ESAIM: M2AN 58 (2024) 723–757
https://doi.org/10.1051/m2an/2024012

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

ENERGY STABLE AND STRUCTURE-PRESERVING SCHEMES FOR THE STOCHASTIC GALERKIN SHALLOW WATER EQUATIONS

Dihan Dai[1], Yekaterina Epshteyn[2] and Akil Narayan[1,*]

Abstract. The shallow water flow model is widely used to describe water flows in rivers, lakes, and coastal areas. Accounting for uncertainty in the corresponding transport-dominated nonlinear PDE models presents theoretical and numerical challenges that motivate the central advances of this paper. Starting with a spatially one-dimensional hyperbolicity-preserving, positivity-preserving stochastic Galerkin formulation of the parametric/uncertain shallow water equations, we derive an entropy-entropy flux pair for the system. We exploit this entropy-entropy flux pair to construct structure-preserving second-order energy conservative, and first- and second-order energy stable finite volume schemes for the stochastic Galerkin shallow water system. The performance of the methods is illustrated on several numerical experiments.

Mathematics Subject Classification. 35L65, 35Q35, 35R60, 65M60, 65M70, 65M08.

Received October 9, 2023. Accepted February 22, 2024.

1. Introduction

The one-dimensional Saint-Venant system of shallow water equations (SWE) is a popular model of water flows where vertical length scales are much smaller than horizontal ones [1]. This system in conservative form is given by,

$$U_t + F(U)_x = S(U), \qquad U = (h, q)^\top \in \mathbb{R}^2, \tag{1.1}$$

where $U = U(x,t)$ is the vector of conservative variables; $h(x,t)$ is the water height (a mass-like variable) and $q(x,t)$ is the water discharge (a momentum-like variable). The flux F and source term S are given by,

$$F(U) = \begin{pmatrix} q \\ \frac{(q)^2}{h} + \frac{gh^2}{2} \end{pmatrix}, \qquad S(U) = \begin{pmatrix} 0 \\ -ghB' \end{pmatrix} \tag{1.2}$$

Keywords and phrases. Stochastic Galerkin method, finite volume method, structure-preserving discretization, shallow water equations, hyperbolic systems of conservation law and balance laws.

[1] Department of Mathematics and Scientific Computing and Imaging (SCI) Institute, The University of Utah, Salt Lake City, UT 84112, USA.

[2] Department of Mathematics, The University of Utah, Salt Lake City, UT 84112, USA.

*Corresponding author: akil@sci.utah.edu

where $B(x)$ is the (assumed known) bottom topography function and $g > 0$ is the gravitational constant. The system (1.1) is supplemented with initial and boundary data that we omit for the time being.

The one-dimensional SWE model (1.1) is a hyperbolic system of partial differential equations (PDE) if $h > 0$, and hence with the non-zero source S, then (1.1) is a nonlinear hyperbolic balance law. Because of this, it inherits the standard challenges in developing numerical methods for such models: solutions generically develop discontinuities in finite time even with smooth initial data, non-uniqueness of weak solutions should be rectified by an implicit or explicit numerical imposition of entropy conditions, and implicit time-integration solvers are challenging to implement due to the nonlinearity [10, 27, 28]. In addition to all this, the SWE has challenges that are somewhat specific to its particular form: positivity of the water height h should be maintained, and numerical schemes should accurately capture near-equilibrium dynamics, which is typically achieved by imposing the *well-balanced* property [3], *i.e.*, that the PDE equilibrium states are exactly captured at the discrete level.

A more nebulous and hence more frustrating challenge is that of *uncertainty* in the model. For example, one may have incomplete, partial information about the initial data or the bottom topography function B. In such cases, one frequently models this data as a random variable or process, and hence the solution U to (1.1) is random. We consider the somewhat more simple situation when the input uncertainty is encoded with a finite-dimensional random variable, in which case (1.1) becomes a parametric model (with the input random variables serving as the parameters). Even with this simplification, the parametric or stochastic nature of the solution exacerbates many of the previously described numerical challenges. A particularly successful approach for handling such problems that we will employ is the polynomial Chaos (PC) method, wherein U is approximated as a polynomial function of the input parameters [34, 35, 44].

The class of *non-intrusive* PC strategies construct the polynomial by collecting an ensemble of solutions to (1.1) at a collection of fixed values of the parameters. This approach is attractive since it can exploit existing and trusted legacy solvers for (1.1), for which there are several effective choices [4, 9, 14, 23–25, 29, 32, 39–43, 45, 47]. However, this approach suffers from the disadvantage that making concrete statements about the quality or properties of the resulting polynomial approximation can be challenging. For example, one cannot guarantee that entropy conditions are satisfied if the polynomial approximation is evaluated away from the parameter ensemble used to construct the approximation.

This paper is concerned with an alternative *intrusive* approach, the stochastic Galerkin (SG) method for PC approximation, which addresses the parametric dependence in a Galerkin fashion, *e.g.*, by enforcing that certain probabilistic moments of (1.1) vanish. This approach has the potential to provide pathways to mathematical rigor of numerical methods through weak enforcement of the parametric dependence. SG methods transform a parametric model (1.1) into a new non-parametric model of larger system size. Since the new SG formulation is non-parametric, one can apply typical deterministic numerical methods for systems of PDEs to solve the SG problem. Such approaches have shown particular success for modeling parametric dependence in elliptic problems; see, *e.g.*, [8]. However, the notable drawback of SG methods when applied to (nonlinear) hyperbolic PDEs is that the new non-parametric SG system need not be a hyperbolic PDE itself, which changes the essential character of the SG system relative to the original system. Despite this challenge, recent work has investigated numerical methods for stochastic Galerkin methods for various types of hyperbolic conservation laws. Some advances include SG-type analysis and algorithms for scalar conservation laws [46] including well-balanced methods [22], Haar wavelet-based SG approaches [20], hyperbolicity preservation through a non-equivalent Roe variables formulation [19], filtering strategies for SG systems [26], limiter-type methods to maintain hyperbolicity [33], hyperbolicity formulations for linear problems [31] or using linearization techniques [38], splitting type approaches for SWE models [7], non-conservative formulations of SG SWE systems [5] and approximate representations through entropic variables [30].

Our focus starts from recent work that has developed an SG formulation for the SWE in conservative form that involves a special SG treatment for the nonlinear, non-polynomial terms [11]. This is a particular distinction of our approach: We require no non-conservative formulation, transformation, numerical filtering/limiting, or linearization. Such an approach can be used to develop a well-balanced, hyperbolicity-preserving, and positivity-preserving finite volume method to solve the SG SWE system. The approach forming our starting

TABLE 1. Notation and terminology used throughout this article.

