(\widehat{U})_x + \widehat{G}(\widehat{U})_y = \widehat{S}(\widehat{U},\widehat{B}). \tag{2.11}$$

Here, $\widehat{U} = \left(\widehat{h}^\top, (\widehat{q^x})^\top, (\widehat{q^y})^\top\right)^\top \in \mathbb{R}^{3K}$, where $\widehat{h}, \widehat{q^x}, \widehat{q^y}$ are length-K vectors whose entries are the coefficients in (2.10). The flux terms are defined by

$$\widehat{F}(\widehat{U}) = \begin{pmatrix} \widehat{q^x} \\ \mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h})\widehat{q^x} + \frac{1}{2}g\mathcal{P}(\widehat{h})\widehat{h} \\ \mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h})\widehat{q^y} \end{pmatrix}, \quad \widehat{G}(\widehat{U}) = \begin{pmatrix} \widehat{q^y} \\ \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h})\widehat{q^x} \\ \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h})\widehat{q^y} + \frac{1}{2}g\mathcal{P}(\widehat{h})\widehat{h} \end{pmatrix}. \tag{2.12}$$

The source term is given by

$$\widehat{S}(\widehat{U}) = \begin{pmatrix} 0 \\ -g\mathcal{P}(\widehat{h})\widehat{B_x} \\ -g\mathcal{P}(\widehat{h})\widehat{B_y} \end{pmatrix}. \tag{2.13}$$

In the deterministic case, the fluxes (2.12) and the source term (2.13) reduce to their deterministic forms as shown in (2.2). A notable observation from the above is that the deterministic term $\frac{q^x q^y}{h}$ have two different stochastic representations in (2.12), i.e., $\mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h})\widehat{q^y} \neq \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h})\widehat{q^x}$. This different treatment of these terms is essential to retaining hyperbolicity; a more detailed discussion of this fact and a brief empirical investigation of the difference between these two stochastic representations is provided in [12].

Recall the deterministic SWE is hyperbolic under the condition that the water height $h > 0$. There is a natural extension of this property to the SG formulation of the SWE.

Theorem 1 (Theorem 3.1 in [12])**.** *If the matrix $\mathcal{P}(\widehat{h})$ is strictly positive definite at every point (x, y, t) in the computational spatial-temporal domain, then the SG formulation* (2.11) *is hyperbolic.*

This result is proven by identifying a symmetrizing similarity transform of the SG SWE flux Jacobians $\frac{\partial \widehat{F}}{\partial \widehat{U}}$ and $\frac{\partial \widehat{G}}{\partial \widehat{U}}$. These flux Jacobians will be useful for us later, so we explicitly provide them below: When $\mathcal{P}(\widehat{h})$ is invertible, the well-defined terms

$$\widehat{u} = \mathcal{P}^{-1}(\widehat{h})\widehat{q^x}, \qquad \widehat{v} = \mathcal{P}^{-1}(\widehat{h})\widehat{q^y}, \tag{2.14}$$

are stochastic representations of the x- and y-directional water velocity variables. I.e., these terms can be interpreted as the vectors of the PCE coefficients of the x-velocity $u \coloneqq q^x/h$ and the y-velocity $v \coloneqq q^y/h$. The flux Jacobians of the SG SWE system (2.11) can then expressed in terms of $K \times K$ blocks as follows:

$$\begin{aligned} \frac{\partial \widehat{F}}{\partial \widehat{U}} &= \begin{pmatrix} 0 & I & 0 \\ g\mathcal{P}(\widehat{h}) - \mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}) & \mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h}) + \mathcal{P}(\widehat{u}) & 0 \\ -\mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}) & \mathcal{P}(\widehat{v}) & \mathcal{P}(\widehat{q^x})\mathcal{P}^{-1}(\widehat{h}) \end{pmatrix}, \\ \frac{\partial \widehat{G}}{\partial \widehat{U}} &= \begin{pmatrix} 0 & 0 & I \\ -\mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}) & \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h}) & \mathcal{P}(\widehat{u}) \\ g\mathcal{P}(\widehat{h}) - \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}) & 0 & \mathcal{P}(\widehat{q^y})\mathcal{P}^{-1}(\widehat{h}) + \mathcal{P}(\widehat{v}) \end{pmatrix}. \end{aligned} \tag{2.15}$$

3 An entropy flux pair for the 2D SG SWE

To construct numerical schemes with desired energy stability properties, it is essential to derive entropy flux pairs for the SG formulation (2.11) of the SWE. We first review entropy flux pairs for the *deterministic* SWE discussed in [21].

3.1 Entropy flux pairs for the deterministic SWE

Solutions of systems of conservation and balance laws can develop shock discontinuities in finite time, even from smooth initial data, potentially leading to non-unique weak solutions defined in the distributional sense. To identify the desired physically relevant solution, an additional entropy admissibility criterion is imposed [2, 10]. For a general balance law in two spatial dimensions,

$$U_t + F(U)_x + G(U)_y = S(U), \tag{3.1}$$

an entropy flux tuple is a tuple $(E(U), H(U), K(U))$ satisfying the *companion balance law*,

$$E(U)_t + H(U)_x + K(U)_y = 0, \tag{3.2}$$

where the entropy $E(U)$ is a scalar function that is convex in U, and H and K represent the corresponding entropy flux functions. For the consistency with the original balance law, the entropy flux pair (E, H, K) is supposed to satisfy the *compatibility condition* as follows:

$$\frac{\partial E}{\partial U}(F_x + G_y - S) = H_x + K_y. \tag{3.3}$$

This condition ensures that multiplying the equation (3.1) by $\frac{\partial E}{\partial U}$ recovers the equation (3.2) for smooth solutions. When the source term vanishes and $(E, H, K) = (E(U), H(U), K(U))$, the compatibility condition (3.3) simplifies to the usual entropy condition for conservation laws. Even though an entropy flux pair may not exist for a general system of balance laws, the companion balance law (3.2) for a hyperbolic system of balance laws, derived from continuum physics, is usually related to the Second Law of thermodynamics, with the total energy of the system often serving as the entropy function. Several related examples can be found in Section 3.3 of [10]. For the deterministic SWE in (2.1), the total energy [21] is given by,

$$E^d(U) = \frac{1}{2}(q^x u + q^y v) + \frac{1}{2}gh^2 + ghB, \qquad u \coloneqq q^x/h, \qquad v \coloneqq q^y/h, \tag{3.4}$$

where $\frac{1}{2}(q^x u + q^y v)$ is the kinetic energy, and $\frac{1}{2}gh^2 + ghB$ is the potential energy. A direct computation of the Hessian confirms that $E^d(U)$ is convex in $U = (h, q^x, q^y)$. If we choose the fluxes H^d and K^d as,

$$H^d(U) \coloneqq \frac{1}{2}(hu^3 + huv^2) + gq^x(h + B), \qquad K^d(U) \coloneqq \frac{1}{2}(hu^2 v + hv^3) + gq^y(h + B), \tag{3.5}$$

then one can verify directly that $E^d_t + H^d_x + K^d_y = 0$, i.e., (E^d, H^d, K^d) satisfies the companion balance law (3.2). Hence, (E^d, H^d, K^d) is an entropy flux tuple for (2.1). In the case of weak solutions with shocks, the entropy admissibility criterion requires that energy dissipates according to a vanishing viscosity principle as follows:

$$E^d(U)_t + H^d(U)_x + K^d(U)_y \le 0. \tag{3.6}$$

By extending the results of deterministic SWE, we construct an entropy flux pair for the SG formulation of the 2D SWE system (2.11) in the next section. This involves identifying a tuple, consisting of an entropy function that is convex in the state variable, that satisfies the companion balance law.

3.2 An entropy flux pair for the 2D SG SWE

In this subsection, we focus on constructing an entropy flux pair for the SG system (2.11). We recall the discussion surrounding (2.11) for the definition of $\widehat{U}$. The main result of this section is stated in the following theorem.

Theorem 2. *Define the entropy function*

$$E(\widehat{U}) = \frac{1}{2}\Big((\widehat{q^x})^\top \widehat{u} + (\widehat{q^y})^\top \widehat{v}\Big) + \frac{1}{2}g\|\widehat{h}\|^2 + g\widehat{h}^\top \widehat{B}, \tag{3.7}$$

and the flux functions

$$\begin{aligned} H(\widehat{U}) =& \frac{1}{2}\Big(\widehat{u}^\top \mathcal{P}(\widehat{q^x})\widehat{u} + \widehat{v}^\top \mathcal{P}(\widehat{q^x})\widehat{v}\Big) + g(\widehat{q^x})^\top(\widehat{h} + \widehat{B}), \\ K(\widehat{U}) =& \frac{1}{2}\Big(\widehat{v}^\top \mathcal{P}(\widehat{q^y})\widehat{v} + \widehat{u}^\top \mathcal{P}(\widehat{q^y})\widehat{u}\Big) + g(\widehat{q^y})^\top(\widehat{h} + \widehat{B}), \end{aligned} \tag{3.8}$$

If $\mathcal{P}(\widehat{h}) > 0$, i.e., strictly positive definite, then (E, H, K) is an entropy flux pair for the SG system of the two-dimensional SWE (2.11).

Note that the stochastic variants of entropy function (3.7) and flux functions (3.8) reduce to the deterministic energy (3.4) and fluxes (3.5), respectively. The proof of theorem 2 is similar to that of [13, Theorem 3.1], which proves a corresponding result for an SG formulation of the SWE in a single spatial dimension. Hence, we relegate most details to appendix A, and provide only an outline of the major results needed. We start from a lemma that computes the gradient of $\widehat{u}$ and $\widehat{v}$. This technical result is also useful later when we construct energy conservative and stable schemes.

Lemma 1 (Gradients of $\widehat{u},\widehat{v}$)**.** *Let $\widehat{q^x},\widehat{q^y}$ be arbitrary, and let $\widehat{h}\in\mathbb{R}^K$ be such that $P(\widehat{h})$ is invertible. Define $\widehat{u},\widehat{v}$ by* (2.14)*, then the gradients of the velocities are*

$$\frac{\partial\widehat{u}}{\partial\widehat{U}} = [-\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}),\mathcal{P}^{-1}(\widehat{h}),0], \qquad \frac{\partial\widehat{v}}{\partial\widehat{U}} = [-\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}),0,\mathcal{P}^{-1}(\widehat{h})]. \tag{3.9}$$

For the proof, see appendix A. This result is a crucial ingredient in the following two lemmas.

Lemma 2 (Convexity of $E(\widehat{U})$)**.** *If $\mathcal{P}(\widehat{h})$ is positive definite, then the function $E(\widehat{U})$ defined* (3.7) *is convex in $\widehat{U}$.*

Lemma 3 (Companion balance law)**.** *If $\widehat{U}$ is smooth, the entropy flux pair (E,H,K) defined in* (3.7) *and* (3.8) *satisfies the two-dimensional companion balance law:*

$$E(\widehat{U})_t + H(\widehat{U})_x + K(\widehat{U})_y = 0. \tag{3.10}$$

The proofs of lemmas 2 and 3 are provided in appendix A. The proof of Theorem 2 follows from Lemmas 2 and 3. I.e., we construct an entropy flux pair (E,H,K), as defined in (3.7) and (3.8), for the SG SWE (2.11). This forms the foundation for developing energy conservative and energy stable semi-discrete finite volume schemes.

For our next goal of constructing energy conservative and energy stable schemes we define the following quantities,

$$\begin{aligned}
\widehat{V} &:= \Big(\frac{\partial E}{\partial\widehat{U}}\Big)^\top = \Big(-\frac{1}{2}\widehat{u}^\top\mathcal{P}(\widehat{u}) - \frac{1}{2}\widehat{v}^\top\mathcal{P}(\widehat{v}) + g(\widehat{h}+\widehat{B})^\top, \widehat{u}^\top, \widehat{v}^\top\Big)^\top,\\
\Psi &:= \widehat{V}^\top\widehat{F} - H = \frac{1}{2}g\widehat{u}^\top\mathcal{P}(\widehat{h})\widehat{h},\\
\Phi &:= \widehat{V}^\top\widehat{G} - K = \frac{1}{2}g\widehat{v}^\top\mathcal{P}(\widehat{h})\widehat{h},
\end{aligned} \tag{3.11}$$

where $\widehat{V}$ is called the entropy variable, and Ψ and Φ are called stochastic energy potentials.

4 Well-balanced energy conservative and energy stable schemes for the 2D SG SWE

In this section, we develop a well-balanced, second-order energy conservative (EC) scheme, as well as first-order and second-order energy stable (ES) schemes for the stochastic Galerkin (SG) formulation for the two-dimensional shallow water equations (SWE) (2.11). These schemes are constructed using the entropy-flux pairs provided in theorem 2 and are designed to provide energy conservation and decay properties. They are stochastic extensions of the methods in [21] for deterministic balance laws, and are two-dimensional extensions of methods in [13] for the SWE SG system in one spatial dimension. We start by defining the well-balanced property of a numerical scheme, which ensures that stochastic "lake-at-rest" steady states are equilibrium states at the discrete level.

Definition 1 (Well-Balanced SG SWE Property [12])**.** *The solution (h_P,q_P^x,q_P^y) of* (2.11) *is said to be well-balanced if it satisfies the stochastic "lake-at-rest" solution,*

$$q_P^x(x,y,t,\xi) = q_P^y(x,y,t,\xi) \equiv 0, \qquad h_P(x,y,t,\xi) + \Pi_P[B](x,y,t,\xi) \equiv C(\xi), \tag{4.1}$$

where $C(\xi)$ is a random scalar depending only on ξ. Such a well-balanced solution describes a still water surface with a flat but stochastic water surface. In terms of the system of PCE coefficients, (4.1) *is equivalent to the following vector equations*

$$\widehat{q^x} = \widehat{q^y} \equiv \mathbf{0}, \qquad \widehat{h} + \widehat{B} \equiv \widehat{C}, \qquad \forall (x, y, t) \in \mathcal{D} \times [0, T], \tag{4.2}$$

with spatial domain $\mathcal{D}$ and terminal time T. Note that the vector equations (4.2) *represent a steady state of the two-spatial-dimensional SG SWE* (2.11)*.*

4.1 Energy conservative schemes

The semi-discrete form for finite volume (FV) schemes for (2.11) on a uniform rectangular mesh reads,

$$\frac{d}{dt}\boldsymbol{U}_{i,j} = -\frac{\mathcal{F}_{i+\frac{1}{2},j} - \mathcal{F}_{i-\frac{1}{2},j}}{\Delta x} - \frac{\mathcal{G}_{i,j+\frac{1}{2}} - \mathcal{G}_{i,j-\frac{1}{2}}}{\Delta y} + \boldsymbol{S}_{i,j}. \tag{4.3}$$

The domain $\mathcal{D}$ is tessellated with rectangular cells $\mathcal{I}_{i,j} \coloneqq [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, with a uniform mesh size $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$. Hence, $\Delta x \Delta y$ is the size (area) of cell $\mathcal{I}_{i,j}$. We consider M cells in each direction so that the rectangular domain so that $i, j \in [M]$. The quantity $\boldsymbol{U}_{i,j}(t)$ denotes the numerical approximation to the cell average of the vector $\widehat{U}$ over cell $\mathcal{I}_{i,j}$ at time t, i.e., $\boldsymbol{U}_{i,j}(t) \approx \frac{1}{\Delta x \Delta y}\int_{\mathcal{I}_{i,j}} \widehat{U}(x, y, t)dxdy$. The numerical fluxes $\mathcal{F}_{i\pm\frac{1}{2},j}$ and $\mathcal{G}_{i,j\pm\frac{1}{2}}$ depend on the neighboring states, e.g., $\boldsymbol{U}_{i,j}$ and $\boldsymbol{U}_{i+1,j}$ for $\mathcal{F}_{i+\frac{1}{2},j}$. Additionally, $\boldsymbol{S}_{i,j} \approx \frac{1}{\Delta x \Delta y}\int_{\mathcal{I}_{i,j}} \widehat{S}(\widehat{U}, \widehat{B})$ represents the discretization of the source term, which should be designed to ensure a well-balanced numerical scheme in the sense of definition 1. Other boldface notations, such as $(\boldsymbol{h}_{i,j}, \boldsymbol{q}^{\boldsymbol{x}}_{i,j}, \boldsymbol{q}^{\boldsymbol{y}}_{i,j}, \boldsymbol{B}_{i,j})$, are defined in a similar manner. The discrete velocities $\boldsymbol{u}_{i,j}$ and $\boldsymbol{v}_{i,j}$ are defined as stochastic variants of (2.14), as follows:

$$\boldsymbol{u}_{i,j} \coloneqq \mathcal{P}^{-1}(\boldsymbol{h}_{i,j})\boldsymbol{q}^{\boldsymbol{x}}_{i,j}, \qquad \boldsymbol{v}_{i,j} \coloneqq \mathcal{P}^{-1}(\boldsymbol{h}_{i,j})\boldsymbol{q}^{\boldsymbol{y}}_{i,j}. \tag{4.4}$$

The discrete entropic quantities, stochastic variants of (3.11), are defined in terms of the discrete conservative variable $\boldsymbol{U}_{i,j}$ and the discrete velocities $\boldsymbol{u}_{i,j}, \boldsymbol{v}_{i,j}$ as follows:

$$\begin{aligned}
\boldsymbol{E}_{i,j} \coloneqq& \frac{1}{2}\Big((\boldsymbol{q}^{\boldsymbol{x}}_{i,j})^\top \boldsymbol{u}_{i,j} + (\boldsymbol{q}^{\boldsymbol{y}}_{i,j})^\top \boldsymbol{v}_{i,j}\Big) + \frac{1}{2}g\|\boldsymbol{h}_{i,j}\|^2 + g\boldsymbol{h}_{i,j}^\top \boldsymbol{B}_{i,j}, \\
\boldsymbol{V}_{i,j} \coloneqq& \left(\frac{\partial \boldsymbol{E}_{i,j}}{\partial \boldsymbol{U}_{i,j}}\right)^\top = \left(-\frac{1}{2}\boldsymbol{u}_{i,j}^\top \mathcal{P}(\boldsymbol{u}_{i,j}) - \frac{1}{2}\boldsymbol{v}_{i,j}^\top \mathcal{P}(\boldsymbol{v}_{i,j}) + g(\boldsymbol{h}_{i,j} + \boldsymbol{B}_{i,j})^\top, \quad \boldsymbol{u}_{i,j}^\top, \quad \boldsymbol{v}_{i,j}^\top\right)^\top, \\
\boldsymbol{\Psi}_{i,j} \coloneqq& \frac{1}{2}g\boldsymbol{u}_{i,j}^\top \mathcal{P}(\boldsymbol{h}_{i,j})\boldsymbol{h}_{i,j}, \quad \boldsymbol{\Phi}_{i,j} \coloneqq \frac{1}{2}g\boldsymbol{v}_{i,j}^\top \mathcal{P}(\boldsymbol{h}_{i,j})\boldsymbol{h}_{i,j}.
\end{aligned} \tag{4.5}$$

Our numerical schemes involve the average and jump quantities at cell interfaces [13, 21]:

$$\begin{aligned}
\overline{\boldsymbol{a}}_{i+\frac{1}{2},j} \coloneqq \frac{\boldsymbol{a}_{i,j} + \boldsymbol{a}_{i+1,j}}{2}, \qquad \overline{\boldsymbol{a}}_{i,j+\frac{1}{2}} \coloneqq \frac{\boldsymbol{a}_{i,j} + \boldsymbol{a}_{i,j+1}}{2}, \\
[\![\boldsymbol{a}]\!]_{i+\frac{1}{2},j} \coloneqq \boldsymbol{a}_{i+1,j} - \boldsymbol{a}_{i,j}, \quad [\![\boldsymbol{a}]\!]_{i,j+\frac{1}{2}} \coloneqq \boldsymbol{a}_{i,j+1} - \boldsymbol{a}_{i,j},
\end{aligned} \tag{4.6}$$

where $\boldsymbol{a}_{i,j}$ denotes any cell averaged quantity over $\mathcal{I}_{i,j}$, e.g., $\boldsymbol{U}_{i,j}$. The expressions (4.6) are equivalent to

$$\begin{aligned}
\boldsymbol{a}_{i,j} &= \overline{\boldsymbol{a}}_{i+\frac{1}{2},j} - \frac{[\![\boldsymbol{a}]\!]_{i+\frac{1}{2},j}}{2} = \overline{\boldsymbol{a}}_{i-\frac{1}{2},j} + \frac{[\![\boldsymbol{a}]\!]_{i-\frac{1}{2},j}}{2} \\
&= \overline{\boldsymbol{a}}_{i,j+\frac{1}{2}} - \frac{[\![\boldsymbol{a}]\!]_{i,j+\frac{1}{2}}}{2} = \overline{\boldsymbol{a}}_{i,j-\frac{1}{2}} + \frac{[\![\boldsymbol{a}]\!]_{i,j-\frac{1}{2}}}{2}.
\end{aligned} \tag{4.7}$$

Then we introduce additional equalities for the interfacial averages and jumps associated with the linear operator $\mathcal{P}$ in the following lemma. These are straightforward generalizations of [13, Lemma 4.1], so we omit the proof.

Lemma 4. *Let $\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}$ be any spatially discrete quantities, then*

$$\mathcal{P}(\overline{\boldsymbol{a}}_{i+\frac{1}{2},j})[\![\boldsymbol{a}]\!]_{i+\frac{1}{2},j} = \frac{1}{2}[\![\mathcal{P}(\boldsymbol{a})\boldsymbol{a}]\!]_{i+\frac{1}{2},j}, \qquad [\![\boldsymbol{a}]\!]^{\top}_{i+\frac{1}{2},j}\overline{\boldsymbol{b}}_{i+\frac{1}{2},j} + [\![\boldsymbol{b}]\!]^{\top}_{i+\frac{1}{2},j}\overline{\boldsymbol{a}}_{i+\frac{1}{2},j} = [\![\boldsymbol{a}^{\top}\boldsymbol{b}]\!]_{i+\frac{1}{2},j}, \tag{4.8a}$$

$$\mathcal{P}(\overline{\boldsymbol{a}}_{i,j+\frac{1}{2}})[\![\boldsymbol{a}]\!]_{i,j+\frac{1}{2}} = \frac{1}{2}[\![\mathcal{P}(\boldsymbol{a})\boldsymbol{a}]\!]_{i,j+\frac{1}{2}}, \qquad [\![\boldsymbol{a}]\!]^{\top}_{i,j+\frac{1}{2}}\overline{\boldsymbol{b}}_{i,j+\frac{1}{2}} + [\![\boldsymbol{b}]\!]^{\top}_{i,j+\frac{1}{2}}\overline{\boldsymbol{a}}_{i,j+\frac{1}{2}} = [\![\boldsymbol{a}^{\top}\boldsymbol{b}]\!]_{i,j+\frac{1}{2}}. \tag{4.8b}$$

To provide the specific definitions for energy conservative and energy stable schemes for systems of balance laws in two spatial dimensions, we recall that the semi-discrete FV form (4.3) is called a *conservative scheme* when the source term vanishes i.e., $\boldsymbol{S}_{i,j} = 0$. In this case, and by summing (4.3) over the cells, we obtain,

$$\frac{d}{dt}\sum_{i,j\in[M]}\boldsymbol{U}_{i,j}(t) = \sum_{j=1}^{M}\frac{\mathcal{F}_{\frac{1}{2},j} - \mathcal{F}_{M+\frac{1}{2},j}}{\Delta x} + \sum_{i=1}^{M}\frac{\mathcal{G}_{i,\frac{1}{2}} - \mathcal{G}_{i,M+\frac{1}{2}}}{\Delta y}. \tag{4.9}$$

This implies that, if periodic boundary conditions are imposed so that $\mathcal{F}_{\frac{1}{2},j} = \mathcal{F}_{M+\frac{1}{2},j}$ and similarly for $\mathcal{G}$, then $\widehat{U}$ remains constant over time. Since the entropy-flux pair variables (E, H, K) are explicit functions of the state variable $\widehat{U}$ and the inputs of the source term, i.e., $\widehat{B}$, the semi-discrete form (4.3) of the balance law (2.11) can be transformed into a semi-discrete form of the corresponding companion balance law (3.10) for the entropy (energy) of the system. The notions of energy conservative and energy stable schemes can be defined through the evolution of the entropy/energy of the system.