K	Number of terms in a PC expansion
$\widehat{U} = (\widehat{h}^\top, \widehat{q}^\top)^\top \in \mathbb{R}^K \times \mathbb{R}^K$	PC state vector and its components for SG SWE
$\boldsymbol{U} = (\boldsymbol{h}^\top, \boldsymbol{q}^\top)^\top \in \mathbb{R}^K \times \mathbb{R}^K$	Quantities related to cell averages/reconstructed values, etc. of the PC state vector and its components for SG SWE
(E, H)	Entropy pair for SG SWE
$\widehat{u}$	PC vector for velocity
$\boldsymbol{u}$	Quantity related to cell averages/reconstructed values, etc. of the PC vector for velocity
$\widehat{V}, \boldsymbol{V}$	The PC vector for the entropic variable and quantities related to cell averages/reconstructed values, etc. of the entropic variable
$\mathcal{P}(\cdot)$	The operator that maps a PC vector to a PC triple-product matrix
$\overline{\boldsymbol{h}}_{i+\frac{1}{2}}$	Arithmetic average of $\boldsymbol{h}_i$ and $\boldsymbol{h}_{i+1}$. Similar notation is applied to other bold letters, *e.g.*, $\overline{\boldsymbol{U}}_{i+\frac{1}{2}}$
$[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}}$	First-order jump defined *via* cell averages
$\langle\!\langle \boldsymbol{h} \rangle\!\rangle_{i+\frac{1}{2}}$	(Notationally) Second-order jump defined *via* reconstructed values
$\widehat{F}$	Flux
$\mathcal{F}$	Numerical flux
$\boldsymbol{Q}^{\mathrm{EC}}, \boldsymbol{Q}^{\mathrm{ES1}}, \boldsymbol{Q}^{\mathrm{ES2}}$	Energy-conservative flux, 1st-order energy-stable flux, and 2nd-order energy stable flux, respectively
$\boldsymbol{w}^{\pm}, \widetilde{\boldsymbol{w}}^{\pm}$	Scaled variables

point has also been extended to two-dimensional SWE systems [12], but we focus on the single spatial dimension case.

1.1. Contributions of this article

We make the following contributions that build on [11]:

- We derive an entropy-entropy flux pair for the spatially one-dimensional hyperbolicity-preserving, positivity-preserving SG SWE system derived in [11], see Theorem 3.1. Entropy-entropy flux pairs are the theoretical starting point for proposing 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 SG SWE, all of which are also well-balanced. See Theorems 4.1–4.3, with the procedure in Algorithm 1. The designed energy conservative and energy stable schemes are stochastic extensions of the schemes developed in [17,18].
- We provide numerical experiments that explore the simulation capabilities of the new schemes. To the best of our knowledge, these are the first schemes for any SG SWE system that boast energy stability, the well-balanced property, while also being positivity- and hyperbolicity-preserving.

An outline of this paper is as follows: Section 2 introduces our notation, along with background on PC methods and the SG SWE system from [11]. Section 3 provides our entropy-entropy pair construction for the SG SWE system. Section 4 provides the statement of the energy conservative and energy stable schemes that we develop, along with proofs of their theoretical properties, as well as their algorithmic details. Section 5 compiles numerical examples that demonstrate the performance of our scheme. Section 6 gives brief summary of the main results and some future research directions. We summarize our notation in this article in Table 1.

2. Preliminaries

2.1. Notation

We use $\|\cdot\|$ to denote the standard Euclidean (ℓ^2) norm operating on vectors. If $f : \mathbb{R}^m \to \mathbb{R}^n$ for $m, n \in \mathbb{N}$, then we write $f(x)$ for $x = (x_1, \ldots, x_m)$, and $f(x) = (f_1, \ldots, f_n)$. We use the following notation for the gradient:

$$f_x \coloneqq \frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_m} \end{pmatrix} \in \mathbb{R}^{n \times m}.$$

When $n = 1$ (*i.e.*, f is scalar-valued) then $\frac{\partial^2 f}{\partial x^2}$ is the $n \times n$ Hessian of f. If $\boldsymbol{A}$ is a square matrix, then we write $\boldsymbol{A} > 0$ and $\boldsymbol{A} \geq 0$ when $\boldsymbol{A}$ is positive definite and positive semi-definite, respectively.

In work on the SWE system (1.1) it is common to introduce the water velocity (equilibrium) variable

$$u \coloneqq \frac{q}{h}, \tag{2.1}$$

and we also make use of this variable in what follows.

2.2. Polynomial chaos expansion

In this section, we briefly review the results and notation for polynomial chaos expansion. More comprehensive results can be found in [13, 35, 44], etc.

Let $\xi \in \mathbb{R}^d$ be a random variable associated with Lebesgue density function ρ. Define the function space

$$L^2_\rho(\mathbb{R}^d) \coloneqq \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| \left(\int_{\mathbb{R}^d} f^2(s)\rho(s)\,\mathrm{d}s \right)^{\frac{1}{2}} < +\infty \right\}.$$

Assuming finite polynomial moments of all orders for ρ, there exists an orthonormal basis $\{\phi_k\}_{k=1}^\infty$ of L^2_ρ, *i.e.*,

$$\langle \phi_k, \phi_\ell \rangle_\rho \coloneqq \int_{\mathbb{R}^d} \phi_k(s)\phi_\ell(s)\rho(s)\,\mathrm{d}s = \delta_{k,\ell}, \qquad \phi_1(\xi) \equiv 1, \tag{2.2}$$

for all $k, \ell \in \mathbb{N}$, where $\delta_{k,\ell}$ is the Kronecker delta. PC seeks a representation of a random field $z(\cdot, \cdot, \xi) \in L^2_\rho$ in terms of a series of orthonormal polynomials for ξ,

$$z(x, t, \xi) \overset{L^2_\rho}{=} \sum_{k=1}^{\infty} \widehat{z}_i(x, t)\phi_i(\xi), \tag{2.3}$$

where x, t are the deterministic spatial and temporal variables, and $\widehat{z}_i(x, t)$ are deterministic Fourier-like coefficients. The equation (2.3) holds true for all $z(x, t; \cdot) \in L^2_\rho$ under mild conditions [15]. In practice, a finite truncation of (2.3) is usually considered. Let P be a K-dimensional polynomial subspace of L^2_ρ,

$$P = \operatorname{span}\{\phi_k, \quad k = 1, \ldots, K\}, \tag{2.4}$$

i.e., we let ϕ_k be an orthonormal basis for P. We make the common assumption that $1 \in P$, and for convenience we assume that,

$$\phi_1(\xi) \equiv 1.$$

A popular choice for P is the total degree space, but several other options are possible.

One choice of K-term PC *approximation* of a random field z in P is the projection of (2.3) onto P:

$$\Pi_P[z](x,t,\xi) \coloneqq \sum_{k=1}^{K} \widehat{z}_k(x,t)\phi_k(\xi). \tag{2.5}$$

Using the orthogonality of the basis function, the statistics of $\Pi_P[z]$ can be expressed in terms of the expansion coefficients. For example, the mean and the variance of $\Pi_P[z]$ are given by:

$$\mathbb{E}[\Pi_P[z](x,t,\xi)] = \widehat{z}_1(x,t), \quad \mathrm{Var}[\Pi_P[z](x,t,\xi)] = \sum_{k=2}^{K} \widehat{z}_k^2(x,t). \tag{2.6}$$

Let $\widehat{z} = (\widehat{z}_1, \cdots, \widehat{z}_k) \in \mathbb{R}^K$ be the vector of the expansion coefficients in (2.6). Define the linear operator $\mathcal{P} : \mathbb{R}^K \to \mathbb{R}^{K\times K}$ as

$$\mathcal{P}(\widehat{z}) \coloneqq \sum_{k=1}^{K} \widehat{z}_k \mathcal{M}_k, \qquad \mathcal{M}_k \in \mathbb{R}^{K\times K}, \qquad (\mathcal{M}_k)_{\ell,m} = \langle \phi_k, \phi_\ell \phi_m \rangle_\rho. \tag{2.7}$$

Fixing $\widehat{z} \in \mathbb{R}^K$, then $\mathcal{P}(\widehat{z})$ is the (symmetric) quadratic form matrix representation of the bilinear operator $(\widehat{a},\widehat{b}) \mapsto \langle a_P\, z_P, b_P\rangle_\rho$, where $z_P \coloneqq \sum_{k=1}^K \widehat{z}_k\phi_k(\xi)$ and similarly for a_P, b_P with $\widehat{a},\widehat{b} \in \mathbb{R}^K$. Using the fact that $(\mathcal{M}_k)_{\ell,m}$ is commutative in (k,m) a direct computation shows:

$$\mathcal{P}(\widehat{a}) = \big(\mathcal{M}_1\widehat{a},\ \mathcal{M}_2\widehat{a},\ \ldots,\ \mathcal{M}_K\widehat{a}\big). \tag{2.8}$$

A useful lemma is given as follows.

Lemma 2.1. *For any two vectors $\widehat{a},\widehat{b} \in \mathbb{R}^K$,*

$$\mathcal{P}(\widehat{a})\widehat{b} = \mathcal{P}\Big(\widehat{b}\Big)\widehat{a}, \qquad \widehat{b}^\top\mathcal{P}(\widehat{a}) = \widehat{a}^\top\mathcal{P}\Big(\widehat{b}\Big). \tag{2.9}$$

The proof is straightforward using (2.7) and (2.8) along with the symmetry of $\mathcal{P}(\cdot)$. This result is a "commutative" property of the operator $\mathcal{P}(\cdot)$. For example: For any $a,b,c \in \mathbb{R}^K$,

$$\frac{\partial}{\partial c} a^\top\mathcal{P}(c)b = a^\top\mathcal{P}(b). \tag{2.10}$$

A *stochastic Galerkin* (SG) formulation of a ξ-parameterized PDE corresponds to making the ansatz that the state variable lies in the space P, and projecting the PDE residual onto the same space. Straightforward applications of this procedure to (nonlinear) hyperbolic PDEs typically do not result in hyperbolic SG formulations.

2.3. Hyperbolic-preserving stochastic Galerkin formulation for shallow water equation

In [11], we have derived a hyperbolicity-preserving stochastic Galerkin formulation for the shallow water equations. We briefly recall the results in this section.

We make the ansatz that h, q lie in the polynomial space P,

$$h \simeq h_P \coloneqq \sum_{k=1}^{K} \widehat{h}_k(x,t)\phi_k(\xi), \tag{2.11a}$$

$$q \simeq q_P \coloneqq \sum_{k=1}^{K} (\widehat{q})_k(x,t)\phi_k(\xi), \tag{2.11b}$$

and use these to formulate a ξ-variable Galerkin projection of the SWE. We make a special choice of how the

Galerkin projection of the nonlinear, non-polynomial term $(q)^2/h$ is truncated, which results in a *new* (stochastic Galerkin) system of balance laws whose state variables are the expansion coefficients in (2.11) [11]:

$$\widehat{U}_t + \left(\widehat{F}\left(\widehat{U}\right)\right)_x = \widehat{S}\left(\widehat{U}\right). \tag{2.12}$$

Here, $\widehat{U} \coloneqq (\widehat{h}^\top, \widehat{q}^\top)^\top \in \mathbb{R}^{2K}$, where $\widehat{h}, \widehat{q}$ are each length-K vectors of the expansion coefficients in (2.11). The flux and the source terms are,

$$\widehat{F}\left(\widehat{U}\right) = \begin{pmatrix} \widehat{q} \\ \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right)\widehat{q} + \frac{1}{2}g\mathcal{P}\left(\widehat{h}\right)\widehat{h} \end{pmatrix}, \qquad \widehat{S}\left(\widehat{U}\right) = \begin{pmatrix} 0 \\ -g\mathcal{P}\left(\widehat{h}\right)\widehat{B_x} \end{pmatrix}, \tag{2.13}$$

cf. (1.2). The flux Jacobian, written in $K \times K$ blocks, is given by

$$\frac{\partial \widehat{F}}{\partial \widehat{U}} = \begin{pmatrix} O & I \\ g\mathcal{P}\left(\widehat{h}\right) - \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right)\mathcal{P}(\widehat{u}) & \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right) + \mathcal{P}(\widehat{u}) \end{pmatrix}. \tag{2.14}$$

We have introduced the term

$$\widehat{u} = \mathcal{P}^{-1}\left(\widehat{h}\right)\widehat{q}, \tag{2.15}$$

which we view as the vector of the PC coefficients of the x-velocity u introduced in (2.1), and is well-defined if $\mathcal{P}(\widehat{h})$ is invertible.

The deterministic SWE are hyperbolic if the water height $h > 0$; there is a natural extension of this property to the SG SWE.

Theorem 2.1 ([11], Thm. 3.1)**.** *If the matrix $\mathcal{P}(\widehat{h})$ is strictly positive definite for every point (x,t) in the computational spatial-temporal domain, then the SG formulation (2.12) is hyperbolic.*

This is proven by identifying a stochastic extension of the known eigenvector matrix for the deterministic SWE flux Jacobian $\frac{\partial F}{\partial U}$, and using this to show that $\frac{\partial \widehat{F}}{\partial \widehat{U}}$ is similar to a symmetric matrix and hence (2.12) is hyperbolic [11].