Definition 2 (Energy conservative and energy stable schemes)**.** *Suppose the system of balance laws* (2.11) *has an entropy flux pair (E, H, K), where $E(\widehat{U})$ represents the system's energy. Then the semi-discrete finite volume (FV) scheme* (4.3) *is an* ***Energy Conservative (EC)*** *scheme if it can be rewritten in the following semi-discrete form for the evolution of the numerical cell averages $\boldsymbol{E}_{i,j}$ of E:*

$$\frac{d}{dt}\boldsymbol{E}_{i,j}(t) = -\frac{1}{\Delta x}(\mathcal{H}_{i+\frac{1}{2},j} - \mathcal{H}_{i-\frac{1}{2},j}) - \frac{1}{\Delta y}(\mathcal{K}_{i,j+\frac{1}{2}} - \mathcal{K}_{i,j-\frac{1}{2}}), \tag{4.10}$$

where $\mathcal{H}_{i+\frac{1}{2},j}$ represents the numerical entropy flux at the interface $(x, y) = (x_{i+\frac{1}{2}}, y_j)$, and $\mathcal{K}_{i,j+\frac{1}{2}}$ represents another numerical entropy flux at the interface location $(x, y) = (x_i, y_{j+\frac{1}{2}})$.

Alternatively, the scheme (4.3) *is called an* ***Energy Stable (ES)*** *scheme under the weaker condition,*

$$\frac{d}{dt}\boldsymbol{E}_{i,j}(t) \le -\frac{1}{\Delta x}(\mathcal{H}_{i+\frac{1}{2},j} - \mathcal{H}_{i-\frac{1}{2},j}) - \frac{1}{\Delta y}(\mathcal{K}_{i,j+\frac{1}{2}} - \mathcal{K}_{i,j-\frac{1}{2}}). \tag{4.11}$$

By summing (4.10) over all cells of the computational domain, a form similar to (4.9) can be obtained, but for energy instead of the state variable $\boldsymbol{U}$, showing that the cumulative energy remains constant over time under periodic boundary conditions. For non-periodic boundary conditions, energy can increase due to the boundary terms; we discuss a notion of *augmented* energy in section 5.5 that attempts to separate this potential intrinsic energy increase effect from any energy increase due to numerical discretizations.

4.2 An energy conservative scheme for the 2D SG SWE (EC)

For the scheme (4.3), we make the following choices for the fluxes and balance terms:

$$
\begin{aligned}
\mathcal{F}_{i+\frac{1}{2},j} = \mathcal{F}^{EC}_{i+\frac{1}{2},j} &= \begin{pmatrix} \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} \\ \frac{1}{2}g(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}})_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{u}}_{i+\frac{1}{2},j})\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} \\ \mathcal{P}(\overline{\boldsymbol{v}}_{i+\frac{1}{2},j})\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} \end{pmatrix}, \\
\mathcal{G}_{i,j+\frac{1}{2}} = \mathcal{G}^{EC}_{i,j+\frac{1}{2}} &= \begin{pmatrix} \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}} \\ \mathcal{P}(\overline{\boldsymbol{u}}_{i,j+\frac{1}{2}})\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}}, \\ \frac{1}{2}g(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}})_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}})\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}} \end{pmatrix}, \\
\boldsymbol{S}_{i,j} &= \begin{pmatrix} 0 \\ -\frac{g}{2\Delta x}(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i-\frac{1}{2},j}) \\ -\frac{g}{2\Delta y}(\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j-\frac{1}{2}}) \end{pmatrix}.
\end{aligned}
\tag{4.12}
$$

These choices ensure that the scheme is energy conservative and well-balanced.

Theorem 3 (Second-order EC well-balanced scheme)**.** *Suppose the bottom topography function B is independent of time. The semi-discrete FV scheme* (4.3) *for the two-spatial-dimensional SG SWE system* (2.11) *with fluxes and source term in* (4.12) *is a well-balanced EC scheme with local truncation error $\mathcal{O}(\Delta x^2 + \Delta y^2)$.*

The proof consists of several parts, which require nontrivial extensions of the one-dimensional results in [13]. However, because the proof components are relatively technical, we place their presentation in the appendix. First, we demonstrate the second-order truncation error through direct computation.

Lemma 5 (Second-order truncation error)**.** *By selecting the numerical fluxes and the source term* (4.12)*, the semi-discrete FV scheme* (4.3) *has a local truncation error of $\mathcal{O}(\Delta x^2 + \Delta y^2)$.*

See appendix B.1 for the detailed proof. Next, we show that the discretization of the source term $\boldsymbol{S}_{i,j}$ in (4.12) satisfies the well-balanced property, which is the extension of results in [13] to two spatial dimensions.

Lemma 6 (Well-balanced property)**.** *Suppose the source term is chosen as* (4.12) *and the bottom topography function B is time-independent, then the semi-discrete FV scheme* (4.3) *is well-balanced.*

The proof of the above result is contained in appendix B.2. The final step in proving the main theorem of this section is to identify a sufficient condition that ensures the numerical fluxes produce an EC scheme. This is a two-spatial-dimensional extension of that in [13] and also a stochastic variant of that in the deterministic case in [21].

Lemma 7 (A sufficient condition for EC schemes)**.** *Let $\boldsymbol{S}_{i,j}$ be selected as in* (4.12)*. Suppose the numerical fluxes $\mathcal{F}_{i+\frac{1}{2},j}, \mathcal{G}_{i,j+\frac{1}{2}}$ satisfy*

$$
\begin{aligned}
[\![\boldsymbol{V}]\!]^\top_{i+\frac{1}{2},j}\mathcal{F}_{i+\frac{1}{2},j} &= [\![\boldsymbol{\Psi}]\!]_{i+\frac{1}{2},j} + g[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{u}_{i+\frac{1}{2},j}, \\
[\![\boldsymbol{V}]\!]^\top_{i,j+\frac{1}{2}}\mathcal{G}_{i,j+\frac{1}{2}} &= [\![\boldsymbol{\Phi}]\!]_{i,j+\frac{1}{2}} + g[\![\boldsymbol{B}]\!]^\top_{i,j+\frac{1}{2}}\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{v}_{i,j+\frac{1}{2}},
\end{aligned}
\tag{4.13}
$$

where the discrete stochastic energy potentials $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ are defined in (4.5)*. Then the corresponding FV scheme* (4.3) *is an EC scheme, i.e., it satisfies* (4.10)*, with the numerical energy fluxes given by,*

$$
\begin{aligned}
\mathcal{H}_{i+\frac{1}{2},j} &\coloneqq \overline{\boldsymbol{V}}^\top_{i+\frac{1}{2},j}\mathcal{F}_{i+\frac{1}{2},j} - \overline{\boldsymbol{\Psi}}_{i+\frac{1}{2},j} - \frac{g}{4}[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{u}]\!]_{i+\frac{1}{2},j} \\
\mathcal{K}_{i,j+\frac{1}{2}} &\coloneqq \overline{\boldsymbol{V}}^\top_{i,j+\frac{1}{2}}\mathcal{G}_{i,j+\frac{1}{2}} - \overline{\boldsymbol{\Phi}}_{i,j+\frac{1}{2}} - \frac{g}{4}[\![\boldsymbol{B}]\!]^\top_{i,j+\frac{1}{2}}\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{v}]\!]_{i,j+\frac{1}{2}}.
\end{aligned}
\tag{4.14}
$$

For the proof, see appendix B.3. Given all essential lemmas above, we can complete the proof of Theorem 3.

Proof of Theorem 3. It follows from Lemmas 5 and 6 that the scheme (4.3) is well-balanced and second-order accurate for smooth solutions. To show that it is EC, it suffices to verify the condition in Lemma 7, which can be accomplished by direct computation:

$$
\begin{aligned}
&[\![\boldsymbol{V}]\!]^{\top}_{i+\frac{1}{2},j}\mathcal{F}^{EC}_{i+\frac{1}{2},j}\\
\overset{(4.5),(4.12)}{=}&\left(g\left([\![\boldsymbol{h}]\!]_{i+\frac{1}{2},j}+[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j}\right)-\frac{1}{2}[\![\mathcal{P}(\boldsymbol{u})\boldsymbol{u}]\!]_{i+\frac{1}{2},j}-\frac{1}{2}[\![\mathcal{P}(\boldsymbol{v})\boldsymbol{v}]\!]_{i+\frac{1}{2},j}\right)^{\top}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}\\
&+[\![\boldsymbol{u}]\!]^{\top}_{i+\frac{1}{2},j}\left(\frac{1}{2}g\left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i+\frac{1}{2},j}+\mathcal{P}(\overline{\boldsymbol{u}}_{i+\frac{1}{2},j})\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}\right)\\
&+[\![\boldsymbol{v}]\!]^{\top}_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{v}}_{i+\frac{1}{2},j})\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}\\
\overset{(4.8a)}{=}&g\left([\![\boldsymbol{h}]\!]_{i+\frac{1}{2},j}+[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j}\right)^{\top}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}+\frac{1}{2}g[\![\boldsymbol{u}]\!]^{\top}_{i+\frac{1}{2},j}\left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i+\frac{1}{2},j}\\
\overset{(4.8a)}{=}&\frac{1}{2}g[\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]^{\top}_{i+\frac{1}{2},j}\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}+g[\![\boldsymbol{B}]\!]^{\top}_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}+\frac{1}{2}g[\![\boldsymbol{u}]\!]^{\top}_{i+\frac{1}{2},j}\left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i+\frac{1}{2},j}\\
\overset{(4.8a)}{=}&\frac{1}{2}g[\![\boldsymbol{u}^{\top}\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i+\frac{1}{2},j}+g[\![\boldsymbol{B}]\!]^{\top}_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}\\
=&[\![\boldsymbol{\Psi}]\!]_{i+\frac{1}{2},j}+g[\![\boldsymbol{B}]\!]^{\top}_{i+\frac{1}{2},j}\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}.
\end{aligned}
\tag{4.15}
$$

Also, by using the definition of $\mathcal{G}^{EC}_{i,j+\frac{1}{2}}$ (4.12), $\boldsymbol{\Phi}$ (4.5), and equalities (4.8a), through an analogous computation as above, one can obtain

$$
[\![\boldsymbol{V}]\!]^{\top}_{i,j+\frac{1}{2}}\mathcal{G}^{EC}_{i,j+\frac{1}{2}}=[\![\boldsymbol{\Phi}]\!]_{i,j+\frac{1}{2}}+g[\![\boldsymbol{B}]\!]^{\top}_{i,j+\frac{1}{2}}\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{v}_{i,j+\frac{1}{2}},
\tag{4.16}
$$

which verifies (4.13). Hence, it shows that the scheme (4.3), with the numerical fluxes and source term defined in (4.12), is energy conservative.

□

4.3 A first-order energy stable scheme (ES1)

The scheme defined in the previous part preserves the energy of the shallow water equation system, which can lead to non-physical oscillations, since the energy is supposed to dissipate in the presence of shocks. Some existing work can resolve this issue by introducing artificial numerical viscosity [20, 21, 22, 51, 52]. We adopt this approach by extending the energy-stable diffusion operators proposed in [20, 21, 22] to the stochastic case in two spatial dimensions.

First, we recall the traditional Roe-type diffusion ("*RD*") applied to a conservation law, which involves the EC fluxes, defined as follows:

$$
\mathcal{F}^{RD}_{i+\frac{1}{2},j}\coloneqq\mathcal{F}^{EC}_{i+\frac{1}{2},j}-\frac{1}{2}\boldsymbol{Q}^{Roe,F}_{i+\frac{1}{2},j}[\![\boldsymbol{U}]\!]_{i+\frac{1}{2},j},\qquad\qquad\mathcal{G}^{RD}_{i,j+\frac{1}{2}}\coloneqq\mathcal{G}^{EC}_{i,j+\frac{1}{2}}-\frac{1}{2}\boldsymbol{Q}^{Roe,G}_{i,j+\frac{1}{2}}[\![\boldsymbol{U}]\!]_{i,j+\frac{1}{2}}
\tag{4.17}
$$

where $\boldsymbol{Q}^{Roe,F}$ and $\boldsymbol{Q}^{Roe,G}$ are positive semi-definite matrices defined by diagonalizing the Jacobians of the interfacial fluxes at the Roe-averaged state $\overline{\boldsymbol{U}}$:

$$
\boldsymbol{Q}^{Roe,F}_{i+\frac{1}{2},j}\coloneqq\boldsymbol{T}^{Roe}_{F}\left|\boldsymbol{\Lambda}^{Roe}_{F}\right|\left(\boldsymbol{T}^{Roe}_{F}\right)^{-1},\qquad\qquad\frac{\partial\widehat{F}}{\partial\widehat{U}}(\overline{\boldsymbol{U}}_{i+\frac{1}{2},j})=\boldsymbol{T}^{Roe}_{F}\boldsymbol{\Lambda}^{Roe}_{F}\left(\boldsymbol{T}^{Roe}_{F}\right)^{-1},
\tag{4.18}
$$

$$
\boldsymbol{Q}^{Roe,G}_{i,j+\frac{1}{2}}\coloneqq\boldsymbol{T}^{Roe}_{G}\left|\boldsymbol{\Lambda}^{Roe}_{G}\right|\left(\boldsymbol{T}^{Roe}_{G}\right)^{-1},\qquad\qquad\frac{\partial\widehat{G}}{\partial\widehat{U}}(\overline{\boldsymbol{U}}_{i,j+\frac{1}{2}})=\boldsymbol{T}^{Roe}_{G}\boldsymbol{\Lambda}^{Roe}_{G}\left(\boldsymbol{T}^{Roe}_{G}\right)^{-1}.
\tag{4.19}
$$

The semi-discrete scheme (4.3) with the numerical fluxes $\mathcal{F}^{RD}_{i+\frac{1}{2},j}$ and $\mathcal{G}^{RD}_{i,j+\frac{1}{2}}$ would behave like

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{U}_{i,j}(t) = &- \frac{\mathcal{F}^{EC}_{i+\frac{1}{2},j} - \mathcal{F}^{EC}_{i-\frac{1}{2},j}}{\Delta x} - \frac{\mathcal{G}^{EC}_{i,j+\frac{1}{2}} - \mathcal{G}^{EC}_{i,j-\frac{1}{2}}}{\Delta y} + \boldsymbol{S}_{i,j} \\
&+ \frac{1}{2\Delta x}\left(\boldsymbol{Q}^{Roe,F}_{i+\frac{1}{2},j}[\![\boldsymbol{U}]\!]_{i+\frac{1}{2},j} - \boldsymbol{Q}^{Roe,F}_{i-\frac{1}{2},j}[\![\boldsymbol{U}]\!]_{i-\frac{1}{2},j}\right) \\
&+ \frac{1}{2\Delta y}\left(\boldsymbol{Q}^{Roe,G}_{i,j+\frac{1}{2}}[\![\boldsymbol{U}]\!]_{i,j+\frac{1}{2}} - \boldsymbol{Q}^{Roe,G}_{i,j-\frac{1}{2}}[\![\boldsymbol{U}]\!]_{i,j-\frac{1}{2}}\right) \\
\approx &- \frac{1}{\Delta x}\left(\widehat{F}(\widehat{U})\big|_{(x_{i+\frac{1}{2}},y_j)} - \widehat{F}(\widehat{U})\big|_{(x_{i-\frac{1}{2}},y_j)}\right) - \frac{1}{\Delta y}\left(\widehat{G}(\widehat{U})\big|_{(x_i,y_{j+\frac{1}{2}})} - \widehat{G}(\widehat{U})\big|_{(x_i,y_{j-\frac{1}{2}})}\right) \\
&+ \Delta x \boldsymbol{Q}_F \widehat{U}_{xx}\big|_{(x_i,y_j)} + \Delta y \boldsymbol{Q}_G \widehat{U}_{yy}\big|_{(x_i,y_j)} + S(\widehat{U})\big|_{(x_i,y_j)},
\end{aligned} \tag{4.20}$$

where $\boldsymbol{Q}_F$ and $\boldsymbol{Q}_G$ are positive-definite matrices, and $\widehat{U}_{xx}$ and $\widehat{U}_{yy}$ represent the second-order spatial derivatives of the state variables in the PDE, thus introducing diffusion into the EC scheme.

It is convenient to ensure energy stability by adding a numerical diffusion term operating on the entropic variables $\boldsymbol{V}$ rather than on the conservative variables $\boldsymbol{U}$. We introduce the following first-order energy stable numerical fluxes:

$$\mathcal{F}^{ES1}_{i+\frac{1}{2},j} \coloneqq \mathcal{F}^{EC}_{i+\frac{1}{2},j} - \frac{1}{2}\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j}, \qquad \mathcal{G}^{ES1}_{i,j+\frac{1}{2}} \coloneqq \mathcal{G}^{EC}_{i,j+\frac{1}{2}} - \frac{1}{2}\boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}}, \tag{4.21}$$

where $\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}$ and $\boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}$ are positive definite matrices identified in a Roe-type manner from the adjacent states $(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i+1,j})$ at the cell interface $(x_{i+\frac{1}{2}}, y_j)$, and $(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i,j+1})$ at the cell interface $(x_i, y_{j+\frac{1}{2}})$, respectively. The term $\boldsymbol{V}_{i,j}$, defined in (4.5), serves as a second-order approximation of the cell average of the entropy variable $\widehat{V}$. The matrices $\boldsymbol{Q}^{ES,F}$ and $\boldsymbol{Q}^{ES,G}$ will be defined in terms of the following Roe-type energy stable operator,

$$\mathcal{Q}_{i+\frac{1}{2},j}(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i+1,j}) \coloneqq \boldsymbol{T}_F|\boldsymbol{\Lambda}_F|\boldsymbol{T}_F^\top \geq 0, \qquad \mathcal{Q}_{i,j+\frac{1}{2}}(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i,j+1}) \coloneqq \boldsymbol{T}_G|\boldsymbol{\Lambda}_G|\boldsymbol{T}_G^\top \geq 0, \tag{4.22}$$

where the matrices $\boldsymbol{T}_F, \boldsymbol{\Lambda}_F, \boldsymbol{T}_G, \boldsymbol{\Lambda}_G$ are defined from the eigendecomposition of the flux Jacobians $\frac{\partial \widehat{F}}{\partial \widehat{U}}$ and $\frac{\partial \widehat{G}}{\partial \widehat{U}}$ evaluated at the following Roe-type averaged state,

$$\frac{\partial \widehat{F}}{\partial \widehat{U}}\left(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2},j}\right) = \boldsymbol{T}_F\boldsymbol{\Lambda}_F\boldsymbol{T}_F^{-1}, \qquad \widetilde{\boldsymbol{U}}_{i+\frac{1}{2},j} \coloneqq \begin{pmatrix} \overline{\boldsymbol{h}}_{i+\frac{1}{2},j} \\ \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} \\ \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{v}}_{i+\frac{1}{2},j}, \end{pmatrix}, \tag{4.23a}$$

$$\frac{\partial \widehat{G}}{\partial \widehat{U}}\left(\widetilde{\boldsymbol{U}}_{i,j+\frac{1}{2}}\right) = \boldsymbol{T}_G\boldsymbol{\Lambda}_G\boldsymbol{T}_G^{-1}, \qquad \widetilde{\boldsymbol{U}}_{i,j+\frac{1}{2}} \coloneqq \begin{pmatrix} \overline{\boldsymbol{h}}_{i,j+\frac{1}{2}} \\ \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{u}}_{i,j+\frac{1}{2}} \\ \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}} \end{pmatrix}. \tag{4.23b}$$

Then we set the diffusion matrices $\boldsymbol{Q}^{ES,F}$ and $\boldsymbol{Q}^{ES,G}$ appearing in the definition (4.21) of the numerical fluxes $\mathcal{F}^{ES1}_{i+\frac{1}{2},j}$ and $\mathcal{G}^{ES1}_{i,j+\frac{1}{2}}$, respectively, as,

$$\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j} \coloneqq \mathcal{Q}_{i+\frac{1}{2},j}(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i+1,j}) = \boldsymbol{T}_F|\boldsymbol{\Lambda}_F|\boldsymbol{T}_F^\top. \tag{4.24a}$$

$$\boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}} \coloneqq \mathcal{Q}_{i,j+\frac{1}{2}}(\boldsymbol{U}_{i,j}, \boldsymbol{U}_{i,j+1}) = \boldsymbol{T}_G|\boldsymbol{\Lambda}_G|\boldsymbol{T}_G^\top, \tag{4.24b}$$

The energy stable scheme in this subsection is constructed by using the numerical fluxes in (4.21), with the diffusion matrices given in (4.24). We codify the properties of this scheme as follows.

Theorem 4 (ES1 scheme)**.** *Consider the semi-discrete FV scheme* (4.3) *with the source term given in* (4.12) *and diffusive numerical fluxes* $\mathcal{F}^{ES1}_{i+\frac{1}{2},j}, \mathcal{G}^{ES1}_{i,j+\frac{1}{2}}$ *defined in* (4.21)*, and the diffusion matrices* $\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}, \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}$ *as defined in* (4.24)*. The resulting scheme is a first-order, well-balanced, and energy stable scheme.*

Proof. Note that we have already shown that the scheme (4.3) with EC fluxes $\mathcal{F}^{EC}_{i+\frac{1}{2},j}$ and $\mathcal{G}^{EC}_{i,j+\frac{1}{2}}$ is second-order accurate. Therefore, using the definition of $\boldsymbol{V}_{i,j}$, it is straightforward to show the following approximation

$$[\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j} \approx \Delta x \widehat{V}_x(x_{i+\frac{1}{2}}, y_j), \qquad [\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}} \approx \Delta y \widehat{V}_y(x_i, y_{j+\frac{1}{2}}), \tag{4.25}$$

which implies that the artificial diffusion term in (4.21) introduces a first-order local truncation error in the scheme.