3. An entropy-entropy flux pair for SG SWE systems

The formulation (2.12) will be considered in what follows. Our goal will be to derive entropy-entropy flux pairs for these formulations. The first step is for us to recall a known entropy-entropy flux pair for the *deterministic* SWE system.

3.1. Entropy-entropy flux pairs for deterministic shallow water equations

It is well-known that solutions to systems of conservation/balance laws can develop shock discontinuities in finite time for generic initial data. Therefore, weak solutions, *i.e.*, solutions in the sense of distributions, are usually considered. However, weak solutions are not necessarily unique, and to mitigate this issue an additional *entropy admissibility criteria* is imposed [2,10] to identify the physically meaningful solution.

For a general balance law in one space dimension

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

its entropy-entropy flux pair $(E(U), H(U))$ satisfies a *companion* balance law

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

where the *entropy* $E(U)$ is a scalar function that is convex in U, and H is an *entropy flux* function. In order to be consistent with the original balance law for smooth U, the entropy-entropy flux pair (E, H) should satisfy the following *compatibility condition*,

$$\frac{\partial E}{\partial U}(F_x - S) = H_x, \tag{3.3}$$

which is simply the condition ensuring that multiplying (3.1) by $\frac{\partial E}{\partial U}$ recovers (3.2) when solutions are smooth. In the case of $S \equiv 0$ and $(E, H) = (E(U), H(U))$, equation (3.3) is the usual entropy condition for conservation laws. For a general system of balance laws in several spatial dimensions, an entropy-entropy flux pair need not exist. However, for a hyperbolic system of balance laws emerging from continuum physics, the companion balance law (3.2) is usually related to the Second Law of thermodynamics, and the total energy of the system often serves as the entropy function. A variety of examples can be found in Section 3.3 from [10]. For the *deterministic* SWE system in (1.1), the total energy [17] is

$$E^d(U) = \underbrace{\frac{1}{2}qu}_{\text{kinetic energy}} + \underbrace{\frac{1}{2}gh^2 + ghB}_{\text{potential energy}} \tag{3.4}$$

where we recall that u is the velocity defined in (2.1). For any smooth solution U, a direct calculation yields,

$$E^d(U)_t + H^d(U)_x = 0, \tag{3.5}$$

where

$$H^d(U) = \frac{1}{2}qu^2 + gqh + gqB. \tag{3.6}$$

This, along with the fact that E^d is convex in U, establishes that (E^d, H^d) is a valid entropy-entropy flux pair for (1.1). For (weak) solutions with shocks, the entropy admissibility criteria is that energy should dissipate in accordance with a vanishing viscosity principle,

$$E^d(U)_t + H^d(U)_x \leq 0. \tag{3.7}$$

In what follows we will identify entropy-entropy flux pairs for the SG SWE model. This amounts to verifying that (i) such a pair satisfies the companion balance law (an equality for smooth solutions) and (ii) that the entropy function is convex in the state variable.

3.2. An entropy-entropy flux pair for the one-dimensional SG SWE

This section is dedicated to identifying an entropy-entropy flux pair for the SG system (2.12). In this section, we will return to the notation $\widehat{U}$ (containing PC expansion coefficients) for the derivation of an entropy-entropy flux pair for the SG system. Our main result in this section is the following entropy entropy-flux pair for the one-dimensional SG SWE:

Theorem 3.1. *Define the function,*

$$E\left(\widehat{U}\right) = \frac{1}{2}\left((\widehat{q})^\top \widehat{u} + g\left\|\widehat{h}\right\|^2\right) + g\widehat{h}^\top \widehat{B}, \tag{3.8a}$$

and also the flux function,

$$H(\widehat{U}) = \frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{q})\widehat{u} + g\widehat{q}^\top \widehat{h} + g\widehat{q}^\top \widehat{B}. \tag{3.8b}$$

If $\mathcal{P}(\widehat{h}) > 0$, then (E, H) is an entropy-entropy flux pair for the one-dimensional SG SWE (2.12).

Recall that $\widehat{u}$ above is defined in (2.15), and contains PC expansion coefficients for the velocity u defined in (2.1). In the absence of uncertainty, equation (3.8a) reduces to the deterministic total energy (3.4). The rest of this section is devoted to proving Theorem 3.1, which amounts to showing that, if $\mathcal{P}(\widehat{h}) > 0$, then E is convex in $\widehat{U}$ and (E, H) satisfy the companion balance law,

$$E\Big(\widehat{U}\Big)_t + H\Big(\widehat{U}\Big)_x = 0, \tag{3.9}$$

for smooth solutions $\widehat{U}$. Note that for non-smooth solutions, equation (3.9) holds with $=$ replaced by $\leq$. We prove Theorem 3.1 with three intermediate results. Our first result is a technical condition that facilitates later computations.

Lemma 3.1 (Gradient of $\widehat{u}$)**.** *Let $\widehat{q} \in \mathbb{R}^K$ be arbitrary, and let $\widehat{h} \in \mathbb{R}^K$ be such that $\mathcal{P}(\widehat{h})$ is invertible. Defining $\widehat{u}$ as in (2.15), then*

$$\frac{\partial \widehat{u}}{\partial \widehat{U}} = \left[\frac{\partial \widehat{u}}{\partial \widehat{h}}, \ \frac{\partial \widehat{u}}{\partial \widehat{q}}\right] = \left[-\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{P}(\widehat{u}), \ \ \mathcal{P}^{-1}\Big(\widehat{h}\Big)\right]. \tag{3.10}$$

Proof. If $A(t)$ is a t-parameterized matrix, then for any t at which A is invertible,

$$\frac{\partial}{\partial t}A^{-1}(t) = -A^{-1}(t)\frac{\partial A(t)}{\partial t}A^{-1}(t).$$

Applying this to $\mathcal{P}$, we have,

$$\frac{\partial \mathcal{P}^{-1}\Big(\widehat{h}\Big)}{\partial \widehat{h}_\ell} = -\mathcal{P}^{-1}\Big(\widehat{h}\Big)\frac{\partial \mathcal{P}\Big(\widehat{h}\Big)}{\partial \widehat{h}_\ell}\mathcal{P}^{-1}\Big(\widehat{h}\Big) \overset{(2.8)}{=} -\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{M}_\ell\mathcal{P}^{-1}\Big(\widehat{h}\Big), \tag{3.11}$$

and hence,

$$\frac{\partial \widehat{u}}{\partial \widehat{h}_\ell} = \frac{\partial \mathcal{P}^{-1}\Big(\widehat{h}\Big)}{\partial \widehat{h}_\ell}\widehat{q} \overset{(3.11)}{=} -\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{M}_\ell\mathcal{P}^{-1}\Big(\widehat{h}\Big)\widehat{q} \overset{(2.15)}{=} \mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{M}_\ell\widehat{u}. \tag{3.12}$$

Therefore,

$$\frac{\partial \widehat{u}}{\partial \widehat{h}} = \left[-\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{M}_1\widehat{u}, \ \ \cdots \ \ -\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{M}_K\widehat{u}\right] \overset{(2.8)}{=} -\mathcal{P}^{-1}\Big(\widehat{h}\Big)\mathcal{P}(\widehat{u}), \tag{3.13}$$

proving the desired relation for $\frac{\partial \widehat{u}}{\partial \widehat{h}}$. The relation for $\frac{\partial \widehat{u}}{\partial \widehat{q}}$ is immediate from the definition (2.15). □

Lemma 3.2 (Convexity of $E(\widehat{U})$)**.** *If $\mathcal{P}(\widehat{h})$ is positive definite, then the function $E(\widehat{U})$ defined in (3.8a) is convex in $\widehat{U}$.*

Proof. Using the definition (2.15) of $\widehat{u}$, note that,

$$E\Big(\widehat{U}\Big) = \underbrace{\frac{1}{2}(\widehat{q})^\top\mathcal{P}^{-1}\Big(\widehat{h}\Big)\widehat{q}}_{f_1(\widehat{U})} + \underbrace{\frac{g}{2}\widehat{h}^\top\widehat{h} + g\widehat{h}^\top\widehat{B}}_{f_2(\widehat{U})}, \tag{3.14}$$

and therefore in particular,

$$\frac{\partial^2 E}{\partial \widehat{U}^2} = \frac{\partial^2 f_1}{\partial \widehat{U}^2} + \frac{\partial^2 f_2}{\partial \widehat{U}^2}. \tag{3.15}$$

We will show that this Hessian is positive definite. Clearly we have,

$$\frac{\partial f_2}{\partial \widehat{U}} = \left(g\widehat{h}^\top + g\widehat{B}^\top,\ 0\right) \in \mathbb{R}^{1\times 2K}, \qquad \frac{\partial^2 f_2}{\partial \widehat{U}^2} = \begin{pmatrix} gI & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{2K\times 2K}. \tag{3.16}$$

Using the previous lemma, we can directly compute,

$$\frac{\partial f_1}{\partial \widehat{h}} = \frac{1}{2}(\widehat{q})^\top \frac{\partial \widehat{u}}{\partial \widehat{h}} \overset{(3.10),(2.15)}{=} -\frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{u}), \qquad \frac{\partial f_1}{\partial \widehat{q}} = (\widehat{q})^\top \mathcal{P}^{-1}\left(\widehat{h}\right) = \widehat{u}^\top, \tag{3.17}$$

which in turn implies,

$$\frac{\partial^2 f_1}{\partial \widehat{q}^2} = \mathcal{P}^{-1}\left(\widehat{h}\right), \qquad \frac{\partial^2 f_1}{\partial \widehat{h}\partial \widehat{q}} \overset{(3.10)}{=} \left(-\mathcal{P}^{-1}\left(\widehat{h}\right)\mathcal{P}(\widehat{u})\right)^\top = -\mathcal{P}(\widehat{u})\mathcal{P}^{-1}\left(\widehat{h}\right),$$

and finally,

$$\frac{\partial^2 f_1}{\partial \widehat{h}^2} = \frac{1}{2}\frac{\partial}{\partial \widehat{h}}\left(-\widehat{u}^\top \mathcal{P}(\widehat{u})\right) \overset{(3.10)}{=} \mathcal{P}(\widehat{u})\mathcal{P}^{-1}\left(\widehat{h}\right)\mathcal{P}(\widehat{u}).$$

Hence, the Hessian of f_1 is,

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

A direct computation of the quadratic form associated to this Hessian using an arbitrary vector $(w_1^\top, w_2^\top)^\top \in R^{2K}$ yields,

$$\left(w_1^\top, w_2^\top\right)\frac{\partial^2 f_1}{\partial \widehat{U}^2}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \left(\mathcal{P}(\widehat{u})w_1 - w_2\right)^\top \mathcal{P}^{-1}\left(\widehat{h}\right)\left(\mathcal{P}(\widehat{u})w_1 - w_2\right) \geq 0.$$

Finally, combining the above with (3.15) and (3.16) yields,

$$\left(w_1^\top, w_2^\top\right)\frac{\partial^2 E}{\partial \widehat{U}^2}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = g\|w_1\|^2 + \left(\mathcal{P}(\widehat{u})w_1 - w_2\right)^\top \mathcal{P}^{-1}\left(\widehat{h}\right)\left(\mathcal{P}(\widehat{u})w_1 - w_2\right),$$

which is non-negative since $\mathcal{P}(\widehat{h})$ is positive-definite. Therefore, E is convex, as desired. In addition, since the above expression vanishes if and only if $w_1 = w_2 = 0$, then E is also strictly convex. □

The final piece needed to prove Theorem 3.1 is to establish that the entropy function E along with the flux function H defined in (3.8b) satisfy the companion balance law.

Lemma 3.3 ((E, H) satisfy the companion balance law)**.** *When $\widehat{U}$ is a smooth function, the pair (E, H) defined in (3.8) satisfies*

$$E\left(\widehat{U}\right)_t + H\left(\widehat{U}\right)_x = 0. \tag{3.18}$$

Proof. The compatibility condition we seek to show, equivalent to (3.18), is,

$$\frac{\partial E}{\partial \widehat{U}}\left(\frac{\partial \widehat{F}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x} - \widehat{S}\right) = \frac{\partial H}{\partial x}, \tag{3.19}$$

cf. (3.3). To proceed we split both entropy functions into two pieces:

$$E\left(\widehat{U}\right) = E_1\left(\widehat{U}\right) + E_2\left(\widehat{U}\right), \qquad E_1\left(\widehat{U}\right) \coloneqq \frac{1}{2}\left((\widehat{q})^\top \widehat{u} + g\left\|\widehat{h}\right\|^2\right), \qquad E_2\left(\widehat{U}\right) \coloneqq g\widehat{h}^\top \widehat{B}, \tag{3.20a}$$

$$H\left(\widehat{U}\right) = H_1\left(\widehat{U}\right) + H_2\left(\widehat{U}\right), \qquad H_1\left(\widehat{U}\right) \coloneqq \frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{q})\widehat{u} + g\widehat{q}^\top \widehat{h}, \qquad H_2\left(\widehat{U}\right) = g\widehat{q}^\top \widehat{B}. \tag{3.20b}$$