To demonstrate that this scheme is well-balanced, we consider the stochastic lake-at-rest initial data described as introduced in (B.9) in section 4.1. This, combined with the definition of $\boldsymbol{V}_{i,j}$ in (4.5), leads to the conclusion that $[\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j} = [\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}} = \boldsymbol{0}$. Since the EC fluxes are shown to be well-balanced in Lemma 6, it follows that the ES1 scheme is also well-balanced.

Finally, we need to show the ES property (4.11). Define the ES1 energy fluxes

$$\begin{aligned} \mathcal{H}^{ES1}_{i+\frac{1}{2},j} &\coloneqq \mathcal{H}^{EC}_{i+\frac{1}{2},j} - \frac{1}{2}\overline{\boldsymbol{V}}^{\top}_{i+\frac{1}{2},j} \boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j} [\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j} \\ \mathcal{K}^{ES1}_{i,j+\frac{1}{2}} &\coloneqq \mathcal{K}^{EC}_{i,j+\frac{1}{2}} - \frac{1}{2}\overline{\boldsymbol{V}}^{\top}_{i,j+\frac{1}{2}} \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}} [\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}}, \end{aligned} \tag{4.26}$$

where the EC energy fluxes $\mathcal{H}^{EC}_{i+\frac{1}{2},j}$ and $\mathcal{K}^{EC}_{i,j+\frac{1}{2}}$ are defined in (4.14). As in the proof of Lemma 7 (appendix B.3), we multiply (4.3) by $\boldsymbol{V}^{\top}_{i,j}$, and use a similar estimate with the ES1 energy fluxes defined above. This leads to the following estimate:

$$\begin{aligned} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) = &- \frac{1}{\Delta x}\left(\mathcal{H}^{ES1}_{i+\frac{1}{2},j} - \mathcal{H}^{ES1}_{i-\frac{1}{2},j}\right) - \frac{1}{\Delta y}\left(\mathcal{K}^{ES1}_{i,j+\frac{1}{2}} - \mathcal{K}^{ES1}_{i,j-\frac{1}{2}}\right) \\ &- \frac{1}{4\Delta x}\left([\![\boldsymbol{V}]\!]^{\top}_{i+\frac{1}{2},j} \boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j} [\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j} + [\![\boldsymbol{V}]\!]^{\top}_{i-\frac{1}{2},j} \boldsymbol{Q}^{ES,F}_{i-\frac{1}{2},j} [\![\boldsymbol{V}]\!]_{i-\frac{1}{2},j}\right) \\ &- \frac{1}{4\Delta y}\left([\![\boldsymbol{V}]\!]^{\top}_{i,j+\frac{1}{2}} \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}} [\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}} + [\![\boldsymbol{V}]\!]^{\top}_{i,j-\frac{1}{2}} \boldsymbol{Q}^{ES,G}_{i,j-\frac{1}{2}} [\![\boldsymbol{V}]\!]_{i,j-\frac{1}{2}}\right). \end{aligned} \tag{4.27}$$

Since the diffusion matrices $\boldsymbol{Q}^{ES,F}_{i\pm\frac{1}{2},j}, \boldsymbol{Q}^{ES,G}_{i,j\pm\frac{1}{2}}$ are positive semi-definite, this scheme satisfies the energy decay property (4.11), resulting in an ES scheme. □

4.4 A second-order energy stable scheme (ES2)

The ES1 scheme was derived by adding a first-order diffusion term to the second-order EC scheme. It is thus natural to develop a second-order ES scheme by incorporating a suitably constructed, second-order accurate diffusion term. This can be achieved through appropriate non-oscillatory second-order polynomial reconstructions of the entropy variable. We follow the idea of [13, 22] to recover a non-oscillatory piecewise linear reconstruction. Let $\boldsymbol{V}^{E}_{i,j}$ and $\boldsymbol{V}^{W}_{i+1,j}$ denote the second-order reconstructions from the east and west, respectively, of the entropy variable $\boldsymbol{V}(x_{i+\frac{1}{2}}, y_j)$. Similarly, let $\boldsymbol{V}^{N}_{i,j}$ and $\boldsymbol{V}^{S}_{i,j+1}$ represent the second-order reconstructions from the north and south of the entropy variable $\boldsymbol{V}(x_i, y_{j+\frac{1}{2}})$. I.e., if $\widetilde{\boldsymbol{V}}_{i,j}(x,y)$ is the polynomial reconstruction of the entropy variable $\boldsymbol{V}$ in the cell $\mathcal{I}_{i,j}$, then,

$$\boldsymbol{V}^{E}_{i,j} \coloneqq \lim_{\substack{x\to x_{i+1/2}\\ y\to y_j}} \widetilde{\boldsymbol{V}}_{i,j}(x,y), \qquad \boldsymbol{V}^{W}_{i,j} \coloneqq \lim_{\substack{x\to x_{i-1/2}\\ y\to y_j}} \widetilde{\boldsymbol{V}}_{i,j}(x,y), \tag{4.28}$$

$$\boldsymbol{V}^{N}_{i,j} \coloneqq \lim_{\substack{x\to x_i\\ y\to y_{j+1/2}}} \widetilde{\boldsymbol{V}}_{i,j}(x,y), \qquad \boldsymbol{V}^{S}_{i,j} \coloneqq \lim_{\substack{x\to x_i\\ y\to y_{j-1/2}}} \widetilde{\boldsymbol{V}}_{i,j}(x,y). \tag{4.29}$$

Using these reconstructions, we can compute second-order accurate jumps of the entropy variable,

$$\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2},j} \coloneqq \boldsymbol{V}^W_{i+1,j} - \boldsymbol{V}^E_{i,j}, \qquad \langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i,j+\frac{1}{2}} \coloneqq \boldsymbol{V}^S_{i,j+1} - \boldsymbol{V}^N_{i,j}. \tag{4.30}$$

Using the second-order jumps in the energy stable fluxes, we obtain the ES2 fluxes,

$$\mathcal{F}^{ES2}_{i+\frac{1}{2},j} \coloneqq \mathcal{F}^{EC}_{i+\frac{1}{2},j} - \frac{1}{2}\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2},j}, \qquad \mathcal{G}^{ES2}_{i,j+\frac{1}{2}} \coloneqq \mathcal{G}^{EC}_{i,j+\frac{1}{2}} - \frac{1}{2}\boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i,j+\frac{1}{2}}, \tag{4.31}$$

where the positive definite matrices $\boldsymbol{Q}^{ES,F}$ and $\boldsymbol{Q}^{ES,G}$ are the same as those in the ES1 scheme (4.21). The remaining part of this section describes how to compute $\boldsymbol{V}^W_{i+1,j}, \boldsymbol{V}^E_{i,j}, \boldsymbol{V}^S_{i,j+1}, \boldsymbol{V}^N_{i,j}$ in a manner that ensures entropy stability.

The first step is to use a scaled version of the entropy variables as follows:

$$\boldsymbol{w}^E_{i,j} \coloneqq (\boldsymbol{T}_F)^\top_{i+\frac{1}{2},j}\boldsymbol{V}_{i,j}, \quad w^W_{i,j} \coloneqq (\boldsymbol{T}_F)^\top_{i-\frac{1}{2},j}\boldsymbol{V}_{i,j}, \quad \boldsymbol{w}^N_{i,j} \coloneqq (\boldsymbol{T}_G)^\top_{i,j+\frac{1}{2}}\boldsymbol{V}_{i,j}, \quad w^S_{i,j} \coloneqq (\boldsymbol{T}_G)^\top_{i,j-\frac{1}{2}}\boldsymbol{V}_{i,j}, \tag{4.32}$$

where the matrices $\boldsymbol{T}_F, \boldsymbol{T}_G$ are defined in (4.23). Next, we perform a second-order total variation-diminishing (TVD) reconstruction of the scaled variable $\boldsymbol{w}$ at the cell interfaces

$$\begin{aligned} \widetilde{\boldsymbol{w}}^E_{i,j} &\coloneqq \boldsymbol{w}^E_{i,j} + \frac{1}{2}\phi(\boldsymbol{\theta}^E_{i,j}) \circ \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i+\frac{1}{2},j}, \quad \widetilde{\boldsymbol{w}}^W_{i,j} \coloneqq \boldsymbol{w}^W_{i,j} - \frac{1}{2}\phi(\boldsymbol{\theta}^W_{i,j}) \circ \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i-\frac{1}{2},j}, \\ \widetilde{\boldsymbol{w}}^N_{i,j} &\coloneqq \boldsymbol{w}^N_{i,j} + \frac{1}{2}\phi(\boldsymbol{\theta}^N_{i,j}) \circ \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j+\frac{1}{2}}, \quad \widetilde{\boldsymbol{w}}^S_{i,j} \coloneqq \boldsymbol{w}^S_{i,j} - \frac{1}{2}\phi(\boldsymbol{\theta}^S_{i,j}) \circ \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j-\frac{1}{2}}, \end{aligned} \tag{4.33}$$

where the jump $\langle\!\langle \cdot \rangle\!\rangle$ is defined in (4.30), $\circ$ denotes the Hadamard (elementwise) product on vectors, and $\boldsymbol{\theta}_{i,j}$ represents the difference quotients, defined by

$$\boldsymbol{\theta}^E_{i,j} \coloneqq \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i-\frac{1}{2},j} \oslash \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i+\frac{1}{2},j}, \qquad \boldsymbol{\theta}^W_{i,j} \coloneqq \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i+\frac{1}{2},j} \oslash \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i-\frac{1}{2},j}, \tag{4.34}$$

$$\boldsymbol{\theta}^N_{i,j} \coloneqq \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j-\frac{1}{2}} \oslash \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j+\frac{1}{2}}, \qquad \boldsymbol{\theta}^S_{i,j} \coloneqq \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j+\frac{1}{2}} \oslash \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i,j-\frac{1}{2}}, \tag{4.35}$$

where $\oslash$ is the Hadamard (elementwise) division between vectors. To preserve the TVD property, as discussed in [13, 22], we select the function ϕ as the minmod limiter, defined by

$$\phi(\theta) = \begin{cases} 0, & \theta < 0, \\ \theta, & 0 \le \theta \le 1, \\ 1, & \text{otherwise}, \end{cases} \tag{4.36}$$

which operates elementwise on vector inputs. Finally, the desired reconstructions for $\boldsymbol{V}^E_{i,j}, \boldsymbol{V}^W_{i,j}, \boldsymbol{V}^N_{i,j}, \boldsymbol{V}^S_{i,j}$ can be computed by inverting the $\boldsymbol{w}$-to-$\boldsymbol{V}$ map, respectively,

$$(\boldsymbol{T}_F)^\top_{i+\frac{1}{2},j}\boldsymbol{V}^E_{i,j} \coloneqq \widetilde{\boldsymbol{w}}^E_{i,j}, \quad (\boldsymbol{T}_F)^\top_{i-\frac{1}{2},j}\boldsymbol{V}^W_{i,j} \coloneqq \widetilde{\boldsymbol{w}}^W_{i,j}, \quad (\boldsymbol{T}_G)^\top_{i,j+\frac{1}{2}}\boldsymbol{V}^N_{i,j} \coloneqq \widetilde{\boldsymbol{w}}^N_{i,j} \quad (\boldsymbol{T}_G)^\top_{i,j-\frac{1}{2}}\boldsymbol{V}^S_{i,j} \coloneqq \widetilde{\boldsymbol{w}}^S_{i,j}. \tag{4.37}$$

The final ES2 scheme defined in (4.31) satisfies the desired properties.

Theorem 5 (ES2 scheme)**.** *The FV scheme* (4.3)*, with the source term in* (4.12) *and the diffusive numerical fluxes* $\mathcal{F}^{ES2}_{i+\frac{1}{2},j}, \mathcal{G}^{ES2}_{i,j+\frac{1}{2}}$ *defined in* (4.31)*, is a second-order, well-balanced, and energy stable scheme.*

The second-order accuracy results from the fact that the jumps $\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2},j}$ and $\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i,j+\frac{1}{2}}$ for each i, j are computed by using second-order accurate reconstructions. The well-balanced property can be shown in the same manner as in the ES1 scheme, following from the definition of $\boldsymbol{V}_{i,j}$ and the assumption that the stochastic lake-at-rest initial data implies that the jumps are zero. To show the energy stability property, we refer to the following lemma from [22] adapted to the two-dimensional case as stated below.

Lemma 8 ([22], Lemma 3.2)**.** *For each i, j, if there exists positive diagonal matrices $\mathbf{\Pi}_{i+\frac{1}{2},j} \geq 0$ and $\mathbf{\Pi}_{i,j+\frac{1}{2}} \geq 0$, s.t., the second-order jumps satisfy,*

$$\begin{aligned} \langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2},j} &= \left((\boldsymbol{T}_F)^{\top}_{i+\frac{1}{2},j} \right)^{-1} \mathbf{\Pi}_{i+\frac{1}{2},j} (\boldsymbol{T}_F)^{\top}_{i+\frac{1}{2},j} [\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j}, \\ \langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i,j+\frac{1}{2}} &= \left((\boldsymbol{T}_G)^{\top}_{i,j+\frac{1}{2}} \right)^{-1} \mathbf{\Pi}_{i,j+\frac{1}{2}} (\boldsymbol{T}_G)^{\top}_{i,j+\frac{1}{2}} [\![\boldsymbol{V}]\!]_{i,j+\frac{1}{2}}, \end{aligned} \tag{4.38}$$

then the scheme (4.3) *with fluxes $\mathcal{F}_{i+\frac{1}{2},j} = \mathcal{F}^{ES2}_{i+\frac{1}{2},j}$ and $\mathcal{G}_{i,j+\frac{1}{2}} = \mathcal{G}^{ES2}_{i,j+\frac{1}{2}}$ is an ES scheme.*

The proof of this lemma, which we omit, is accomplished by multiplying the FV scheme (4.3) by $\boldsymbol{V}^{\top}_{i,j}$ and following a similar approach as it in the proof of Lemma 7. We turn to complete the proof of Theorem 5.

Proof of Theorem 5. By Lemma 8, it suffices to establish the equalities (4.38) to demonstrate the ES property of the ES2 scheme. The definition of (4.33) implies

$$\begin{aligned} \langle\!\langle \widetilde{\boldsymbol{w}} \rangle\!\rangle^{l}_{i+\frac{1}{2},j} &= \left(1 - \frac{1}{2}\phi\big((\boldsymbol{\theta}^{W}_{i+1,j})^{l}\big) - \frac{1}{2}\big((\boldsymbol{\theta}^{E}_{i,j})^{l}\big) \right) \langle\!\langle \boldsymbol{w} \rangle\!\rangle^{l}_{i+\frac{1}{2},j}, \\ \langle\!\langle \widetilde{\boldsymbol{w}} \rangle\!\rangle^{l}_{i,j+\frac{1}{2}} &= \left(1 - \frac{1}{2}\phi\big((\boldsymbol{\theta}^{S}_{i,j+1})^{l}\big) - \frac{1}{2}\big((\boldsymbol{\theta}^{N}_{i,j})^{l}\big) \right) \langle\!\langle \boldsymbol{w} \rangle\!\rangle^{l}_{i,j+\frac{1}{2}}, \end{aligned} \tag{4.39}$$

where l denotes the index of the components of the corresponding vectors. The equality above, together with the definitions in (4.32), (4.37), and the linearity of $\langle\!\langle \cdot \rangle\!\rangle$ and $[\![\cdot]\!]$, implies (4.38) with the diagonal matrices $\mathbf{\Pi}_{i+\frac{1}{2},j}$ and $\mathbf{\Pi}_{i,j+\frac{1}{2}}$ defined by

$$(\mathbf{\Pi}_{i+\frac{1}{2},j})_{l,l} = 1 - \frac{1}{2}\phi\big((\boldsymbol{\theta}^{W}_{i+1,j})^{l}\big) - \frac{1}{2}\big((\boldsymbol{\theta}^{E}_{i,j})^{l}\big), \qquad (\mathbf{\Pi}_{i,j+\frac{1}{2}})_{l,l} = 1 - \frac{1}{2}\phi\big((\boldsymbol{\theta}^{S}_{i,j+1})^{l}\big) - \frac{1}{2}\big((\boldsymbol{\theta}^{N}_{i,j})^{l}\big). \tag{4.40}$$

Note that the diagonal matrices $\mathbf{\Pi}_{i+\frac{1}{2},j}$ and $\mathbf{\Pi}_{i,j+\frac{1}{2}}$ are positive semi-definite, since the minmod limiter satisfies $0 \leq \phi(\theta) \leq 1$. Consequently, the ES2 scheme is energy stable, completing the proof. □

5 Algorithmic details and pseudocode

We have presented energy conservative and energy stable schemes (EC, ES1, ES2) for the semi-discrete form (4.3), including the numerical construction and theoretical results. We now discuss details of the corresponding fully discrete algorithms. Complete algorithmic pseudocode is given in algorithm 1.

5.1 Desingularization

In the construction of numerical fluxes, the computation of $\boldsymbol{u}_{i,j}, \boldsymbol{v}_{i,j}$ requires $\big(\mathcal{P}(\boldsymbol{h}_{i,j})\big)^{-1}$, which is valid when $\mathcal{P}(\boldsymbol{h}_{i,j})$ is positive-definite. However, this matrix could be ill-conditioned in some situations. To mitigate error from ill-conditioned operations, we employ a desingularization procedure introduced in [33], whose stochastic variant was introduced in [11, 12, 13]. Suppose $\mathcal{P}(\boldsymbol{h}_{i,j})$ has the eigenvalue decomposition,

$$\mathcal{P}(\boldsymbol{h}_{i,j}) = \boldsymbol{Q}\mathbf{\Pi}\boldsymbol{Q}^{\top}, \tag{5.1}$$

where $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_K)$ is a diagonal positive matrix. Then the desingularization procedure approximates $\big(\mathcal{P}(\boldsymbol{h}_{i,j})\big)^{-1}\boldsymbol{q}_{i,j}$ by perturbing the eigenvalues π_k to regularize the velocities:

$$\boldsymbol{u}_{i,j} = \boldsymbol{Q}\widetilde{\mathbf{\Pi}}^{-1}\boldsymbol{Q}^{\top}\boldsymbol{q}^{x}_{i,j}, \qquad \widetilde{\mathbf{\Pi}} = \mathrm{diag}(\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K), \qquad \widetilde{\pi}_i = \frac{\sqrt{\pi_i^4 + \max\{\pi_i^4, \epsilon^4\}}}{\sqrt{2}\pi_i}, \tag{5.2}$$

$$\boldsymbol{v}_{i,j} = \boldsymbol{Q}\widetilde{\mathbf{\Pi}}^{-1}\boldsymbol{Q}^{\top}\boldsymbol{q}^{y}_{i,j}, \qquad \widetilde{\mathbf{\Pi}} = \mathrm{diag}(\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K), \qquad \widetilde{\pi}_i = \frac{\sqrt{\pi_i^4 + \max\{\pi_i^4, \epsilon^4\}}}{\sqrt{2}\pi_i}, \tag{5.3}$$

where ϵ is a small positive constant. To ensure the consistency of the scheme, the discharge variables must be recomputed when the desingularization procedure is activated,

$$\boldsymbol{q}^x_{i,j} \leftarrow \mathcal{P}(\boldsymbol{h}_{i,j})\boldsymbol{u}_{i,j}, \qquad\qquad \boldsymbol{q}^y_{i,j} \leftarrow \mathcal{P}(\boldsymbol{h}_{i,j})\boldsymbol{v}_{i,j}. \tag{5.4}$$

5.2 Hyperbolicity-preserving criterion

The hyperbolicity and existence of the entropy flux pair require that $\mathcal{P}(\widehat{h})$ is positive-definite, i.e., $\mathcal{P}(\widehat{h}) > 0$. This corresponds to the condition $\mathcal{P}(\boldsymbol{h}_{i,j}) > 0$ holding for each cell $\mathcal{I}_{i,j}$. While this may be true at the current time, extra conditions must be imposed that this is true at the next time step. A sufficient condition, [11, Theorem 3.2], is the following condition for every $n \in [N]$:

$$\widehat{h}_{i,j}(\xi_n) \coloneqq \sum_{k=1}^{K} (\boldsymbol{h}_{i,j})_k \phi_k(\xi_n) > 0, \qquad\qquad \boldsymbol{h}_{i,j} = \left((\boldsymbol{h}_{i,j})_1, \ldots, (\boldsymbol{h}_{i,j})_K\right)^\top, \tag{5.5}$$

where $\{\xi_n\}_{n=1}^N$ is a nodal set in $\mathbb{R}^d$ for a positive-weight quadrature rule with sufficient accuracy relative to the ξ-polynomial space $P = \operatorname{span}(\phi_1, \ldots, \phi_K)$. We enforce (5.5) by restricting the time step size in the time evolution procedure, as is done in [11, Lemma 4.2]. Suppose (5.5) is satisfied at the current time level, and the forward Euler method is applied. To enforce (5.5) in the next time level, we require

$$\Delta t < \lambda \coloneqq \min_{i,j} \min_{n\in[N]} \left| \frac{\widehat{h}_{i,j}(\xi_n)}{\frac{\widehat{\mathcal{F}}^h_{i+\frac{1}{2},j}(\xi_n) - \widehat{\mathcal{F}}^h_{i-\frac{1}{2},j}(\xi_n)}{\Delta x} + \frac{\widehat{\mathcal{G}}^h_{i,j+\frac{1}{2}}(\xi_n) - \widehat{\mathcal{G}}^h_{i,j-\frac{1}{2}}(\xi_n)}{\Delta y}} \right|, \tag{5.6}$$

where

$$\widehat{\mathcal{F}}^h_{i+\frac{1}{2},j}(\xi) \coloneqq \sum_{k=1}^{K} \left(\mathcal{F}^h_{i+\frac{1}{2},j}\right)_k \phi_k(\xi), \qquad \mathcal{F}_{i+\frac{1}{2},j} = \left((\mathcal{F}^h_{i+\frac{1}{2},j})^\top, (\mathcal{F}^{q^x}_{i+\frac{1}{2},j})^\top, (\mathcal{F}^{q^y}_{i+\frac{1}{2},j})^\top\right)^\top \in \mathbb{R}^{3K} \tag{5.7}$$

$$\widehat{\mathcal{G}}^h_{i,j+\frac{1}{2}}(\xi) \coloneqq \sum_{k=1}^{K} \left(\mathcal{G}^h_{i,j+\frac{1}{2}}\right)_k \phi_k(\xi), \qquad \mathcal{G}_{i,j+\frac{1}{2}} = \left((\mathcal{G}^h_{i,j+\frac{1}{2}})^\top, (\mathcal{G}^{q^x}_{i,j+\frac{1}{2}})^\top, (\mathcal{G}^{q^y}_{i,j+\frac{1}{2}})^\top\right)^\top \in \mathbb{R}^{3K}. \tag{5.8}$$

In addition to enforcing a global time step restriction across all cells for hyperbolicity, we must also ensure that Δt satisfies the wave speed CFL condition to account for the hyperbolic nature of the system, cf. [11, Equation (4.15)].