From (3.14), (3.16), and (3.17), we have already computed the gradient of E:

$$\frac{\partial E_1}{\partial \widehat{U}} = \left(-\frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{u}) + g\widehat{h}^\top, \;\; \widehat{u}^\top\right), \qquad \frac{\partial E_2}{\partial \widehat{U}} = \left(g\widehat{B}^\top, \;\; 0\right). \tag{3.21}$$

Combining these expressions with the flux Jacobian in (2.14) and the source term in (2.13) yields,

$$-\frac{\partial E_1}{\partial \widehat{U}}\widehat{S} + \frac{\partial E_2}{\partial \widehat{U}}\left(\frac{\partial \widehat{F}}{\partial \widehat{U}}\frac{\partial \widehat{U}}{\partial x} - \widehat{S}\right) = g\widehat{B}^\top \frac{\partial \widehat{q}}{\partial x} + g\widehat{q}^\top \widehat{B}_x \overset{(3.20b)}{=} \frac{\partial H_2}{\partial x}. \tag{3.22a}$$

Note then that if we are able to show,

$$\frac{\partial E_1}{\partial \widehat{U}}\frac{\partial \widehat{F}}{\partial \widehat{U}} = \frac{\partial H_1}{\partial \widehat{U}}, \tag{3.22b}$$

then the expressions (3.22) are equivalent to (3.19). Therefore, we are left only to show (3.22b). A direct computation with (3.21) and (2.14) yields,

$$\frac{\partial E_1}{\partial \widehat{U}}\frac{\partial \widehat{F}}{\partial \widehat{U}} = \left(g\widehat{q}^\top - \widehat{u}^\top \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right)\mathcal{P}(\widehat{u}), \quad g\widehat{h}^\top + \frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{u}) + \widehat{u}^\top \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right)\right).$$

On the other hand, we have the expressions,

$$\begin{aligned}
\frac{\partial}{\partial \widehat{h}}\frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{q})\widehat{u} &= \widehat{u}^\top \mathcal{P}(\widehat{q})\frac{\partial \widehat{u}}{\partial \widehat{h}} \overset{(3.10)}{=} -\widehat{u}^\top \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right)\mathcal{P}(\widehat{u}), \\
\frac{\partial}{\partial \widehat{q}}\frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{q})\widehat{u} &= \widehat{u}^\top \mathcal{P}(\widehat{q})\frac{\partial \widehat{u}}{\partial \widehat{q}} + \frac{1}{2}\left(\frac{\partial}{\partial \widehat{q}} z^\top \mathcal{P}(\widehat{q}) z\right)\Big|_{z \leftarrow \widehat{u}} \overset{(2.10),(3.10)}{=} \widehat{u}^\top \mathcal{P}(\widehat{q})\mathcal{P}^{-1}\left(\widehat{h}\right) + \frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{u})
\end{aligned}$$

and using these to compute $\frac{\partial H_1}{\partial \widehat{U}}$ shows that (3.22b) is true, completing the proof. □

The proof of Theorem 3.1 is complete: Lemmas 3.2 and 3.3 imply that (E, H) as defined in (3.8) are an entropy-entropy flux pair for (2.12).

Remark 3.1. The quantities,

$$\widehat{V} \coloneqq \left(\frac{\partial E}{\partial \widehat{U}}\right)^\top = \left(-\frac{1}{2}\widehat{u}^\top \mathcal{P}(\widehat{u}) + g\left(\widehat{h} + \widehat{B}\right)^\top, \widehat{u}^\top\right)^\top, \qquad \Psi \coloneqq \widehat{V}\widehat{F} - H \overset{(2.15),(3.8b)}{=} \frac{1}{2}g\widehat{u}^\top \mathcal{P}\left(\widehat{h}\right)\widehat{h}, \tag{3.23}$$

are called the *entropy variable* and *stochastic energy potential*, respectively. These variables serve important roles in the construction of the energy conservative and the energy stable schemes that we develop later.

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

In this section, we present several well-balanced energy conservative and energy stable numerical scheme for the SG SWE. The schemes designed below are stochastic extensions of the schemes developed in [17]. Our entropy-entropy flux pairs developed in Section 3 will be crucial ingredients for energy conservative and energy stable schemes for the SG formulation (2.12)–(2.15).

We also need to specify the well-balanced property we are interested in: By "well-balanced", we mean that the scheme can preserve the stochastic "lake-at-rest" state exactly at the discrete level.

Definition 4.1 (Well-Balanced SG SWE Property, [11])**.** We say that a solution (h_P, q_P) to (2.12) is well-balanced if it satisfies the *stochastic* "lake-at-rest" solution,

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

where $C(\xi)$ is a random scalar depending only on ξ, Π_P corresponds to a polynomial truncation, cf. (2.5), and subscripts P refer to the (stochastic) discrete solutions in the subspace P. In terms of our previous notation for P-expansion coefficients, equation (4.1) is equivalent to the following vector equation

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

where $\mathcal{D}$ is the spatial domain and T is the terminal time.

We emphasize that even without introducing the lake-at-rest definition (4.1), the vector equation (4.2) itself is a steady state of the SG system (2.12).

4.1. Energy conservative schemes

We consider the semi-discrete form for FV schemes for (2.12) over a uniform mesh in the x variable:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}_i = -\frac{\mathcal{F}_{i+\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}}{\Delta x} + \boldsymbol{S}_i. \tag{4.3}$$

Here, $\boldsymbol{U}_i \approx \frac{1}{\Delta x}\int_{\mathcal{I}_i} \widehat{U}(x,t)\,\mathrm{d}x$ is the approximation of the cell averages of $\widehat{U}$ over cells $\mathcal{I}_i := [x_{i-1/2}, x_{i+1/2}]$ at time t, and $\Delta x = |\mathcal{I}_i| = x_{i+1/2} - x_{i-1/2}$. The terms $\mathcal{F}_{i\pm1/2}$ are numerical fluxes at the boundaries of the cells, which are functions of neighboring states, *e.g.*, $\mathcal{F}_{i+1/2}$ is a function of $\boldsymbol{U}_i$ and $\boldsymbol{U}_{i+1}$. The term $\boldsymbol{S}_i \approx \frac{1}{\Delta x}\int_{\mathcal{I}_i} \widehat{S}(\widehat{U}, \widehat{B})\,\mathrm{d}x$ is a discretization of the source term, which we will design below to be well-balanced. To reiterate our notation: normal typeset capital letters (sometimes with "hat" notation) refers to degrees of freedom associated to discretizing *only* the stochastic variable ξ, *i.e.*, $(\widehat{U}, \widehat{h}, \widehat{q}, \widehat{B})$. Boldface notation with subscripts i refers to degrees of freedom associated to a subsequent discretization of the spatial variable x over cell $\mathcal{I}_i$, *i.e.*, $(\boldsymbol{U}_i, \boldsymbol{h}_i, \boldsymbol{q}_i, \boldsymbol{B}_i)$. We define the discrete velocity variable $\boldsymbol{u}_i$ in a manner analogous to (2.15):

$$\boldsymbol{u}_i := \mathcal{P}(\boldsymbol{h}_i)^{-1}\boldsymbol{q}_i. \tag{4.4}$$

Discrete entropic quantities are derived from the discrete conservative variables $\boldsymbol{U}_i$ and velocity variable $\boldsymbol{u}_i$. *I.e.*, the following are direct generalizations of the definition of $E(\widehat{U})$ in (3.8a), and of $(\widehat{V}, \Psi)$ in (3.23):

$$\boldsymbol{E}_i := \frac{1}{2}\left(\boldsymbol{q}_i^\top \boldsymbol{u}_i + g\|\boldsymbol{h}_i\|^2\right) + g\boldsymbol{h}_i^\top \boldsymbol{B}_i, \tag{4.5a}$$

$$\boldsymbol{V}_i := \left(\frac{\partial \boldsymbol{E}_i}{\partial \boldsymbol{U}_i}\right)^\top = \left(-\frac{1}{2}\boldsymbol{u}_i^\top \mathcal{P}(\boldsymbol{u}_i) + g(\boldsymbol{h}_i + \boldsymbol{B}_i)^\top, \ \ \boldsymbol{u}_i^\top\right)^\top, \tag{4.5b}$$

$$\boldsymbol{\Psi}_i := \boldsymbol{V}_i \widehat{F}(\boldsymbol{U}_i) - H(\boldsymbol{U}_i) = \frac{1}{2} g \boldsymbol{u}_i^\top \mathcal{P}(\boldsymbol{h}_i)\boldsymbol{h}_i. \tag{4.5c}$$

We now introduce some notation that is used in [17] for averages and jumps at cell interfaces:

$$\overline{\boldsymbol{a}}_{i+1/2} \coloneqq \frac{1}{2}(\boldsymbol{a}_{i+1} + \boldsymbol{a}_i), \qquad [\![\boldsymbol{a}]\!]_{i+1/2} \coloneqq \boldsymbol{a}_{i+1} - \boldsymbol{a}_i, \tag{4.6}$$

where $\boldsymbol{a}_i$ is the cell average over $\mathcal{I}_i$. The expressions above are equivalent to,

$$\boldsymbol{a}_i = \overline{\boldsymbol{a}}_{i+1/2} - \frac{[\![\boldsymbol{a}]\!]_{i+1/2}}{2} = \overline{\boldsymbol{a}}_{i-1/2} + \frac{[\![\boldsymbol{a}]\!]_{i-1/2}}{2}, \tag{4.7}$$

and all these expressions are valid regardless of the size of $\boldsymbol{a}$ (*e.g.*, both row and column vectors are allowed). We will require some additional technical results for interfacial averages and jumps.

Lemma 4.1. *Let $\boldsymbol{a}_i, \boldsymbol{b}_i$ be any spatially discrete quantities. Then:*

$$\mathcal{P}\left(\overline{\boldsymbol{a}}_{i+\frac{i}{2}}\right)[\![\boldsymbol{a}]\!]_{i+\frac{i}{2}} = \frac{1}{2}[\![\mathcal{P}(\boldsymbol{a})\boldsymbol{a}]\!]_{i+\frac{i}{2}} \tag{4.8a}$$

$$[\![\boldsymbol{a}]\!]^\top_{i+\frac{i}{2}}\overline{\boldsymbol{b}}_{i+\frac{i}{2}} + [\![\boldsymbol{b}]\!]^\top_{i+\frac{i}{2}}\overline{\boldsymbol{a}}_{i+\frac{i}{2}} = [\![\boldsymbol{a}^\top\boldsymbol{b}]\!]_{i+\frac{i}{2}}. \tag{4.8b}$$

Proof. Due to linearity of $\mathcal{P}$, then,

$$\mathcal{P}\left(\overline{\boldsymbol{a}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{a}]\!]_{i+\frac{1}{2}} \overset{(2.7)}{=} \mathcal{P}\left(\frac{1}{2}(\boldsymbol{a}_{i+1} + \boldsymbol{a}_i)\right)(\boldsymbol{a}_{i+1} - \boldsymbol{a}_i) \overset{(2.9)}{=} \frac{1}{2}(\mathcal{P}(\boldsymbol{a}_{i+1})\boldsymbol{a}_{i+1} - \mathcal{P}(\boldsymbol{a}_i)\boldsymbol{a}_i) = \frac{1}{2}[\![\mathcal{P}(\boldsymbol{a})\boldsymbol{a}]\!]_{i+\frac{1}{2}},$$

which proves (4.8a). Similarly, equation (4.8b) can be proven directly:

$$\begin{aligned}[\![\boldsymbol{a}]\!]^\top_{i+\frac{1}{2}}\overline{\boldsymbol{b}}_{i+\frac{1}{2}} + [\![\boldsymbol{b}]\!]^\top_{i+\frac{1}{2}}\overline{\boldsymbol{a}}_{i+\frac{1}{2}} &= \frac{1}{2}\Big((\boldsymbol{a}_{i+1} - \boldsymbol{a}_i)^\top(\boldsymbol{b}_{i+1} + \boldsymbol{b}_i) + (\boldsymbol{b}_{i+1} - \boldsymbol{b}_i)^\top(\boldsymbol{a}_{i+1} + \boldsymbol{a}_i)\Big) \\ &= \boldsymbol{a}^\top_{i+1}\boldsymbol{b}_{i+1} - \boldsymbol{a}_i^\top\boldsymbol{b}_i = [\![\boldsymbol{a}^\top\boldsymbol{b}]\!]_{i+\frac{1}{2}}.\end{aligned}$$

□

We now make particular definitions for energy conservative and energy stable schemes for one-dimensional systems of balance laws. To provide context, with no source terms (*i.e.*, $\boldsymbol{S}_i = 0$) then spatial discretizations of the form (4.3) are called *conservative* schemes since they imply,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i\in[M]}\Delta x\boldsymbol{U}_i(t) = \left[\mathcal{F}_{1/2} - \mathcal{F}_{M+1/2}\right], \qquad \text{(vanishing source, } \boldsymbol{S}_i = 0), \tag{4.9}$$

and in particular with periodic boundary conditions, then this implies that the cumulative amount of $\widehat{U}$ in the system is constant in time[1].