5.3 Adaptive time step size

The time step restriction for the hyperbolicity-preserving procedure (5.6) is derived by assuming a forward Euler time integration method. We can extend this to higher-order strong stability-preserving Runge-Kutta (SSP-RK) methods, which are convex combinations of forward Euler methods with multiple intermediate stages [27]. We employ the adaptive time-stepping procedure in [6], which allows Forward Euler time step restrictions to be applied to SSP-RK procedures. Computationally, the procedure computes updates of the time step restriction at each intermediate stage of the RK process.

5.4 Numerical imposition of boundary conditions

On a rectangular computational domain $[a, b] \times [c, d]$, let $Q_{i,j}$, $i, j \in [M]$, denote the (numerical) cell averages of any physical quantity on the mesh. Periodic boundary conditions are straightforward to implement,

$$Q_{i,0} = Q_{i,M}, \qquad Q_{i,1} = Q_{i,M+1}, \qquad Q_{0,j} = Q_{M,j}, \qquad Q_{1,j} = Q_{M+1,j}, \qquad \forall\ i, j \in [M]$$

Periodic boundary conditions can be unphysical, and one way to mitigate finite-domain effects is to impose "outflow" boundary conditions that emulate physical continuation of the domain beyond the

computational grid. A simple strategy for outflow boundary condition is to set ghost cell values through extrapolation from interior cells. The straightforward zeroth-order extrapolation uses a constant function for the extrapolation, i.e., sets as boundary conditions,

$$Q_{i,0} = Q_{i,1}, \qquad Q_{i,M+1} = Q_{i,M}, \qquad Q_{0,j} = Q_{1,j}, \qquad Q_{M+1,j} = Q_{M,j}, \qquad \forall\, i,j \in [M] \tag{5.9}$$

Higher-order extrapolation can improve accuracy, but zeroth-order extrapolation is often preferred for stability [39].

5.5 Measuring energy change through augmented energy

Definition 2 that labels schemes as energy conservative and energy stable relies on such properties holding in the *interior* of the computational domain, but does not guarantee energy conservation/stability if boundary fluxes work to increase energy inside the domain. In such cases, energy will increase through no fault of the scheme but instead as a natural physical evolution of the model. Such situations can occur even for standard procedures that implement outflow boundary conditions. We discuss this in more detail below.

Under periodic boundary conditions, summing the equality/inequality in the definition of the EC or ES scheme yields,

$$\sum_{i,j\in[M]} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) = 0, \qquad \text{or,} \qquad \sum_{i,j\in[M]} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) \le 0, \tag{5.10}$$

respectively. With outflow boundary conditions, the summation of the semi-discrete formula differs from the periodic case. For example, in a rectangular computational domain with outflow boundary conditions on all four sides, summing the energy change across all cells to enforce energy conservation or stability corresponds to the respective assertions,

$$\sum_{i,j\in[M]} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) = \frac{1}{\Delta x}\sum_{j\in[M]}(\mathcal{H}_{\frac{1}{2},j} - \mathcal{H}_{M+\frac{1}{2},j}) + \frac{1}{\Delta y}\sum_{i\in[M]}(\mathcal{K}_{i,\frac{1}{2}} - \mathcal{K}_{i,M+\frac{1}{2}}), \qquad \text{(EC)} \tag{5.11a}$$

$$\sum_{i,j\in[M]} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) \le \frac{1}{\Delta x}\sum_{j\in[M]}(\mathcal{H}_{\frac{1}{2},j} - \mathcal{H}_{M+\frac{1}{2},j}) + \frac{1}{\Delta y}\sum_{i\in[M]}(\mathcal{K}_{i,\frac{1}{2}} - \mathcal{K}_{i,M+\frac{1}{2}}), \qquad \text{(ES)} \tag{5.11b}$$

From these expressions, we see that since the boundary fluxes are imposed by a relatively arbitrary problem setup (outflow boundary conditions, Dirichlet boundary conditions, etc.), then we cannot expect the above equality or inequality to result in non-increasing energy, regardless of the scheme developed. This is a concern that arises in practice: Consider the SWE SG system in one spatial dimension (formally, from the two-dimensional formulation one can set all y-derivatives to 0 and make all quantities y-independent), with the following piecewise linear bottom topography B, water velocity u, and water surface w:

$$B(x) = \begin{cases} 0, & x < 0.5, \\ 0.1x - 0.05, & 0.5 \le x \le 1.5, \\ 0.1, & \text{otherwise}, \end{cases} \qquad u(x,0) = 0.3, \qquad w(x,0) = 1, \tag{5.12}$$

over the computational domain is $[0,2]$. We impose the one-dimensional version of the outflow boundary conditions (5.9) on both sides of the domain, and we integrate up to terminal time $T = 0.07$. The results are shown in Figure 1, and it is evident that the relative energy of the three schemes (EC, ES1, ES2) is increasing, see the bottom-left plot. I.e., this increase in relative energy is caused by the fact that this setup results in the total energy of the system increasing due to the outflow boundary conditions.

This increase in energy for the true solution makes it difficult to evaluate the effectiveness of the EC and ES schemes. An alternative measure of energy is the *augmented energy*, described below, which accounts for boundary effects on the total energy of the system. If H_1 and H_M denote the entropy

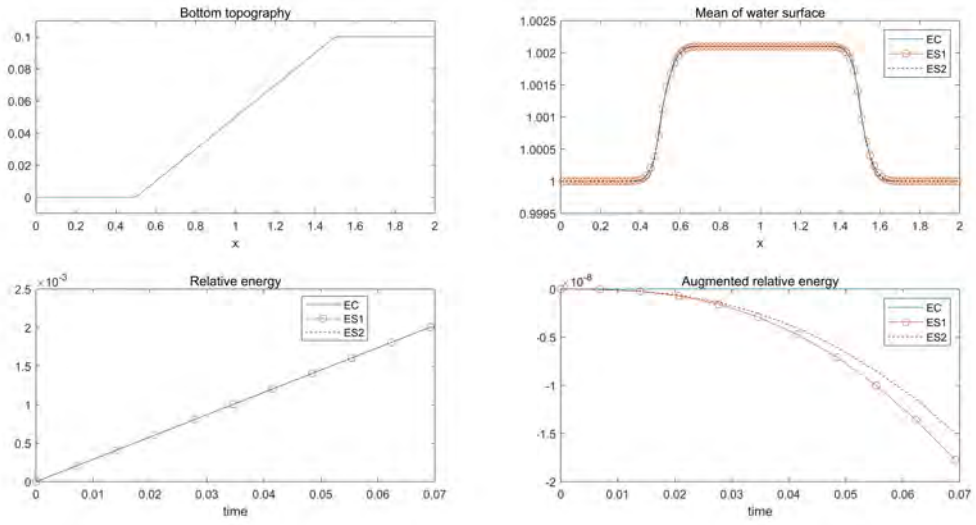

Figure 1: Illustration of energy versus augmented energy that incorporates energy fluxes at the boundary. Top left: Bottom topography B corresponding to (5.12). Top right: Mean of the water surface $w = h + B$ at terminal time $T = 0.07$. Bottom left: The standard relative energy $\frac{\boldsymbol{E}(t)-\boldsymbol{E}(0)}{\boldsymbol{E}(0)}$ increases in time for any accurate scheme due to influx of energy at the boundaries. Bottom right: Measuring the relative change in augmented energy $\frac{\widetilde{\boldsymbol{E}}(t)-\widetilde{\boldsymbol{E}}(0)}{\widetilde{\boldsymbol{E}}(0)}$ that offsets energy change due to boundary effects restores expected behavior for relative change of energy when using EC and ES schemes.

flux values at the one-dimensional boundary values, then we define the augmented energy $\widetilde{\boldsymbol{E}}$ over the computational domain to evolve as,

$$\frac{d}{dt}\widetilde{\boldsymbol{E}}(t) \coloneqq \sum_{i=1}^{M} \frac{d}{dt}\boldsymbol{E}_i(t) - \frac{1}{\Delta x}(H_1 - H_M), \qquad \widetilde{\boldsymbol{E}}(0) = \sum_{i=1}^{M} \boldsymbol{E}_i(0). \tag{5.13}$$

I.e., the augmented energy evolves according to the sum over all cells of the energy derivative, along with the energy change associated with the entropy boundary fluxes. The equation (5.13) can be time integrated since the right-hand side can be evaluated. By evaluating the relative augmented energy for the setup (5.12), shown in the bottom right plot of Figure 1, we observe that the EC scheme exhibits no (augmented) energy change, as expected, and the ES1 and ES2 schemes exhibit small energy decrease, with ES2 less dissipative than ES1, again as expected.

In this paper, for two-dimensional numerical examples, we generally implement outflow boundary conditions using zero-order extrapolation on the left and right boundaries and periodic boundary conditions on the upper and lower boundaries, resulting in the derivative of the augmented energy Our two-dimensional numerical examples in the next section largely use periodic boundary conditions on the upper and lower boundaries and zeroth-order outflow boundary conditions on the left and right boundaries, corresponding to the augmented energy,

$$\frac{d}{dt}\widetilde{\boldsymbol{E}}(t) \coloneqq \sum_{i,j=1}^{M} \frac{d}{dt}\boldsymbol{E}_{i,j}(t) - \sum_{j=1}^{M} \frac{1}{\Delta x}(H_{1,j} - H_{M,j}). \tag{5.14}$$

6 Numerical experiments

We will illustrate the performance of our EC and ES schemes through various examples, investigating convergence rates, energy conservation and decay properties, the well-balanced property, and the ability to handle multivariate random variables.

Algorithm 1 Pseudocode of the fully discrete EC/ES schemes for the 2D SG SWE

Input: Scheme type: *scheme* = EC, ES1, or ES2.
Input: Bottom topography B, initial data U at $t = 0$, terminal time T.
Initialize: Get PCE coefficients $\boldsymbol{U}_{i,j}$ with given polynomial index set Λ, set $t = 0$.

while $t < T$ **do**
  Compute $\boldsymbol{B}_{i,j}$ from B for all i, j.
  Compute $\boldsymbol{u}_{i,j}, \boldsymbol{v}_{i,j}$ for all i, j.
  Compute $\mathcal{F}^{EC}_{i+\frac{1}{2},j}, \mathcal{G}^{EC}_{i,j+\frac{1}{2}}$, and $\boldsymbol{S}_{i,j}$ for all i, j, by (4.12).
  if *scheme* is EC **then**
    $\mathcal{F}_{i+\frac{1}{2},j} \leftarrow \mathcal{F}^{EC}_{i+\frac{1}{2},j}, \mathcal{G}_{i,j+\frac{1}{2}} \leftarrow \mathcal{G}^{EC}_{i,j+\frac{1}{2}}$, for all i, j.
  else for all i, j
    Compute entropy variable $\boldsymbol{V}_{i,j}$ by (4.5).
    Compute $\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}, \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}$ for all i, j by (4.24a), (4.24b).
    if *scheme* is ES1 **then**
      Compute $\mathcal{F}_{i+\frac{1}{2},j} \leftarrow \mathcal{F}^{ES1}_{i+\frac{1}{2},j}, \mathcal{G}_{i,j+\frac{1}{2}} \leftarrow \mathcal{G}^{ES1}_{i,j+\frac{1}{2}}$ by (4.21) using $\mathcal{F}^{EC}_{i+\frac{1}{2},j}, \mathcal{G}^{EC}_{i,j+\frac{1}{2}}, \boldsymbol{V}_{i,j}$, and $\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}, \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}$.
    else if *scheme* is ES2 **then**
      Construct $\boldsymbol{V}^E_{i,j}, \boldsymbol{V}^W_{i,j}, \boldsymbol{V}^N_{i,j}, \boldsymbol{V}^S_{i,j}$ by (4.32), (4.33), and (4.37).
      Compute $\mathcal{F}_{i+\frac{1}{2},j} \leftarrow \mathcal{F}^{ES2}_{i+\frac{1}{2},j}, \mathcal{G}_{i,j+\frac{1}{2}} \leftarrow \mathcal{G}^{ES2}_{i,j+\frac{1}{2}}$
      by (4.31) using $\mathcal{F}^{EC}_{i+\frac{1}{2},j}, \mathcal{G}^{EC}_{i,j+\frac{1}{2}}, \boldsymbol{V}^E_{i,j}, \boldsymbol{V}^W_{i,j}, \boldsymbol{V}^N_{i,j}, \boldsymbol{V}^S_{i,j}$, and $\boldsymbol{Q}^{ES,F}_{i+\frac{1}{2},j}, \boldsymbol{Q}^{ES,G}_{i,j+\frac{1}{2}}$.
    end if
  end if
  Initialize λ and Δt by (5.6) and CFL condition.
  Determine Δt using the adaptive time step size procedure in Section 5.3.
  Use a SSP RK method with Δt determined adaptively, updating $\boldsymbol{h}_{i,j}, \boldsymbol{q}^x_{i,j}, \boldsymbol{q}^y_{i,j}$.
  $t \leftarrow t + \Delta t$
end while

Let ξ be a one-dimensional random variable from the Beta parametric family,

$$\xi \sim \text{Beta}(\beta+1, \alpha+1) \text{ on } [-1,1] \ : \ \rho(\xi) \propto (1-\xi)^{\alpha}(1+\xi)^{\beta}, \qquad \xi \in [-1,1], \tag{6.1}$$

Hence, the parameters $\alpha, \beta > -1$ can be chosen freely and control the mass concentration at $\xi = 1$ and $\xi = -1$, respectively. For this density ρ, the associated orthonormal polynomial basis is the family of Jacobi polynomials with parameters (α, β). In particular, $\alpha = \beta = 0$ corresponds to the uniform distribution $\mathcal{U}(-1,1)$, with Legendre polynomial basis functions. We will also consider a two-dimensional random variable $\xi = (\xi^{(1)}, \xi^{(2)})$ with two independent and identically distributed one-dimensional random variables $\xi^{(1)}, \xi^{(2)}$, i.e., $\xi^{(1)}, \xi^{(2)} \overset{\text{iid}}{\sim} \text{Beta}(\beta+1, \alpha+1)$. In this case the density is $\rho(\xi) \coloneqq \rho(\xi^{(1)})\rho(\xi^{(2)})$, and the orthonormal basis is obtained by tensorizing one-dimensional orthonormal polynomials.

For all two-dimensional examples, we implement outflow boundary conditions using zero-order extrapolation on the left and right boundaries and periodic boundary conditions on the upper and lower boundaries. To compare the energy conservation and decay properties of our EC and ES schemes, we will measure both the relative energy and the augmented relative energy change, with the latter introduced in section 5.5:

$$\text{relative change in (original) energy} = \frac{\boldsymbol{E}(t) - \boldsymbol{E}(0)}{\boldsymbol{E}(0)}, \tag{6.2}$$

$$\text{relative change in augmented energy} = \frac{\widetilde{\boldsymbol{E}}(t) - \widetilde{\boldsymbol{E}}(0)}{\widetilde{\boldsymbol{E}}(0)}, \tag{6.3}$$

where $\boldsymbol{E}(t) = \sum_{i,j} \boldsymbol{E}_{i,j}(t)$.

Throughout this section on numerical experiments, we assume the gravitational constant $g = 1$ and generally use $K = 4$ terms in the PCE procedure, except in the accuracy test and in the example involving a two-dimensional random variable. For the visualization of results, we will plot the water surface $w \coloneqq h + B$ instead of the conservative variable h corresponding to the water height. We generally apply the first- and second-order energy stable schemes (ES1, ES2) in the following numerical examples and further apply the second-order energy conservative scheme (EC) in the accuracy test.

6.1 Accuracy test for the energy conservative and the energy stable schemes

We begin by examining the order of accuracy of the proposed EC and ES schemes for the two-dimensional stochastic shallow water system. This test is a stochastic modification of the test originally developed in [4]. The initial conditions for the water surface and velocities are deterministic and defined as follows:

$$w(x,y,0,\xi) = 1, \qquad u(x,y,0,\xi) = 0.3, \qquad v(x,y,0,\xi) = 0, \tag{6.4}$$

where w represents the water surface, and u and v are the velocities in the x- and y- directions, respectively. Consider a stochastic elliptic-shaped hump bottom

$$B(x,y,\xi) = 0.5\exp(-25(x-1)^2 - 50(y-0.5)^2) + 0.1(\xi+1), \tag{6.5}$$

where $\xi \sim \mathcal{U}(-1,1)$. The computational domain is $[0,2] \times [0,1]$. A reference solution is computed on a 800×800 uniform rectangular grid for each scheme. For the time evolution solver, we utilize the third-order Strong Stability-Preserving (SSP) Runge-Kutta (RK3) method [27].

We illustrate the order of accuracy of our schemes using the water height h by computing the error between the reference solution and the numerical test solution with the L^1 norm in physical space and the L^2 norm in stochastic space.

$$\begin{aligned} \text{Error}(h_d) &= \|h_d(x,y,t,\xi) - h_{ref}(x,y,t,\xi)\|_{L^1(\mathcal{D};L^2_\rho(\mathbb{R}^d))}, \\ \|h(x,y,t,\xi)\|_{L^1(\mathcal{D};L^2_\rho(\mathbb{R}^d))} &\coloneqq \int_{\mathcal{D}} \|h(x,y,t,\xi)\|_{L^2_\rho} dxdy \end{aligned} \tag{6.6}$$

The contour plots of the reference solution for the mean of water surface and its standard deviation at $t = 0.07$ are shown in Figure 2.

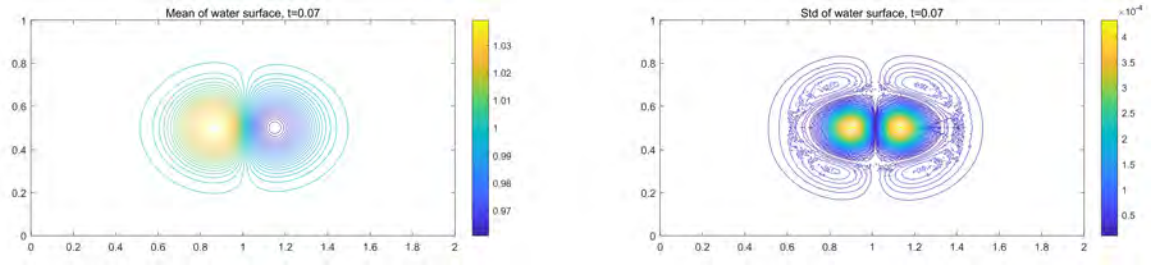

Figure 2: Results for section 6.1: Contours for the water surface of the reference solution of ES2. Left: mean. Right: standard deviation. $T = 0.07$.

We compute the orders of convergence for the numerical solutions with $K = 2, T = 0.07$, as summarized in Table 1. The results indicate that the EC scheme has the smallest error and the highest order of accuracy, while the ES1 scheme has the largest error and the lowest order of accuracy. All schemes yield results consistent with theoretical expectations. We avoid using larger values of K due to the occurrence of oscillations in stochastic space, which leads to instability in the higher modes of the PCE. This is a known general problem of the SG method when used for the hyperbolic conservation/balance laws problems, e.g. [41] and it will be part of our future research to address it.

Scheme	Grid size	Error	Order
ES1	100×100	2.1447e-04	
	200×200	7.3671e-05	1.5417
	400×400	2.2557e-05	1.7075
ES2	100×100	1.5434e-04	
	200×200	3.9852e-05	1.9534
	400×400	1.0528e-05	1.9203
EC	100×100	1.4880e-04	
	200×200	3.6890e-05	2.0121
	400×400	8.8995e-06	2.0514

Table 1: Accuracy test of section 6.1: Order of convergence for $K = 2, T = 0.07$. Reference solution computed with grid size 800×800.

We investigated the EC scheme in this accuracy test because it produces smooth solutions for short time evolution up to $T = 0.07$. However, in the following numerical examples, we avoid using the EC scheme, as it introduces unphysical numerical oscillations into non-smooth solutions since energy should be dissipated across discontinuities, similar to the observations reported for the models in 1D physical space [13].

6.2 Gaussian-shape hump with stochastic bottom

We consider the example with a random variable applied to the position of the hump in the bottom topography,

$$B(x, y, \xi) = 0.8 \exp(-5(x - 0.9 + 0.1\xi)^2 - 50(y - 0.5)^2), \qquad \xi \sim \mathcal{U}(-1, 1). \tag{6.7}$$

The discontinuous initial water surface is given by

$$w(x,y,0,\xi) = \begin{cases} 1.01, & 0.05 < x < 0.15, \\ 1, & \text{otherwise}, \end{cases} \tag{6.8}$$

with zero initial velocities $u(x,y,0) = v(x,y,0) = 0$.

The computational domain is $[0,2] \times [0,1]$. The test is a modification of the tests from [38, 4, 12]. We compute the numerical solutions of the ES schemes on a 200×200 mesh grid at times $t = 0.6, 0.9, 1.2, 1.5, 1.8$. The results include contour plots representing the mean water surface, accompanied by disk glyphs whose radius is proportional to the uncertainties at the corresponding cells. According to the plots in Figure 3, the water propagates to the right initially. After interacting with the hump, the water splits and propagates in all directions, generating more wave structures. The uncertainties are concentrated near the peak of the water surface until the water reaches the hump, after which they spread out along the wave structure. The results in Figure 3 and 4 show that both the ES1 and ES2 schemes dissipate energy and produce similar wave structures. In addition, the ES2 scheme achieves higher resolution and with less energy dissipation than the ES1 scheme which delivers a more smeared/diffuse solution, as theoretically expected.

6.3 Gaussian-shape hump with stochastic initial water surface

To demonstrate the robustness of our schemes, we consider a similar example of the Gaussian-shape hump with a stochastic initial water surface, as discussed in Section 6.2. Consider the deterministic bottom topography

$$B(x,y) = 0.8 \exp(-5(x-0.9)^2 - 50(y-0.5)^2), \tag{6.9}$$

and a stochastic initial water surface

$$w(x,y,0,\xi) = \begin{cases} 1 + 0.01(\xi+1), & 0.05 < x < 0.15, \\ 1, & \text{otherwise}, \end{cases} \tag{6.10}$$

where $\xi \sim \mathcal{U}(-1,1)$, and with zero initial velocities $u(x,y,0) = v(x,y,0) = 0$. We implement the same settings as described in Section 6.2 and present the results in Figures 5 and 6. The plots display similar contours of the mean water surface as those in the previous example in Section 6.2. Both the ES1 and ES2 schemes demonstrate energy dissipation; however, the ES2 scheme provides better resolution and less energy dissipation compared to the ES1 scheme. Additionally, the uncertainties are more uniformly distributed across the water surface in proportion to the mean water surface, as compared to the distribution of uncertainties in Section 6.2.

6.4 A submerged flat plateau

We consider the deterministic initial water surface with a small deterministic perturbation of size 10^{-4} as follows:

$$w(x,y,0,\xi) = \begin{cases} 1.0001, & -0.4 < x < -0.3, \\ 1, & \text{otherwise}, \end{cases} \qquad u(x,y,0) = v(x,y,0) = 0, \tag{6.11}$$

and with a stochastic bottom function

$$B(x,y) = \begin{cases} 0.9998, & r \le 0.1, \\ 9.997(0.2-r) + 0.0001(\xi+1), & 0.1 < r \le 0.2, \\ 0.0001, & \text{otherwise}, \end{cases} \tag{6.12}$$

where $r \coloneqq \sqrt{x^2+y^2} + 0.0001$ and $\xi \sim \mathcal{U}(-1,1)$.

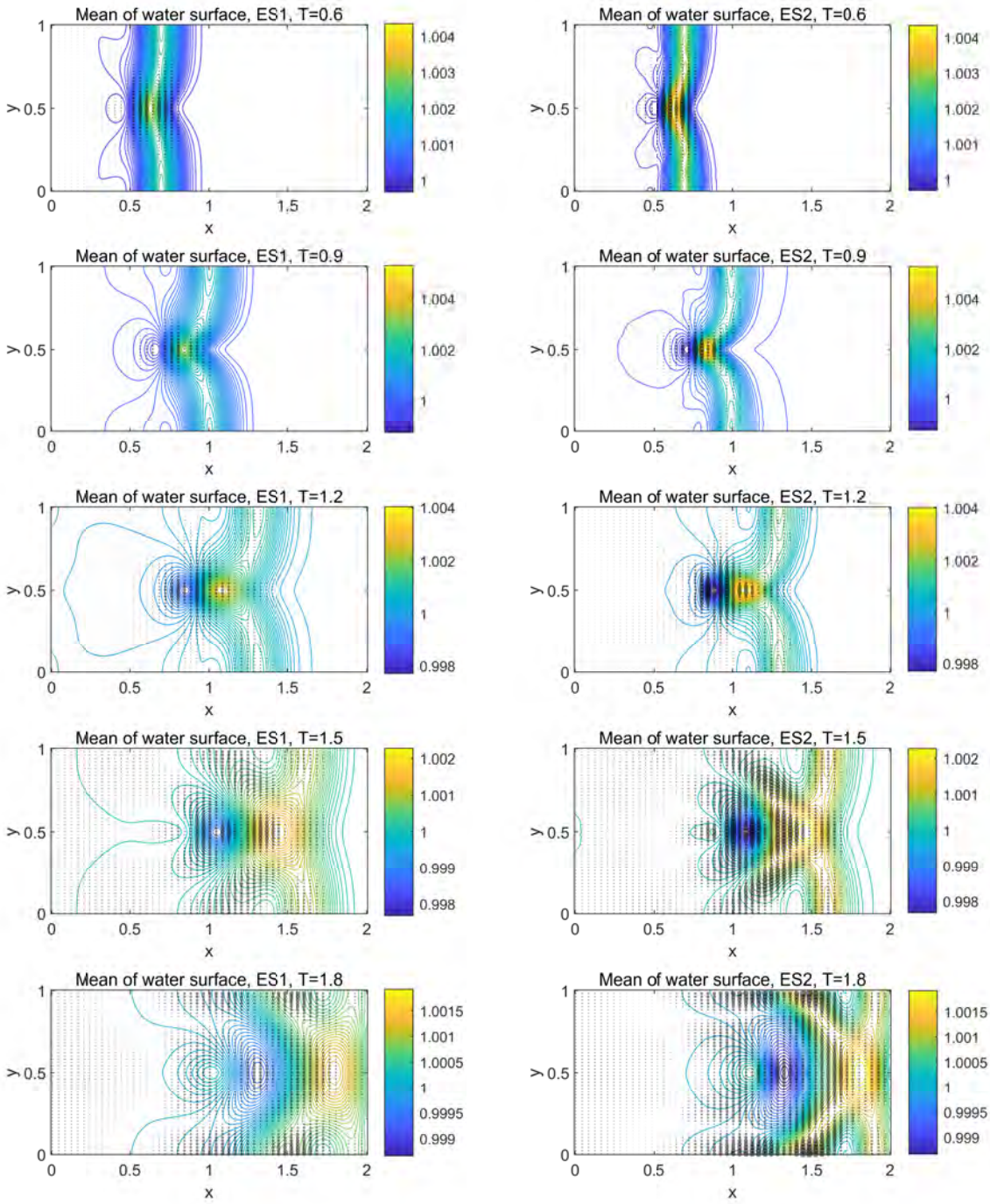

Figure 3: Results for section 6.2. Mean of water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 4.6524e-04, 7.3933e-04, 3.5168e-04, 1.3633e-04, 1.0856e-04, respectively. Right: ES2, the maximum standard deviation is 1.0398e-03, 2.1739e-03, 8.9851e-04, 3.4970e-04, 2.9730e-04, respectively. 200×200, $t = 0.6, 0.9, 1.2, 1.5, 1.8$.

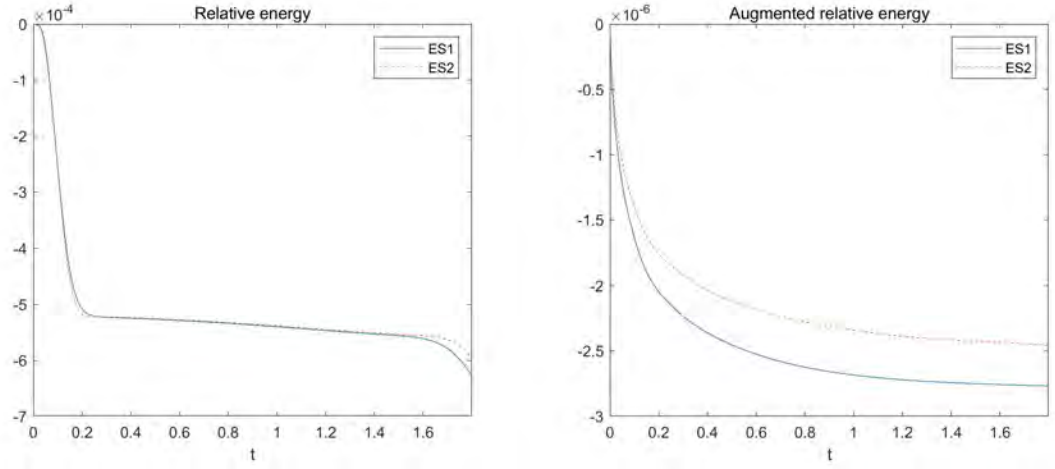

Figure 4: Results for section 6.2. Left: relative change in energy. Right: relative change in augmented energy.

The bottom is a flat plateau close to the water surface with a small perturbation of magnitude 10^{-4}. This topography is continuous but lacks smoothness in the physical domain. The test is a modification of the test from [4, 12]. We compute the solutions of the ES schemes over the computational domain $[-0.5, 0.5] \times [-0.5, 0.5]$ at times $t = 0.2, 0.35, 0.5, 0.65$. The initial deterministic perturbation generates two waves. The left-going wave travels out of the computational domain, while the right-going wave interacts with the plateau, leading to complex wave patterns as the interaction progresses. The results in Figure 7 demonstrate that our schemes effectively capture the stochastic lake-at-rest solution, indicating that they are well-balanced. Both the ES1 and ES2 schemes dissipate energy, with the ES2 scheme providing higher resolution and less energy dissipation compared to the ES1 scheme. Note that the uncertainties of the ES1 scheme are not visible, since the ES1 standard deviation is much smaller than that of the ES2 scheme.

The initial deterministic perturbation on the water surface is 10^{-4}. Throughout the time evolution, the maximum mean water surface displacement from the steady state remains within this initial perturbation threshold, and the standard deviation remains small. These results demonstrate the capability of both the ES1 and ES2 schemes to accurately capture the lake-at-rest solution under small perturbations.

For comparison, we also present a non-well-balanced ES2 scheme by modifying the numerical source term to

$$\boldsymbol{S}_{i,j} = \begin{pmatrix} 0 \\ -\frac{g}{2\Delta x}\mathcal{P}(\overline{\boldsymbol{h}}_{i,j})\big(\llbracket \boldsymbol{B} \rrbracket_{i+\frac{1}{2},j} + \llbracket \boldsymbol{B} \rrbracket_{i-\frac{1}{2},j}\big) \\ -\frac{g}{2\Delta y}\mathcal{P}(\overline{\boldsymbol{h}}_{i,j})\big(\llbracket \boldsymbol{B} \rrbracket_{i,j+\frac{1}{2}} + \llbracket \boldsymbol{B} \rrbracket_{i,j-\frac{1}{2}}\big) \end{pmatrix}, \tag{6.13}$$

corresponding to a central differencing scheme for the derivative of bottom topography.

We can observe from Figure 9 that the wave structure produced by the non-well-balanced ES2 scheme differs significantly from that of the well-balanced schemes in Figure 7. The non-well-balanced scheme generates substantially larger perturbations compared to the well-balanced solutions. Note that even with mesh refinement, the non-well-balanced scheme cannot produce physically accurate wave structure, and the perturbations remain larger than the initial value throughout the time evolution.

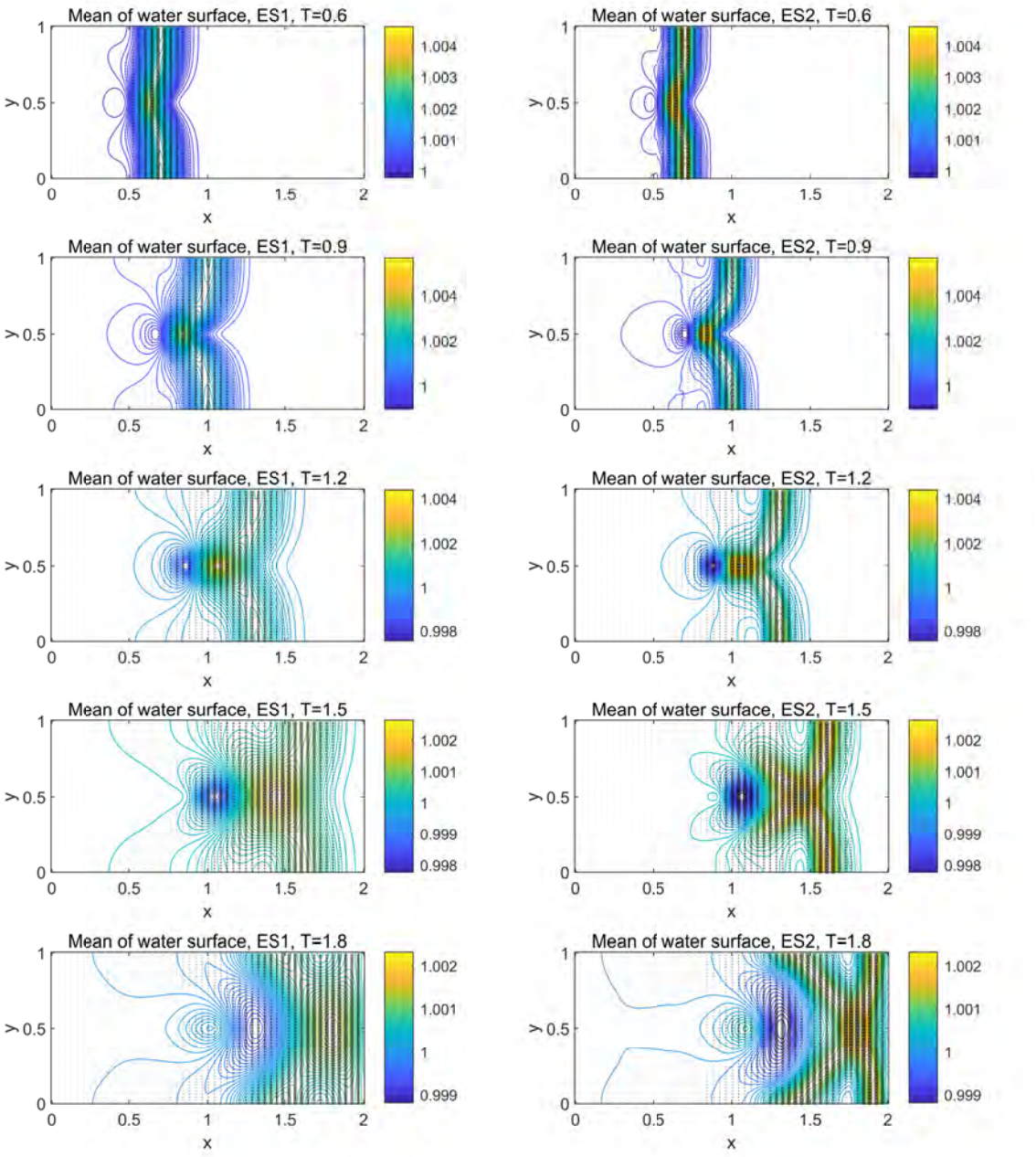

Figure 5: Results for section 6.3. Mean of water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 1.8216e-03, 1.9375e-03, 1.5528e-03, 1.0162e-03, 8.2437e-04, respectively. Right: ES2, the maximum standard deviation is 2.6787e-03, 3.2736e-03, 2.5036e-03, 1.5027e-03, 1.3321e-03, respectively. 200×200, $t = 0.6, 0.9, 1.2, 1.5, 1.8$.

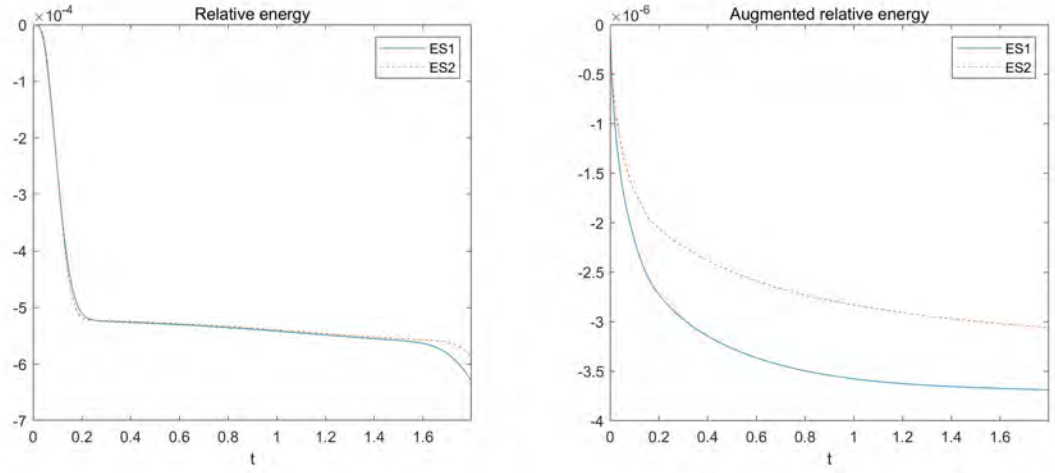

Figure 6: Results for section 6.3. Left: relative change in energy. Right: relative change in augmented energy.

6.5 Perturbation to lake-at-rest

We consider another example of a perturbation to lake-at-rest solution with a stochastic initial water surface

$$w(x,y,0,\xi) = \begin{cases} 1 + 0.001(\xi+1), & |xy| \le 0.05, \\ 1, & \text{otherwise}, \end{cases} \qquad u(x,y,0) = v(x,y,0) = 0, \tag{6.14}$$

and with the deterministic bottom topography

$$B(x,y) = \begin{cases} 0.25(\cos(5\pi(xy+0.35))+1), & -0.55 < xy < -0.15, \\ 0.125(\cos(10\pi(xy-0.35))+1), & 0.25 < xy < 0.45, \\ 0, & \text{otherwise}. \end{cases} \tag{6.15a}$$

The computational domain is $[-1,1]\times[-1,1]$ and $\xi \sim \mathcal{U}(-1,1)$. The test is a 2D modification of the test from [8, 13].

The results presented in Figure 12 indicate that the water propagates towards the four vertices of the rectangular computational domain. The wave moves slower after reaching the raised bottom. As in previous examples, both ES1 and ES2 schemes are well-balanced and produce a similar wave structure, with the ES2 scheme creating a more accurate solution with a higher resolution and less diffusion. Furthermore, uncertainties are distributed in proportion to the mean water surface in the ES schemes, as they originate from the initial water surface rather than the bottom topography.

To investigate the impact of the bottom topography bump on the results, we slightly modify the bottom topography to

$$B(x,y) = \begin{cases} 0.45(\cos(5\pi(xy+0.35))+1), & -0.55 < xy < -0.15 \\ 0.125(\cos(10\pi(xy-0.35))+1), & 0.25 < xy < 0.45 \\ 0, & \text{otherwise}, \end{cases} \tag{6.15b}$$

and plot results for the ES1 and ES2 schemes in fig. 13. Comparing the solutions in Figure 13 with those in Figure 12 at $T = 0.6$ and $T = 0.8$, one can observe the significant influence of the bottom topography on the wave structure.

For comparison, we also simulate results using an ES2 scheme that employs the non-well-balanced source term discretization in (6.13). From the results shown in Figure 14, the non-well-balanced ES2 scheme produces a different wave structure and generates a larger displacement of the water surface

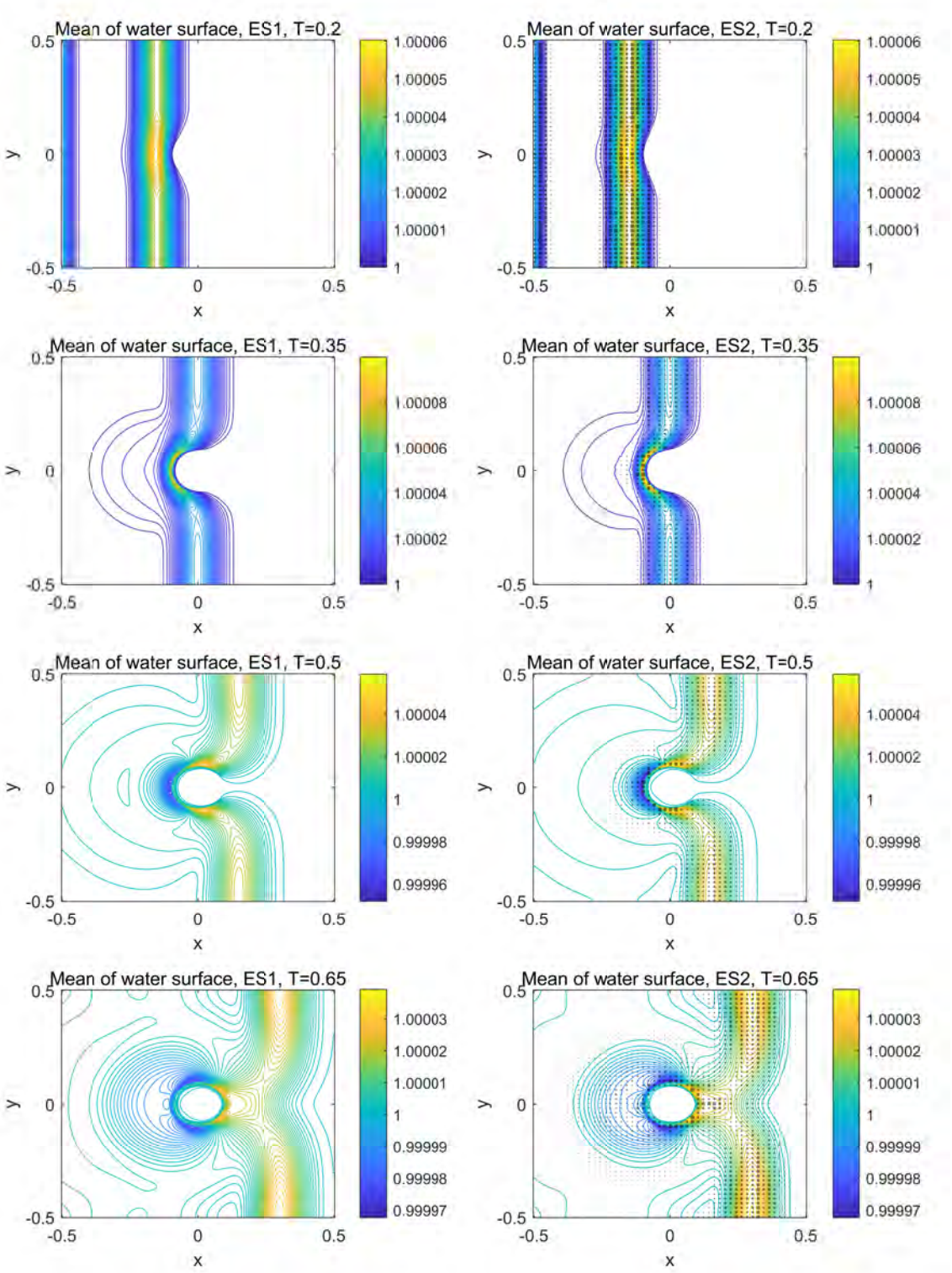

Figure 7: Results for section 6.4. Mean of water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 1.7485e-08, 1.9120e-07, 1.3268e-07, 1.0392e-07, respectively. Right: ES2, the maximum standard deviation is 6.6484e-07, 1.8990e-06, 1.4625e-06, 8.1986e-07, respectively. 200×200, $t = 0.2, 0.35, 0.5, 0.65$.

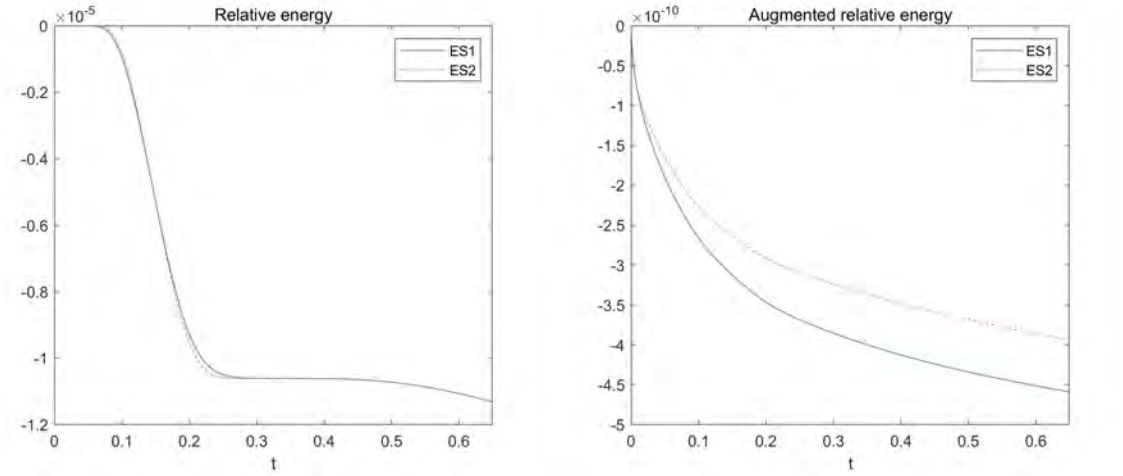

Figure 8: Results of section 6.4. Left: relative change in energy. Right: relative change in augmented energy.

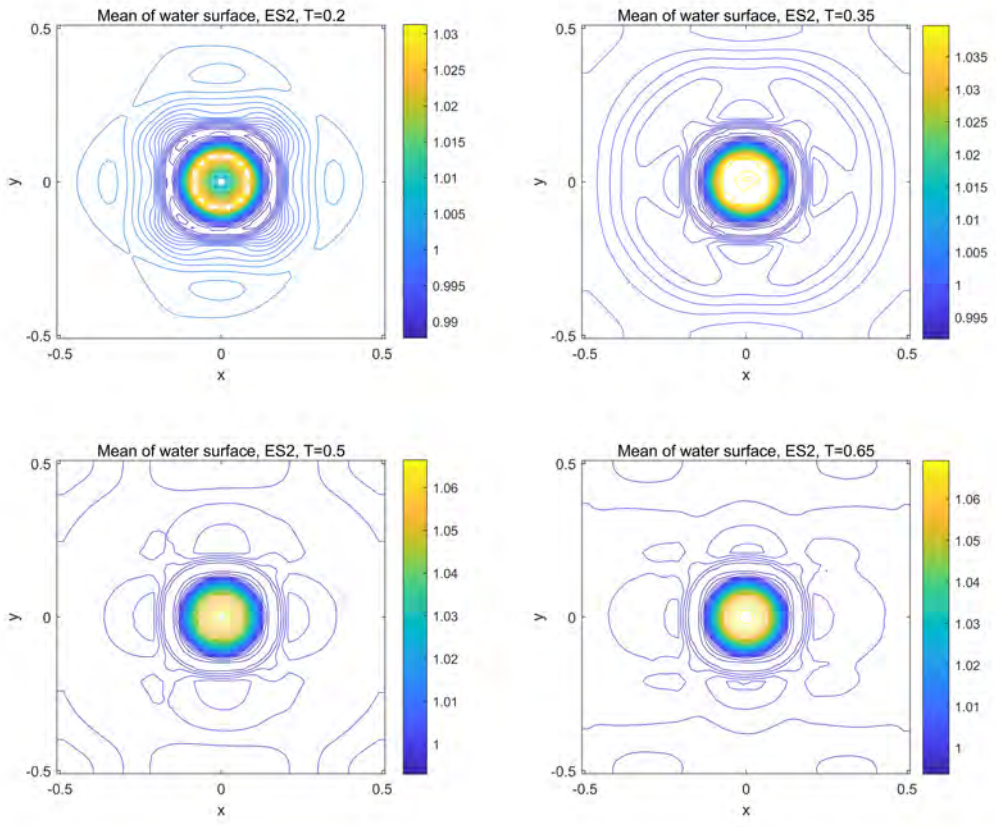

Figure 9: Results for section 6.4. Mean of water surface. Non-well-balanced ES2 scheme. 50×50, $t = 0.2, 0.35, 0.5, 0.65$.

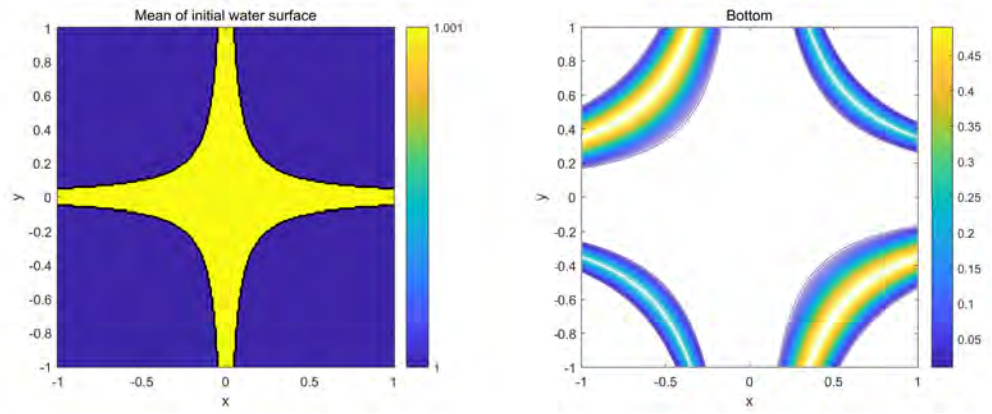

Figure 10: Results for section 6.5. Left: initial water surface. Right: bottom topography. 200×200

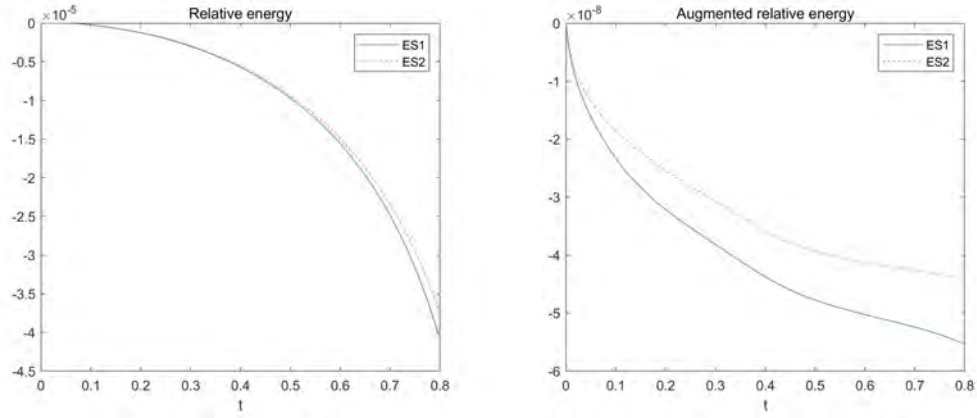

Figure 11: Results for section 6.5. Left: relative change in energy. Right: relative change in augmented energy. 200×200

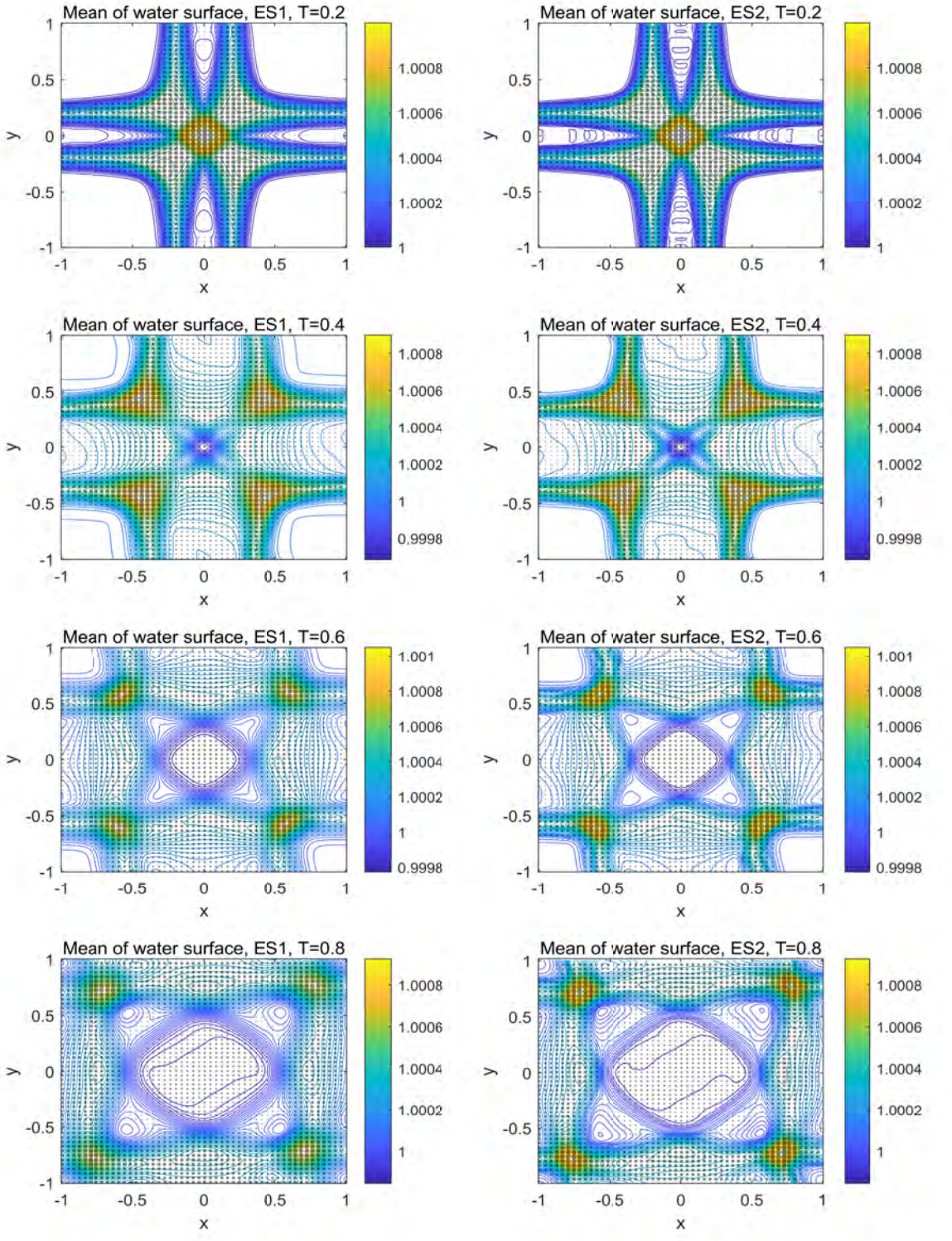

Figure 12: Results for section 6.5 with bottom topography B given by (6.15a). Contour plots for the water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 5.7522e-04, 4.6740e-04, 4.8587e-04, 3.9092e-04. Right: ES2, the maximum standard deviation is 5.7687e-04, 5.0798e-04, 5.8621e-04, 5.1600e-04. The mesh is 200×200 with snapshots in time shown at $T = 0.2, 0.4, 0.6, 0.8$.

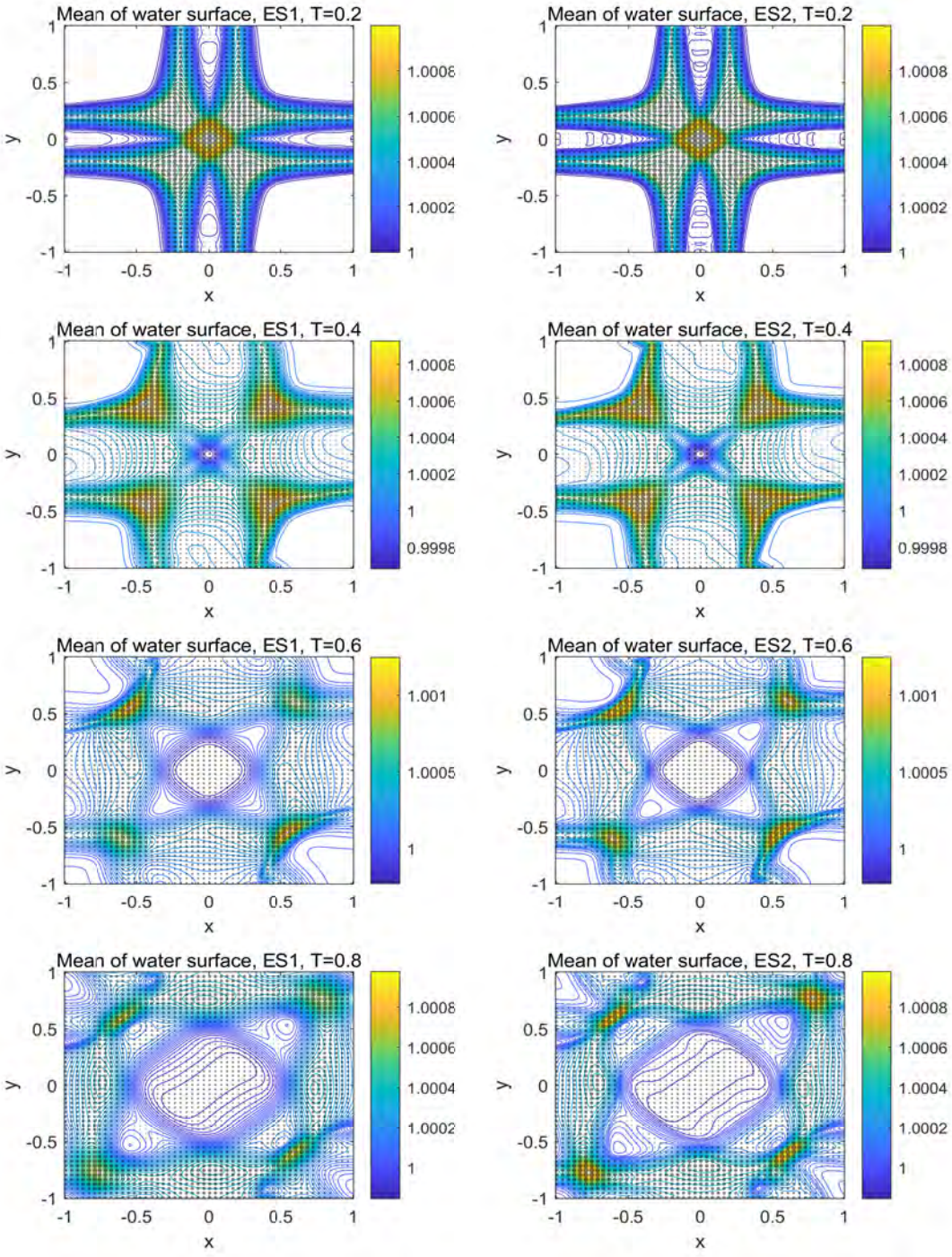

Figure 13: Results for section 6.5 with bottom topography B given by (6.15b). Contour plots for the water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 5.7522e-04, 4.8313e-04, 5.9729e-04, 4.0329e-04. Right: ES2, the maximum standard deviation is 5.7688e-04, 5.2388e-04, 7.0756e-04, 5.4857e-04. The mesh is 200×200 with snapshots in time shown at $T = 0.2, 0.4, 0.6, 0.8$.

compared to the well-balanced ES1 and ES2 schemes. Again, note that refining the mesh does not produce the accurate wave structure, and the perturbations during the time evolution are still larger than the initial perturbation for the non-well-balanced scheme.

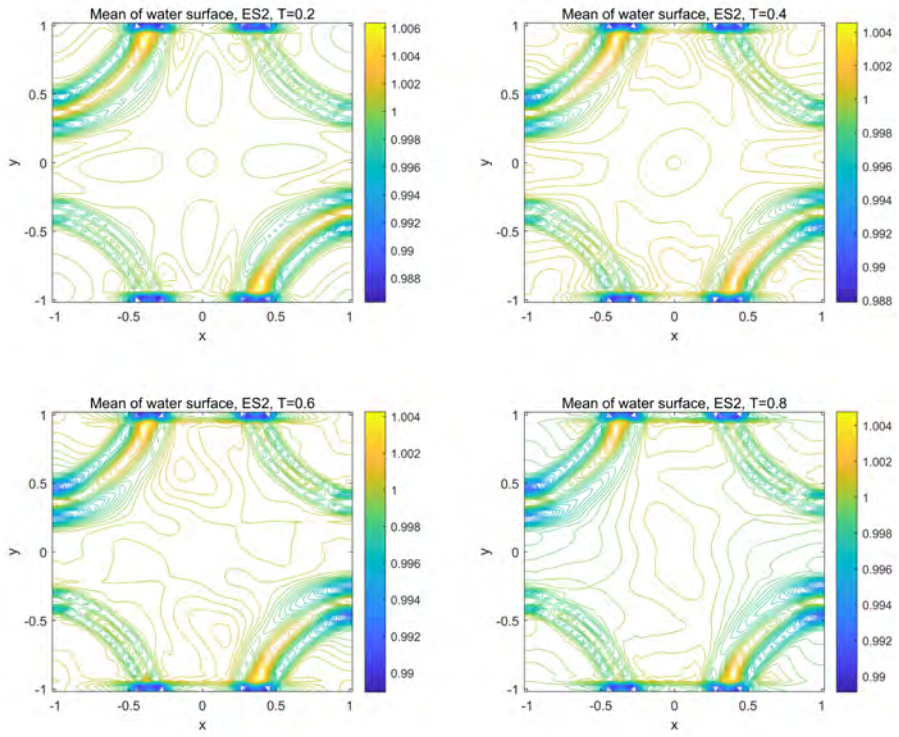

Figure 14: Results for section 6.5. Mean of water surface. Non-well-balanced ES2 scheme. 50×50, $t = 0.2, 0.4, 0.6, 0.8$.

6.6 Gaussian-shape hump with two-dimensional random variable

As a final test, we consider a two-dimensional in the stochastic space variant of the example used for the accuracy test provided in Section 6.1, with a deterministic initial water surface

$$w(x, y, 0, \xi) = 1, \qquad u(x, y, 0, \xi) = 0.3, \qquad v(x, y, 0, \xi) = 0, \tag{6.16}$$

and a stochastic bottom

$$B(x, y, \xi) = 0.5 \exp(-12.5(\xi_1 + 1)(x - 1)^2 - 25(\xi_2 + 1)(y - 0.5)^2). \tag{6.17}$$

Here $\xi = (\xi_1, \xi_2)$ is a two-dimensional random vector with independent components, where ξ_1 has a Beta$(\beta + 1, \alpha + 1)$ distribution on $[-1, 1]$ with parameters $(\alpha, \beta) = (3, 1)$, and ξ_2 is uniformly distributed on $[-1, 1]$. The uncertainties are in the width of the Gaussian-shape hump. To simulate the SG SWE system, we use $K_1 = K_2 = 3$ polynomial terms for the parameters ξ_1 and ξ_2, resulting in a tensor-product space of polynomials with dimension $K = K_1 K_2 = 9$. We compute solutions for $(x, y) \in [0, 2] \times [0, 1]$ up to time $T = 0.07$. Due to the effect of the two-dimensional random variable, the solution in Figure 15 has a more complex structure compared to the structure of the solution in the test in Figure 2. The largest uncertainties are observed near the peak and bottom of the mean water surface, while the smallest uncertainties occur in the region between them. The energy plots in Figure 16 demonstrate that measuring augmented energy is necessary to account for energy-related boundary effects in this example.

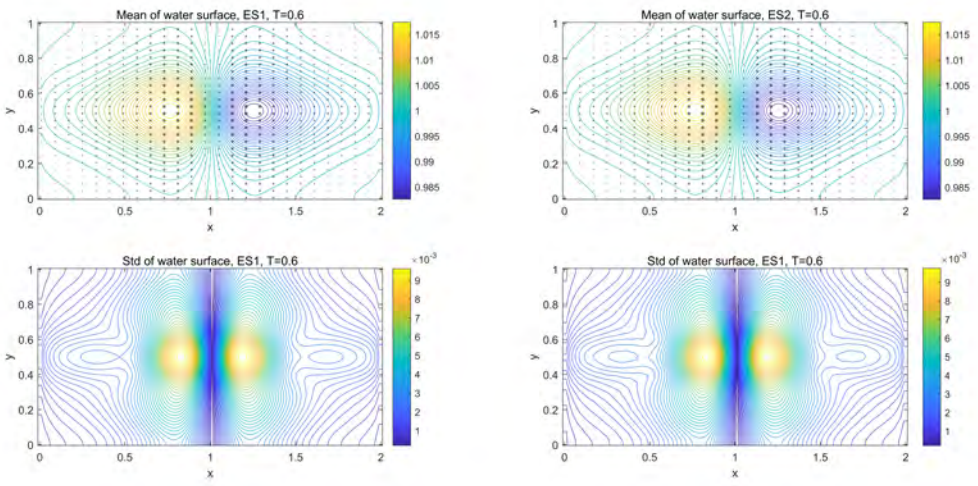

Figure 15: Results for section 6.6. Contour plots for the water surface. Disk-glyph over mean contours, where the radii of the disks indicate the magnitude of the standard deviation. Left: ES1, the maximum standard deviation is 9.8031e-03. Right: ES2, the maximum standard deviation is 9.9456e-03. $100 \times 100, T = 0.07$.

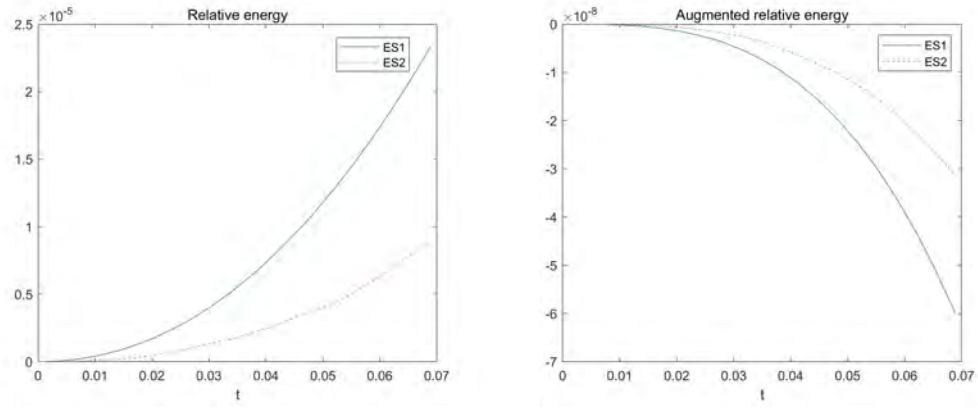

Figure 16: Results for section 6.6. Left: relative change in energy. Right: relative change in augmented energy. 100×100.

7 Conclusion

In this work we derived an entropy flux pair for the two-dimensional hyperbolicity-preserving and positivity-preserving SG SWE system developed in [12]. The entropy flux pair facilitates the formulation of entropy admissibility criteria to obtain the desired physical solution among non-unique weak solutions. Using this entropy pair, we constructed second-order energy conservative and first- and second-order energy stable finite volume schemes for the two-spatial-dimensional SG SWE, with the well-balanced property. We presented several numerical experiments demonstrating the efficiency and robustness of our schemes. Currently, our schemes are limited to second-order spatial accuracy; however, they can be extended to higher orders by developing higher-order energy conservative finite volume schemes, more accurate diffusion operators, or employing discontinuous Galerkin methods. Another possibility for future research is to extend the developed framework to models associated with dry/wet interfaces.

CRediT authorship contribution statement

Yekaterina Epshteyn: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Akil Narayan:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Yinqian Yu:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work of Yekaterina Epshteyn, Akil Narayan and Yinqian Yu was partially supported by NSF DMS-2207207. AN was also partially supported by NSF DMS-1848508.

References

[1] Ivo Babuska, Raúl Tempone, and Georgios E Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 42(2):800–825, 2004.

[2] Sylvie Benzoni-Gavage and Denis Serre. *Multi-dimensional Hyperbolic Partial Differential Equations: First-order Systems and Applications*. Oxford University Press, 2007.

[3] Alfredo Bermudez and Ma Elena Vazquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[4] Bryson, Steve, Epshteyn, Yekaterina, Kurganov, Alexander, and Petrova, Guergana. Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM: M2AN*, 45(3):423–446, 2011.

[5] Chen Chen, Clint Dawson, and Eirik Valseth. Cross-mode stabilized stochastic shallow water systems using stochastic finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 405:115873, 2023.

[6] Alina Chertock, Shumo Cui, Alexander Kurganov, and Tong Wu. Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *International Journal for Numerical Methods in Fluids*, 78(6):355–383, 2015.

[7] Alina Chertock, Shi Jin, and Alexander Kurganov. An operator splitting based stochastic Galerkin method for the one-dimensional compressible Euler equations with uncertainty. *Preprint*, pages 1–21, 2015.

[8] Alina Chertock, Shi Jin, and Alexander Kurganov. A well-balanced operator splitting based stochastic Galerkin method for the one-dimensional Saint-Venant system with uncertainty. *Preprint*, 2015.

[9] Nelida Črnjarić-Žic, Senka Vuković, and Luka Sopta. Balanced finite volume WENO and central WENO schemes for the shallow water and the open-channel flow equations. *Journal of Computational Physics*, 200(2):512–548, 2004.

[10] Constantine M Dafermos. *Hyperbolic Conservation Laws in Continuum Physics*. Springer, 2016.

[11] Dihan Dai, Yekaterina Epshteyn, and Akil Narayan. Hyperbolicity-preserving and well-balanced stochastic Galerkin method for shallow water equations. *SIAM Journal on Scientific Computing*, 43(2):A929–A952, 2021.

[12] Dihan Dai, Yekaterina Epshteyn, and Akil Narayan. Hyperbolicity-preserving and well-balanced stochastic Galerkin method for two-dimensional shallow water equations. *Journal of Computational Physics*, 452:110901, 2022.

[13] Dihan Dai, Yekaterina Epshteyn, and Akil Narayan. Energy stable and structure-preserving schemes for the stochastic Galerkin shallow water equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 58(2):723–757, 2024.

[14] B De St Venant. Théorie du mouvement non-permanent des eaux avec application aux crues des rivières et à l'introduntion des Marees dans leur lit. *Academic de Sci. Comptes Redus*, 73:148–154, 237–240, 1871.

[15] Bert J Debusschere, Habib N Najm, Philippe P Pébay, Omar M Knio, Roger G Ghanem, and Olivier P Le Maître. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM Journal on Scientific Computing*, 26(2):698–719, 2004.

[16] Bruno Després, Gaël Poëtte, and Didier Lucor. Robust uncertainty propagation in systems of conservation laws with the entropy closure method. In *Uncertainty Quantification in Computational Fluid Dynamics*, pages 105–149. Springer, 2013.

[17] Martin Eigel, Claude Jeffrey Gittelson, Christoph Schwab, and Elmar Zander. Adaptive stochastic Galerkin FEM. *Computer Methods in Applied Mechanics and Engineering*, 270:247–269, 2014.

[18] Yekaterina Epshteyn and Thuong Nguyen. Adaptive central-upwind scheme on triangular grids for the Saint–Venant system. *Communications in Mathematical Sciences*, 21(3):671–708, 2023.

[19] Oliver G Ernst, Antje Mugler, Hans-Jörg Starkloff, and Elisabeth Ullmann. On the convergence of generalized polynomial chaos expansions. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(2):317–339, 2012.

[20] Ulrik Fjordholm, Siddhartha Mishra, and Eitan Tadmor. Energy preserving and energy stable schemes for the shallow water equations. *London Mathematical Society Lecture Note Series*, 1(363):93–139, 2009.

[21] Ulrik S. Fjordholm, Siddhartha Mishra, and Eitan Tadmor. Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *Journal of Computational Physics*, 230(14):5587–5609, 2011.

[22] Ulrik S. Fjordholm, Siddhartha Mishra, and Eitan Tadmor. Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. *SIAM Journal on Numerical Analysis*, 50(2):544–573, 2012.

[23] Stephan Gerster and Michael Herty. Entropies and symmetrization of hyperbolic stochastic Galerkin formulations. *Communications in Computational Physics*, 27:639–671, 2020.

[24] Stephan Gerster, Michael Herty, and Aleksey Sikstel. Hyperbolic stochastic Galerkin formulation for the p-system. *Journal of Computational Physics*, 395:186–204, 2019.

[25] Stephan Gerster, Aleksey Sikstel, and Giuseppe Visconti. Haar-type stochastic Galerkin formulations for hyperbolic systems with Lipschitz continuous flux function. *arXiv preprint arXiv:2203.11718*, 2022.

[26] RG Ghanem. Stochastic finite elements: A spectral approach. *Courier Corporation*, 1991.

[27] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[28] Jingwei Hu and Shi Jin. A stochastic Galerkin method for the Boltzmann equation with uncertainty. *Journal of Computational Physics*, 315:150–168, 2016.

[29] Shi Jin and Ruiwen Shu. A study of hyperbolicity of kinetic stochastic Galerkin system for the isentropic Euler equations with uncertainty. *Chinese Annals of Mathematics, Series B*, 40(5):765–780, 2019.

[30] Shi Jin, Dongbin Xiu, and Xueyu Zhu. A well-balanced stochastic Galerkin method for scalar hyperbolic balance laws with random inputs. *Journal of Scientific Computing*, 67:1198–1218, 2016.

[31] Alexander Kurganov. Finite-volume schemes for shallow-water equations. *Acta Numerica*, 27:289–351, 2018.

[32] Alexander Kurganov and Doron Levy. Central-upwind schemes for the Saint-Venant system. *ESAIM: Mathematical Modelling and Numerical Analysis*, 36(3):397–425, 2002.

[33] Alexander Kurganov and Guergana Petrova. A Second-Order Well-Balanced Positivity Preserving Central-Upwind Scheme for the Saint-Venant System. *Communications in Mathematical Sciences*, 5(1):133 – 160, 2007.

[34] Jonas Kusch, Ryan G McClarren, and Martin Frank. Filtered stochastic Galerkin methods for hyperbolic equations. *Journal of Computational Physics*, 403:109073, 2020.

[35] Olivier Le Maître and Omar M Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Springer Science & Business Media, 2010.

[36] OP Le Maître, Omar M Knio, Habib N Najm, and Roger G Ghanem. Uncertainty propagation using Wiener–Haar expansions. *Journal of Computational Physics*, 197(1):28–57, 2004.

[37] Randall J LeVeque. *Numerical methods for conservation laws*, volume 214. Springer, 1992.

[38] Randall J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm. *Journal of Computational Physics*, 146(1):346–365, 1998.

[39] Randall J LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge University Press, 2002.

[40] Xin Liu, Jason Albright, Yekaterina Epshteyn, and Alexander Kurganov. Well-balanced positivity preserving central-upwind scheme with a novel wet/dry reconstruction on triangular grids for the Saint-Venant system. *Journal of Computational Physics*, 374:213–236, 2018.

[41] Fabian Meyer, Christian Rohde, and Jan Giesselmann. A posteriori error analysis for random scalar conservation laws using the stochastic Galerkin method. *IMA Journal of Numerical Analysis*, 40(2):1094–1121, 02 2019.

[42] Siddhartha Mishra, Ch Schwab, and Jonas Sukys. Multilevel Monte Carlo finite volume methods for shallow water equations with uncertain topography in multi-dimensions. *SIAM Journal on Scientific Computing*, 34(6):B761–B784, 2012.

[43] Fabio Nobile, Raúl Tempone, and Clayton G Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

[44] Gaël Poëtte. Contribution to the mathematical and numerical analysis of uncertain systems of conservation laws and of the linear and nonlinear Boltzmann equation. *Ph.D. thesis, Université de Bordeaux*, 2019.

[45] Gaël Poëtte, Bruno Després, and Didier Lucor. Uncertainty quantification for systems of conservation laws. *Journal of Computational Physics*, 228(7):2443–2467, 2009.

[46] Roland Pulch and Dongbin Xiu. Generalised polynomial chaos for a class of linear conservation laws. *Journal of Scientific Computing*, 51:293–312, 2012.

[47] Benedict D Rogers, Alistair GL Borthwick, and Paul H Taylor. Mathematical balancing of flux gradient and source terms prior to using Roe's approximate Riemann solver. *Journal of Computational Physics*, 192(2):422–451, 2003.

[48] Louisa Schlachter and Florian Schneider. A hyperbolicity-preserving stochastic Galerkin approximation for uncertain hyperbolic systems of equations. *Journal of Computational Physics*, 375:80–98, 2018.

[49] Ruiwen Shu, Jingwei Hu, and Shi Jin. A stochastic Galerkin method for the Boltzmann equation with multi-dimensional random inputs using sparse wavelet bases. *Numerical Mathematics: Theory, Methods and Applications*, 10(2):465–488, 2017.

[50] Timothy John Sullivan. *Introduction to Uncertainty Quantification*, volume 63. Springer, 2015.

[51] Eitan Tadmor. The numerical viscosity of entropy stable schemes for systems of conservation laws. I. *Mathematics of Computation*, 49(179):91–103, 1987.

[52] Eitan Tadmor. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numerica*, 12:451–512, 2003.

[53] Julie Tryoen, Olivier Le Maître, Michael Ndjinga, and Alexandre Ern. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *Journal of Computational Physics*, 229(18):6485–6511, 2010.

[54] Xiaoliang Wan and George Em Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 209(2):617–642, 2005.

[55] Norbert Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

[56] Kailiang Wu, Huazhong Tang, and Dongbin Xiu. A stochastic Galerkin method for first-order quasilinear hyperbolic systems with uncertainty. *Journal of Computational Physics*, 345:224–244, 2017.

[57] Y. Xing. Chapter 13 - Numerical methods for the nonlinear shallow water equations. In Rémi Abgrall and Chi-Wang Shu, editors, *Handbook of Numerical Methods for Hyperbolic Problems*, volume 18 of *Handbook of Numerical Analysis*, pages 361–384. Elsevier, 2017.

[58] Yulong Xing. High order finite volume WENO schemes for the shallow water flows through channels with irregular geometry. *Journal of Computational and Applied Mathematics*, 299:229–244, 2016.

[59] Yulong Xing and Chi-Wang Shu. High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *Journal of Computational Physics*, 208(1):206–227, 2005.

[60] Yulong Xing and Chi-Wang Shu. High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Journal of Computational Physics*, 214(2):567–598, 2006.

[61] Yulong Xing and Chi-Wang Shu. A new approach of high order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Comput. Phys*, 1(1):100–134, 2006.

[62] Dongbin Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

[63] Dongbin Xiu and Jan S Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3):1118–1139, 2005.

[64] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

[65] Dongbin Xiu and Jie Shen. Efficient stochastic Galerkin methods for random diffusion equations. *Journal of Computational Physics*, 228(2):266–281, 2009.

[66] Xinghui Zhong and Chi-Wang Shu. Entropy stable Galerkin methods with suitable quadrature rules for hyperbolic systems with random inputs. *Journal of Scientific Computing*, 92(1):14, 2022.

[67] Jian G Zhou, Derek M Causon, Clive G Mingham, and David M Ingram. The surface gradient method for the treatment of source terms in the shallow-water equations. *Journal of Computational Physics*, 168(1):1–25, 2001.

[68] Tao Zhou and Tao Tang. Galerkin methods for stochastic hyperbolic problems using bi-orthogonal polynomials. *Journal of Scientific Computing*, 51:274–292, 2012.

A Proof of theorem 2

Theorem 2 is a direct result of convexity of the function $E(\widehat{U})$, proven in appendix A.2, and also of the fact that (E, H, K) satisfy the companion balance law, proven in appendix A.3. The proofs of these results require the technical result of lemma 1, so we begin with that result.

A.1 Proof of lemma 1

The proof is based on the calculations in Lemma 3.1 of [13], where the partial derivative of the inverse of a parameterized matrix is computed. Let $\widehat{h}_l$ be an arbitrary component of $\widehat{h}$, then compute the partial derivative

$$\frac{\partial \mathcal{P}^{-1}(\widehat{h})}{\partial \widehat{h}_l} = -\mathcal{P}^{-1}(\widehat{h})\frac{\partial \mathcal{P}(\widehat{h})}{\partial \widehat{h}_l}\mathcal{P}^{-1}(\widehat{h}) = -\mathcal{P}^{-1}(\widehat{h})\mathcal{M}_l\mathcal{P}^{-1}(\widehat{h}). \tag{A.1}$$

Multiply $\widehat{q^x}$ on both sides, then

$$\frac{\partial \widehat{u}}{\partial \widehat{h}_l} = \frac{\partial \mathcal{P}^{-1}(\widehat{h})}{\partial \widehat{h}_l}\widehat{q^x} = -\mathcal{P}^{-1}(\widehat{h})\mathcal{M}_l\mathcal{P}^{-1}(\widehat{h})\widehat{q^x} = -\mathcal{P}^{-1}(\widehat{h})\mathcal{M}_l\widehat{u}, \tag{A.2}$$

resulting in

$$\frac{\partial \widehat{u}}{\partial \widehat{h}} = -\mathcal{P}^{-1}(\widehat{h})[\mathcal{M}_1\widehat{u}, \ldots, \mathcal{M}_K\widehat{u}] \overset{(2.9)}{=} -\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}). \tag{A.3}$$

Similarly, we also get

$$\frac{\partial \widehat{v}}{\partial \widehat{h}} = -\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}). \tag{A.4}$$

In addition, by definition (2.14), it is straightforward that

$$\frac{\partial \widehat{u}}{\partial \widehat{q^x}} = \frac{\partial \widehat{v}}{\partial \widehat{q^y}} = \mathcal{P}^{-1}(\widehat{h}), \qquad \frac{\partial \widehat{v}}{\partial \widehat{q^x}} = \frac{\partial \widehat{u}}{\partial \widehat{q^y}} = 0. \tag{A.5}$$

This completes the proof.

A.2 Proof of lemma 2

The proof is straightforward and follows from computing the Hessian to show that it is positive definite. Recall the entropy function,

$$E(\widehat{U}) = \frac{1}{2}\Big((\widehat{q^x})^\top\widehat{u} + (\widehat{q^y})^\top\widehat{v}\Big) + \frac{1}{2}g\|\widehat{h}\|^2 + g\widehat{h}^\top\widehat{B}, \tag{A.6}$$

where we denote the terms separately,

$$f_1 \coloneqq \frac{1}{2}(\widehat{q^x})^\top\widehat{u}, \qquad f_2 \coloneqq \frac{1}{2}(\widehat{q^y})^\top\widehat{v}, \qquad f_3 \coloneqq \frac{1}{2}g\|\widehat{h}\|^2 + g\widehat{h}^\top\widehat{B}. \tag{A.7}$$

By Lemma 1, we compute the first-order partial derivatives of f_1,

$$\begin{aligned}
&\frac{\partial f_1}{\partial \widehat{h}} = \frac{1}{2}(\widehat{q^x})^\top\frac{\partial \widehat{u}}{\partial \widehat{h}} \overset{(3.9)}{=} -\frac{1}{2}(\widehat{q^x})^\top\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}) = -\frac{1}{2}\Big(\mathcal{P}^{-1}(\widehat{h})\widehat{q^x}\Big)^\top\mathcal{P}(\widehat{u}) = -\frac{1}{2}\widehat{u}^\top\mathcal{P}(\widehat{u}),\\
&\frac{\partial f_1}{\partial \widehat{q^x}} \overset{(3.9)}{=} \widehat{u}^\top, \quad \frac{\partial f_1}{\partial \widehat{q^y}} = 0,
\end{aligned} \tag{A.8}$$

and the second-order partial derivatives,

$$\begin{aligned}
&\frac{\partial^2 f_1}{\partial \widehat{h}^2} = -\frac{1}{2}\frac{\partial}{\partial \widehat{h}}\big(\widehat{u}^\top\mathcal{P}(\widehat{u})\big) \overset{(3.9)}{=} \mathcal{P}(\widehat{u})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}),\\
&\frac{\partial^2 f_1}{\partial \widehat{h}\partial \widehat{q^x}} \overset{(3.9)}{=} -\mathcal{P}(\widehat{u})\mathcal{P}^{-1}(\widehat{h})\\
&\frac{\partial^2 f_1}{\partial \widehat{q^x}^2} \overset{(3.9)}{=} \mathcal{P}^{-1}(\widehat{h}),
\end{aligned} \tag{A.9}$$

with the remaining elements in the Hessian of f_1 equal to zero, i.e.,

$$\frac{\partial^2 f_1}{\partial \widehat{U}^2} = \begin{pmatrix} \mathcal{P}(\widehat{u})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}) & -\mathcal{P}(\widehat{u})\mathcal{P}^{-1}(\widehat{h}) & 0 \\ -\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{u}) & \mathcal{P}^{-1}(\widehat{h}) & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A.10}$$

Similarly to f_1, we compute the Hessian of f_2 as follows,

$$\frac{\partial^2 f_2}{\partial \widehat{U}^2} = \begin{pmatrix} \mathcal{P}(\widehat{v})\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}) & 0 & -\mathcal{P}(\widehat{v})\mathcal{P}^{-1}(\widehat{h}) \\ 0 & 0 & 0 \\ -\mathcal{P}^{-1}(\widehat{h})\mathcal{P}(\widehat{v}) & 0 & \mathcal{P}^{-1}(\widehat{h}) \end{pmatrix}. \tag{A.11}$$

In addition, a direct computation gives

$$\frac{\partial f_3}{\partial \widehat{U}} = \Big(g(\widehat{h}+\widehat{B})^\top, 0, 0\Big) \qquad \Longrightarrow \qquad \frac{\partial^2 f_3}{\partial \widehat{U}^2} = \begin{pmatrix} gI & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A.12}$$

Finally, for any vector $\big(w_1^\top, w_2^\top, w_3^\top\big)^\top \in \mathbb{R}^{3K}$, we compute the quadratic form associated with the Hessian matrices as follows:

$$\begin{aligned} \big(w_1^\top, w_2^\top, w_3^\top\big)\frac{\partial^2 f_1}{\partial \widehat{U}^2}\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} &= \big(\mathcal{P}(\widehat{u})w_1 - w_2\big)^\top \mathcal{P}^{-1}(\widehat{h})\big(\mathcal{P}(\widehat{u})w_1 - w_2\big) \geq 0, \\ \big(w_1^\top, w_2^\top, w_3^\top\big)\frac{\partial^2 f_2}{\partial \widehat{U}^2}\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} &= \big(\mathcal{P}(\widehat{v})w_1 - w_3\big)^\top \mathcal{P}^{-1}(\widehat{h})\big(\mathcal{P}(\widehat{v})w_1 - w_3\big) \geq 0, \\ \big(w_1^\top, w_2^\top, w_3^\top\big)\frac{\partial^2 f_3}{\partial \widehat{U}^2}\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} &= g\|w_1\|^2 \geq 0. \end{aligned} \tag{A.13}$$

This results in that the quadratic form associated with the Hessian of E being non-negative, provided that $\mathcal{P}$ is positive definite. In addition, the quadratic form vanishes if and only if $w_1 = w_2 = w_3 = \mathbf{0}$, implying that E is strictly convex in $\widehat{U}$, completing the proof.

A.3 Proof of lemma 3

Instead of directly demonstrating (3.10), we will present the equivalent compatibility condition,

$$\frac{\partial E}{\partial \widehat{U}}\Big(\frac{\partial \widehat{F}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x} + \frac{\partial \widehat{G}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial y} - \widehat{S}\Big) = \frac{\partial H}{\partial x} + \frac{\partial K}{\partial y}. \tag{A.14}$$

We divide each of the entropy and flux functions into two parts, respectively,

$$E(\widehat{U}) = E_1 + E_2, \qquad E_1 = \frac{1}{2}\Big((\widehat{q^x})^\top\widehat{u} + (\widehat{q^y})^\top\widehat{v}\Big) + \frac{1}{2}g\|\widehat{h}\|^2, \qquad E_2 = g\widehat{h}^\top\widehat{B}, \tag{A.15}$$

$$H(\widehat{U}) = H_1 + H_2, \qquad H_1 = \frac{1}{2}\Big(\widehat{u}^\top\mathcal{P}(\widehat{q^x})\widehat{u} + \widehat{v}^\top\mathcal{P}(\widehat{q^x})\widehat{v}\Big) + g(\widehat{q^x})^\top\widehat{h}, \qquad H_2 = g(\widehat{q^x})^\top\widehat{B}, \tag{A.16}$$

$$K(\widehat{U}) = K_1 + K_2, \qquad K_1 = \frac{1}{2}\Big(\widehat{v}^\top\mathcal{P}(\widehat{q^y})\widehat{v} + \widehat{u}^\top\mathcal{P}(\widehat{q^y})\widehat{u}\Big) + g(\widehat{q^y})^\top\widehat{h} \qquad K_2 = g(\widehat{q^y})^\top\widehat{B}. \tag{A.17}$$

The computation in Lemma 2 gives,

$$\frac{\partial E_1}{\partial \widehat{U}} = \Big(-\frac{1}{2}\widehat{u}^\top\mathcal{P}(\widehat{u}) - \frac{1}{2}\widehat{v}^\top\mathcal{P}(\widehat{v}) + g\widehat{h}^\top, \widehat{u}^\top, \widehat{v}^\top\Big), \qquad \frac{\partial E_2}{\partial \widehat{U}} = \big(g\widehat{B}^\top, 0, 0\big), \tag{A.18}$$

which implies

$$-\frac{\partial E_1}{\partial \widehat{U}}\widehat{S} = g(\widehat{q^x})^\top\widehat{B}_x + g(\widehat{q^y})^\top\widehat{B}_y. \tag{A.19}$$

The equations above, the flux Jacobians in (2.15), and the source term in (2.13), indicate

$$-\frac{\partial E_1}{\partial \widehat{U}}\widehat{S} + \frac{\partial E_2}{\partial \widehat{U}}\Big(\frac{\partial \widehat{F}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x} + \frac{\partial \widehat{G}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial y} - \widehat{S}\Big) \tag{A.20}$$

$$= g(\widehat{q^x})^\top\widehat{B}_x + g(\widehat{q^y})^\top\widehat{B}_y + g\widehat{B}^\top\Big(\frac{\partial \widehat{q^x}}{\partial x} + \frac{\partial \widehat{q^y}}{\partial y}\Big) \tag{A.21}$$

$$= \Big(g(\widehat{q^x})^\top\widehat{B}_x + g\widehat{B}^\top\frac{\partial \widehat{q^x}}{\partial x}\Big) + \Big(g(\widehat{q^y})^\top\widehat{B}_y + g\widehat{B}^\top\frac{\partial \widehat{q^y}}{\partial y}\Big) \tag{A.22}$$

$$= \frac{\partial H_2}{\partial x} + \frac{\partial K_2}{\partial y}. \tag{A.23}$$

Therefore, the compatibility condition (A.14) is equivalent to the following condition,

$$\frac{\partial E_1}{\partial \widehat{U}}\Big(\frac{\partial \widehat{F}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x}+\frac{\partial \widehat{G}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial y}\Big)=\frac{\partial H_1}{\partial x}+\frac{\partial K_1}{\partial y}=\frac{\partial H_1}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x}+\frac{\partial K_1}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial y}. \tag{A.24}$$

It suffices to show

$$\frac{\partial E_1}{\partial \widehat{U}}\frac{\partial \widehat{F}}{\partial \widehat{U}}=\frac{\partial H_1}{\partial \widehat{U}}, \qquad \frac{\partial E_1}{\partial \widehat{U}}\frac{\partial \widehat{G}}{\partial \widehat{U}}=\frac{\partial K_1}{\partial \widehat{U}}, \tag{A.25}$$

which can be computed directly from (A.18), (2.15), (A.15), and (3.9).

B Proofs for section 4.2

In this section we provide proofs for lemmas 5 to 7 in section 4.2, which are the crucial ingredients to proving that the choice of numerical fluxes (4.12) yields a second-order, well-balanced, EC scheme.

B.1 Proof of lemma 5

Assume $(\widehat{U},\widehat{B})$ are spatially smooth functions and $(\boldsymbol{U}_i,\boldsymbol{B}_i)$ are the exact cell averages. It suffices to compare the numerical fluxes $\mathcal{F}_{i+\frac12,j},\mathcal{G}_{i,j+\frac12}$ and the numerical source term $\boldsymbol{S}_{i,j}$ to the exact fluxes $\widehat{F}(\widehat{U})$ at $(x_{i+\frac12},y_j)$, and $\widehat{G}(\widehat{U})$ at $(x_i,y_{j+\frac12})$, and the exact source function $\widehat{S}(\widehat{U})$ at (x_i,y_j), respectively. We can rewrite $\widehat{F}(\widehat{U})$ and $\widehat{G}(\widehat{U})$ in (2.12) as

$$\begin{aligned}
\widehat{F}(\widehat{U}) &\overset{(2.14)}{=} \begin{pmatrix} \mathcal{P}(\widehat{h})\widehat{u} \\ \mathcal{P}(\widehat{q^x})\widehat{u}+\frac12 g\mathcal{P}(\widehat{h})\widehat{h} \\ \mathcal{P}(\widehat{q^x})\widehat{v} \end{pmatrix} \overset{(2.9),(2.14)}{=} \begin{pmatrix} \mathcal{P}(\widehat{h})\widehat{u} \\ \mathcal{P}(\widehat{u})\mathcal{P}(\widehat{h})\widehat{u}+\frac12 g\mathcal{P}(\widehat{h})\widehat{h} \\ \mathcal{P}(\widehat{v})\mathcal{P}(\widehat{h})\widehat{u} \end{pmatrix}, \\
\widehat{G}(\widehat{U}) &\overset{(2.14)}{=} \begin{pmatrix} \mathcal{P}(\widehat{h})\widehat{v} \\ \mathcal{P}(\widehat{q^y})\widehat{u} \\ \mathcal{P}(\widehat{q^y})\widehat{v}+\frac12 g\mathcal{P}(\widehat{h})\widehat{h} \end{pmatrix} \overset{(2.9),(2.14)}{=} \begin{pmatrix} \mathcal{P}(\widehat{h})\widehat{v} \\ \mathcal{P}(\widehat{u})\mathcal{P}(\widehat{h})\widehat{v} \\ \mathcal{P}(\widehat{v})\mathcal{P}(\widehat{h})\widehat{v}+\frac12 g\mathcal{P}(\widehat{h})\widehat{h} \end{pmatrix}.
\end{aligned} \tag{B.1}$$

Notice the following quantitative approximations in space:

$$\overline{\boldsymbol{U}}_{i+\frac12,j}=\widehat{U}(x_{i+\frac12},y_j)+\mathcal{O}(\Delta x^2), \qquad \overline{\boldsymbol{U}}_{i,j+\frac12}=\widehat{U}(x_i,y_{j+\frac12})+\mathcal{O}(\Delta y^2), \tag{B.2a}$$

$$\overline{\boldsymbol{u}}_{i+\frac12,j}=\widehat{u}(x_{i+\frac12},y_j)+\mathcal{O}(\Delta x^2), \qquad \overline{\boldsymbol{u}}_{i,j+\frac12}=\widehat{u}(x_i,y_{j+\frac12})+\mathcal{O}(\Delta y^2), \tag{B.2b}$$

$$\overline{\boldsymbol{v}}_{i+\frac12,j}=\widehat{v}(x_{i+\frac12},y_j)+\mathcal{O}(\Delta x^2), \qquad \overline{\boldsymbol{v}}_{i,j+\frac12}=\widehat{v}(x_i,y_{j+\frac12})+\mathcal{O}(\Delta y^2), \tag{B.2c}$$

$$[\![\boldsymbol{U}]\!]_{i+\frac12,j}=\Delta x\widehat{U}_x(x_{i+\frac12},y_j)+\mathcal{O}(\Delta x^2), \qquad [\![\boldsymbol{U}]\!]_{i,j+\frac12}=\Delta y\widehat{U}_y(x_i,y_{j+\frac12})+\mathcal{O}(\Delta y^2), \tag{B.2d}$$

$$\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac12,j})=\mathcal{P}(\widehat{h}(x_{i+\frac12},y_j))+\mathcal{O}(\Delta x^2), \qquad \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac12})=\mathcal{P}(\widehat{h}(x_i,y_{j+\frac12}))+\mathcal{O}(\Delta y^2), \tag{B.2e}$$

$$\mathcal{P}(\overline{\boldsymbol{u}}_{i+\frac12,j})=\mathcal{P}(\widehat{u}(x_{i+\frac12},y_j))+\mathcal{O}(\Delta x^2), \qquad \mathcal{P}(\overline{\boldsymbol{u}}_{i,j+\frac12})=\mathcal{P}(\widehat{u}(x_i,y_{j+\frac12}))+\mathcal{O}(\Delta y^2), \tag{B.2f}$$

$$\mathcal{P}(\overline{\boldsymbol{v}}_{i+\frac12,j})=\mathcal{P}(\widehat{v}(x_{i+\frac12},y_j))+\mathcal{O}(\Delta x^2), \qquad \mathcal{P}(\overline{\boldsymbol{v}}_{i,j+\frac12})=\mathcal{P}(\widehat{v}(x_i,y_{j+\frac12}))+\mathcal{O}(\Delta y^2). \tag{B.2g}$$

With the quantitative approximations above, it is straightforward to show that the numerical fluxes (4.12) are second-order approximations of the exact fluxes (B.1) respectively, i.e.,

$$\mathcal{F}^{EC}_{i+\frac12,j}=\widehat{F}(\widehat{U})|_{x_{i+\frac12},y_j}+\mathcal{O}(\Delta x^2), \qquad \mathcal{G}^{EC}_{i,j+\frac12}=\widehat{G}(\widehat{U})|_{x_i,y_{j+\frac12}}+\mathcal{O}(\Delta y^2). \tag{B.3}$$

Moreover, to show that the scheme (4.3) has a second-order spatial local truncation error, it suffices to show that $\frac{\mathcal{F}^{EC}_{i+\frac12,j}-\mathcal{F}^{EC}_{i-\frac12,j}}{\Delta x}$ is a second-order approximation of $\widehat{F}(\widehat{U})_x$ at (x_i,y_j), and $\frac{\mathcal{G}^{EC}_{i,j+\frac12}-\mathcal{G}^{EC}_{i,j-\frac12}}{\Delta y}$ is a second-order approximation of $\widehat{G}(\widehat{U})_y$ at (x_i,y_j), and the numerical source term $\boldsymbol{S}_{i,j}$ is a second-order approximation of $\widehat{S}(\widehat{U})$ at (x_i,y_j). This can be shown directly by applying the Taylor's expansion to the numerical fluxes at (x_i,y_j). For example, consider the first K-block term corresponding to $\widehat{h}$ in $\frac{\mathcal{F}^{EC}_{i+\frac12,j}-\mathcal{F}^{EC}_{i-\frac12,j}}{\Delta x}$, assuming $\boldsymbol{U}_{i,j}=\widehat{U}(x_i,y_j)$,

$$\frac{(\mathcal{F}^{EC}_{i+\frac12,j})^h-(\mathcal{F}^{EC}_{i-\frac12,j})^h}{\Delta x}=\frac{\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac12,j})\overline{\boldsymbol{u}}_{i+\frac12,j}-\mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac12,j})\overline{\boldsymbol{u}}_{i-\frac12,j}}{\Delta x}\overset{(B.2b),(B.2e)}{=}\big((\mathcal{P}(\widehat{h})\widehat{u})_x\big)_{i,j}+\mathcal{O}(\Delta x^2), \tag{B.4}$$

where $\big((\mathcal{P}(\widehat{h})\widehat{u})_x\big)_{i,j}$ is the partial derivative with respect to x of the exact flux at (x_i,y_j). More detailed computations are as follows for one of the terms: We compute the local truncation error of $\frac{(\mathcal{F}^{EC}_{i+\frac12,j})^h-(\mathcal{F}^{EC}_{i-\frac12,j})^h}{\Delta x}$ to the exact flux

$\big(\widehat{F}^h(\widehat{U})\big)_x = \big(\mathcal{P}(\widehat{h})\widehat{u}\big)_x$. First, by (4.12),

$$\begin{aligned}\frac{(\mathcal{F}^{EC}_{i+\frac{1}{2},j})^h - (\mathcal{F}^{EC}_{i-\frac{1}{2},j})^h}{\Delta x} &= \frac{\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} - \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})\overline{\boldsymbol{u}}_{i-\frac{1}{2},j}}{\Delta x}\\ =\frac{1}{4\Delta x}\bigg(&\Big(\mathcal{P}(\widehat{h}_{i,j}) + \mathcal{P}(\widehat{h}_{i+1,j})\Big)(\widehat{u}_{i,j} + \widehat{u}_{i+1,j}) - \Big(\mathcal{P}(\widehat{h}_{i,j}) + \mathcal{P}(\widehat{h}_{i-1,j})\Big)(\widehat{u}_{i,j} + \widehat{u}_{i-1,j})\bigg).\end{aligned} \tag{B.5}$$

Then, we apply the Taylor's expansion of the crossing product terms in the second equality, i.e., $\mathcal{P}(\widehat{h}_{i,j})\widehat{u}_{i+1,j}$, $\mathcal{P}(\widehat{h}_{i+1,j})\widehat{u}_{i,j}$, $\mathcal{P}(\widehat{h}_{i,j})\widehat{u}_{i-1,j}$ and $\mathcal{P}(\widehat{h}_{i-1,j})\widehat{u}_{i,j}$ at (x_i, y_j). This results in

$$\begin{aligned}\text{(B.5)} =&\frac{1}{4\Delta x}\bigg(\mathcal{P}(\widehat{h}_{i,j})\widehat{u}_{i,j} + \mathcal{P}(\widehat{h}_{i,j})\Big(\widehat{u}_{i,j} + \Delta x(\widehat{u}_x)_{i,j} + \frac{\Delta x^2}{2}(\widehat{u}_{xx})_{i,j}) + \frac{\Delta x^3}{6}(\widehat{u}_{xxx})_{i,j}) + \mathcal{O}(\Delta x^4)\Big)\\ &+ \widehat{u}_{i,j}\Big(\mathcal{P}(\widehat{h}_{i,j}) + \Delta x(\mathcal{P}(\widehat{h}_x))_{i,j} + \frac{\Delta x^2}{2}(\mathcal{P}(\widehat{h}_{xx}))_{i,j}) + \frac{\Delta x^3}{6}(\mathcal{P}(\widehat{h}_{xxx}))_{i,j}) + \mathcal{O}(\Delta x^4)\Big) + (\mathcal{P}(\widehat{h})\widehat{u})_{i+1,j}\bigg)\\ &- \frac{1}{4\Delta x}\bigg(\mathcal{P}(\widehat{h}_{i,j})\widehat{u}_{i,j} + \mathcal{P}(\widehat{h}_{i,j})\Big(\widehat{u}_{i,j} - \Delta x(\widehat{u}_x)_{i,j} + \frac{\Delta x^2}{2}(\widehat{u}_{xx})_{i,j}) - \frac{\Delta x^3}{6}(\widehat{u}_{xxx})_{i,j}) + \mathcal{O}(\Delta x^4)\Big)\\ &+ \widehat{u}_{i,j}\Big(\mathcal{P}(\widehat{h}_{i,j}) - \Delta x(\mathcal{P}(\widehat{h}_x))_{i,j} + \frac{\Delta x^2}{2}(\mathcal{P}(\widehat{h}_{xx}))_{i,j}) - \frac{\Delta x^3}{6}(\mathcal{P}(\widehat{h}_{xxx}))_{i,j}) + \mathcal{O}(\Delta x^4)\Big) + (\mathcal{P}(\widehat{h})\widehat{u})_{i-1,j}\bigg).\end{aligned} \tag{B.6}$$

After combining the terms in the same order, we get

$$\begin{aligned}\text{(B.6)} =&\frac{1}{4\Delta x}\Big(2\Delta x\mathcal{P}(\widehat{h}_{i,j})(\widehat{u}_x)_{i,j} + 2\Delta x\mathcal{P}(\widehat{h}_x)_{i,j}(\widehat{u})_{i,j} + \frac{\Delta x^3}{3}(\widehat{u}_{xxx})_{i,j}\Big)\\ &+ \frac{1}{4\Delta x}\Big((\mathcal{P}(\widehat{h})\widehat{u})_{i+1,j} - (\mathcal{P}(\widehat{h})\widehat{u})_{i-1,j}\Big) + \mathcal{O}(\Delta x^3).\end{aligned} \tag{B.7}$$

Finally, we again apply the Taylor's expansion to $(\mathcal{P}(\widehat{h})\widehat{u})_{i+1,j} - (\mathcal{P}(\widehat{h})\widehat{u})_{i-1,j}$ to obtain a second-order approximation of $\big(\mathcal{P}(\widehat{h})\widehat{u}\big)_x$ at (x_i, y_j)

$$\begin{aligned}\text{(B.7)} =&\frac{1}{2}((\mathcal{P}(\widehat{h})\widehat{u})_x)_{i,j} + \frac{\Delta x^2}{12}(\widehat{u}_{xxx})_{i,j} + \frac{1}{4\Delta x}\Big(2\Delta x((\mathcal{P}(\widehat{h})\widehat{u})_x)_{i,j} + \frac{\Delta x^3}{3}\big((\mathcal{P}(\widehat{h})\widehat{u})_{xxx}\big)_{i,j}\Big) + \mathcal{O}(\Delta x^3)\\ =&((\mathcal{P}(\widehat{h})\widehat{u})_x)_{i,j} + \frac{\Delta x^2}{12}\Big((\widehat{u}_{xxx})_{i,j} + \big((\mathcal{P}(\widehat{h})\widehat{u})_{xxx}\big)_{i,j}\Big) + \mathcal{O}(\Delta x^3).\end{aligned} \tag{B.8}$$

Similarly, other terms in the numerical fluxes can be evaluated by the Taylor's expansion to verify the second-order local truncation error of the energy conservative scheme.

In this way, one can directly compute the local truncation error to show that the scheme (4.3), with the numerical fluxes and source term (4.12), is second-order spatially accurate.

B.2 Proof of lemma 6

Suppose we are given the initial data

$$\boldsymbol{u}_{i,j} = \boldsymbol{v}_{i,j} \equiv \boldsymbol{0}, \qquad \boldsymbol{h}_{i,j} + \boldsymbol{B}_{i,j} = \text{const vector}, \qquad \forall i, j. \tag{B.9}$$

The time-independent bottom topography leads to $\frac{d}{dt}\boldsymbol{B}_{i,j} = 0, \forall i, j$. To prove the well-balanced property, it suffices to show

$$\frac{d}{dt}\boldsymbol{h}_{i,j} \equiv 0, \qquad \frac{d}{dt}\boldsymbol{q}^x_{i,j} \equiv 0, \qquad \frac{d}{dt}\boldsymbol{q}^y_{i,j} \equiv 0, \qquad \forall i, j. \tag{B.10}$$

By the straightforward substitution and the fact that the discretization of the velocities are both zero vectors, due to (B.10),

$$\begin{aligned}\frac{d}{dt}\boldsymbol{h}_{i,j} =& -\frac{1}{\Delta x}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j} - \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})\overline{\boldsymbol{u}}_{i-\frac{1}{2},j}\right)\\ &-\frac{1}{\Delta y}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j+\frac{1}{2}} - \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})\overline{\boldsymbol{v}}_{i,j-\frac{1}{2}}\right)\\ =&\boldsymbol{0}.\end{aligned} \tag{B.11}$$

Before investigating the time derivative of the discharge, we first introduce useful identities:

$$\begin{aligned}\Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i+\frac{1}{2},j} - \Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i-\frac{1}{2},j} &\overset{(4.6)}{=} \frac{1}{2}\left([\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i+\frac{1}{2},j} + [\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i-\frac{1}{2},j}\right)\\ &\overset{(4.8a)}{=} \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{h}]\!]_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})[\![\boldsymbol{h}]\!]_{i-\frac{1}{2},j},\\ \Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i,j+\frac{1}{2}} - \Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i,j-\frac{1}{2}} &\overset{(4.6)}{=} \frac{1}{2}\left([\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i,j+\frac{1}{2}} + [\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i,j-\frac{1}{2}}\right)\\ &\overset{(4.8a)}{=} \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{h}]\!]_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})[\![\boldsymbol{h}]\!]_{i,j-\frac{1}{2}}.\end{aligned} \tag{B.12}$$

Now by straightforward substitution, using identities (B.12) and the initial data, we obtain the following result:

$$
\begin{aligned}
\frac{d}{dt}\boldsymbol{q}^x_{i,j} &\overset{(4.12)}{=} -\frac{g}{2\Delta x}\left(\left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i+\frac{1}{2},j} - \left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i-\frac{1}{2},j}\right)\\
&\quad -\frac{g}{2\Delta x}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i-\frac{1}{2},j}\right)\\
&\overset{(B.12)}{=} -\frac{g}{2\Delta x}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i-\frac{1}{2},j}\right) \overset{(B.9)}{=} \boldsymbol{0},\\
\frac{d}{dt}\boldsymbol{q}^y_{i,j} &\overset{(4.12)}{=} -\frac{g}{2\Delta y}\left(\left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i,j+\frac{1}{2}} - \left(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\right)_{i,j-\frac{1}{2}}\right)\\
&\quad -\frac{g}{2\Delta y}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j-\frac{1}{2}}\right)\\
&\overset{(B.12)}{=} -\frac{g}{2\Delta y}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i,j-\frac{1}{2}}\right) \overset{(B.9)}{=} \boldsymbol{0},
\end{aligned}
\tag{B.13}
$$

which completes the proof.

B.3 Proof of lemma 7

By multiplying both sides of (4.3) by $\boldsymbol{V}_{i,j}^\top$ and using the definition $\boldsymbol{V}_{i,j} \coloneqq (\frac{\partial \boldsymbol{E}_{i,j}}{\partial \boldsymbol{U}_{i,j}})^\top$, we obtain

$$
\frac{d}{dt}\boldsymbol{E}_{i,j} = -\frac{1}{\Delta x}\left(\boldsymbol{V}_{i,j}^\top \mathcal{F}_{i+\frac{1}{2},j} - \boldsymbol{V}_{i,j}^\top \mathcal{F}_{i-\frac{1}{2},j}\right) - \frac{1}{\Delta y}\left(\boldsymbol{V}_{i,j}^\top \mathcal{G}_{i,j+\frac{1}{2}} - \boldsymbol{V}_{i,j}^\top \mathcal{G}_{i,j-\frac{1}{2}}\right) + \boldsymbol{V}_{i,j}^\top \boldsymbol{S}_{i,j}. \tag{B.14}
$$

We estimate these terms separately. The first term on the right-hand side can be expanded as follows:

$$
\begin{aligned}
\boldsymbol{V}_{i,j}^\top \mathcal{F}_{i+\frac{1}{2},j} &\overset{(4.7)}{=} \overline{\boldsymbol{V}}_{i+\frac{1}{2},j}^\top \mathcal{F}_{i+\frac{1}{2},j} - \frac{1}{2}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2},j}^\top \mathcal{F}_{i+\frac{1}{2},j}\\
&\overset{(4.13)(4.14)}{=} \mathcal{H}_{i+\frac{1}{2},j} + \overline{\boldsymbol{\Psi}}_{i+\frac{1}{2},j} + \frac{g}{4}[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j}^\top \mathcal{P}(\overline{h}_{i+\frac{1}{2},j})[\![\boldsymbol{u}]\!]_{i+\frac{1}{2},j}\\
&\quad - \frac{1}{2}[\![\boldsymbol{\Psi}]\!]_{i+\frac{1}{2},j} - \frac{g}{2}[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j}^\top \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\overline{\boldsymbol{u}}_{i+\frac{1}{2},j}\\
&\overset{(4.7)}{=} \mathcal{H}_{i+\frac{1}{2},j} + \boldsymbol{\Psi}_{i,j} - \frac{g}{2}[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j}^\top \mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})\boldsymbol{u}_{i,j}.
\end{aligned}
\tag{B.15}
$$

Using an analogous computation, we can also obtain the second term,

$$
\boldsymbol{V}_{i,j}^\top \mathcal{F}_{i-\frac{1}{2},j} = \mathcal{H}_{i-\frac{1}{2},j} + \boldsymbol{\Psi}_{i,j} + \frac{g}{2}[\![\boldsymbol{B}]\!]_{i-\frac{1}{2},j}^\top \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})\boldsymbol{u}_{i,j}. \tag{B.16}
$$

Additionally, in a similar manner, and with the corresponding quantities in (4.13) and (4.14), we derive

$$
\begin{aligned}
\boldsymbol{V}_{i,j}^\top \mathcal{G}_{i,j+\frac{1}{2}} &= \mathcal{K}_{i,j+\frac{1}{2}} + \boldsymbol{\Phi}_{i,j} - \frac{g}{2}[\![\boldsymbol{B}]\!]_{i,j+\frac{1}{2}}^\top \mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})\boldsymbol{v}_{i,j},\\
\boldsymbol{V}_{i,j}^\top \mathcal{G}_{i,j-\frac{1}{2}} &= \mathcal{K}_{i,j-\frac{1}{2}} + \boldsymbol{\Phi}_{i,j} + \frac{g}{2}[\![\boldsymbol{B}]\!]_{i,j-\frac{1}{2}}^\top \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})\boldsymbol{v}_{i,j}.
\end{aligned}
\tag{B.17}
$$

Finally, a direct computation shows

$$
\begin{aligned}
\boldsymbol{V}_{i,j}^\top \boldsymbol{S}_{i,j} = &-\frac{g\boldsymbol{u}_{i,j}^\top}{2\Delta x}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i+\frac{1}{2},j} + \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2},j})[\![\boldsymbol{B}]\!]_{i-\frac{1}{2},j}\right)\\
&-\frac{g\boldsymbol{v}_{i,j}^\top}{2\Delta y}\left(\mathcal{P}(\overline{\boldsymbol{h}}_{i,j+\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j+\frac{1}{2}} + \mathcal{P}(\overline{\boldsymbol{h}}_{i,j-\frac{1}{2}})[\![\boldsymbol{B}]\!]_{i,j-\frac{1}{2}}\right)
\end{aligned}
\tag{B.18}
$$

The formula (B.14), together with the combination of expressions (B.15), (B.16), (B.17), and (B.18), shows that the scheme (4.3) is energy conservative, with energy fluxes defined in (4.14).