To translate this concept to the notion of an energy conservative scheme, note that an entropy-entropy flux pair (E, H) introduced in Section 3 is explicitly a function of the state $\widehat{U}$ and inputs in the source term (here, $\widehat{B}$). Hence, the semi-discrete form (4.3) can be transformed into a semi-discrete form for the companion balance law (3.9). Then we call (4.3) energy conservative if it implies a conservative scheme for the companion balance law that describes the evolution of the entropy (energy).

Definition 4.2 (Energy conservative and energy stable schemes)**.** Suppose that the system of balance laws (2.12) has an entropy-entropy flux pair (E, H) where $E(\widehat{U})$ can be interpreted as energy for the system. Then

[1]For non-periodic boundary conditions, the energy would increase/decrease depending on the boundary conditions and their corresponding impact on the boundary fluxes.

the semi-discrete FV scheme (4.3) is an Energy Conservative (EC) scheme if it can be rewritten as the following semi-discrete form for the evolution of the numerical cell averages $\boldsymbol{E}_i$ of E:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}_i(t) = -\frac{1}{\Delta x}\big(\mathcal{H}_{i+1/2} - \mathcal{H}_{i-1/2}\big), \qquad i \in [M], \tag{4.10}$$

where $\mathcal{H}_{i+1/2}$ is some numerical entropy flux at the interface location $x = x_{i+1/2}$. The scheme (4.3) is called an Energy Stable (ES) scheme if

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}_i(t) \le -\frac{1}{\Delta x}\big(\mathcal{H}_{i+1/2} - \mathcal{H}_{i-1/2}\big), \qquad i \in [M]. \tag{4.11}$$

Note that the definitions above are cell-wise conditions that are stronger than a global condition such as (4.9).

4.2. An EC scheme for the SG SWE

In this section we present an EC scheme for the one-dimensional SG SWE system (2.12). We use the conservative scheme (4.3), with the following choices of flux and source terms:

$$\mathcal{F}_{i+1/2} = \mathcal{F}^{\mathrm{EC}}_{i+1/2} := \begin{pmatrix} \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)\overline{\boldsymbol{u}}_{i+\frac{1}{2}} \\ \frac{g}{2}\Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i+\frac{1}{2}} + \mathcal{P}\Big(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\Big)\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)\overline{\boldsymbol{u}}_{i+\frac{1}{2}} \end{pmatrix}, \tag{4.12a}$$

$$\boldsymbol{S}_i = \begin{pmatrix} \boldsymbol{0} \\ -\frac{g}{2\Delta x}\Big(\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{B}]\!]_{i+\frac{1}{2}} + \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\Big)[\![\boldsymbol{B}]\!]_{i-\frac{1}{2}}\Big) \end{pmatrix}. \tag{4.12b}$$

Above, the interfacial averages $\overline{\boldsymbol{u}}_{i+1/2}$ are computed as defined in (4.6). Our main result for this scheme is as follows.

Theorem 4.1 (EC Scheme)**.** *Suppose the bottom topography function B is independent of time. Consider the semi-discrete scheme (4.3) for the SG SWE system (2.12). Suppose that the flux and source terms are selected as in (4.12). Then, this is a well-balanced EC scheme with local truncation error $\mathcal{O}(\Delta x^2)$.*

The remainder of this section is devoted to the proof, which requires some intermediate steps. First, we show that $\boldsymbol{S}_i$ is a well-balanced choice for the source term discretization.

Lemma 4.2. *Suppose $\boldsymbol{S}_i$ is chosen as in (4.12b). If the bottom topography B is independent of time, then (4.3) is a well-balanced scheme in the sense of Definition 4.1.*

Proof. Given initial data

$$\boldsymbol{u}_i \equiv \boldsymbol{0}, \quad \boldsymbol{h}_i + \boldsymbol{B}_i = \text{const vector}, \qquad \forall i, \tag{4.13}$$

the well-balanced property with time-independent bottom topography (see Def. 4.1) requires that, for every i,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{h}_i \equiv 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_i \equiv 0. \tag{4.14}$$

We first notice that,

$$\begin{aligned} \Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i+\frac{1}{2}} - \Big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\Big)_{i-\frac{1}{2}} &= \frac{1}{2}\Big([\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i+\frac{1}{2}} + [\![\mathcal{P}(\boldsymbol{h})\boldsymbol{h}]\!]_{i-\frac{1}{2}}\Big) \\ &\overset{(4.8a)}{=} \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} + \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\Big)[\![\boldsymbol{h}]\!]_{i-\frac{1}{2}}. \end{aligned} \tag{4.15}$$

Note that with initialization (4.13), then $\boldsymbol{u}_i = 0$, and hence $\overline{\boldsymbol{u}}_{i+1/2} = 0$. Therefore the semi-discrete scheme (4.3) with the flux and source terms in (4.12) yields,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{h}_i =& -\frac{1}{\Delta x}\Big(\mathcal{P}(\overline{\boldsymbol{h}}_{i+\frac{1}{2}})\overline{\boldsymbol{u}}_{i+\frac{1}{2}} - \mathcal{P}(\overline{\boldsymbol{h}}_{i-\frac{1}{2}})\overline{\boldsymbol{u}}_{i-\frac{1}{2}}\Big) = \boldsymbol{0},\\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_i =& -\frac{g}{2\Delta x}\Big(\big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\big)_{i+\frac{1}{2}} - \big(\overline{\mathcal{P}(\boldsymbol{h})\boldsymbol{h}}\big)_{i-\frac{1}{2}}\Big)\\
&-\frac{g}{2\Delta x}\Big(\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{B}]\!]_{i+\frac{1}{2}} + \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\Big)[\![\boldsymbol{B}]\!]_{i-\frac{1}{2}}\Big)\\
\overset{(4.15)}{=}& -\frac{g}{2\Delta x}\Big(\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i+\frac{1}{2}} + \mathcal{P}\Big(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\Big)[\![\boldsymbol{h}+\boldsymbol{B}]\!]_{i-\frac{1}{2}}\Big) \overset{(4.13)}{=} \boldsymbol{0},
\end{aligned} \tag{4.16}$$

which establishes (4.14). □

Lemma 4.3. *The flux and source terms in* (4.12) *commit a local truncation error of* $\mathcal{O}(\Delta x^2)$.

The proof is direct, by assuming $(\boldsymbol{U}_i, \boldsymbol{B}_i)$ are exact cell averages of spatially smooth functions $(\widehat{U}, \widehat{B})$ and then comparing $\mathcal{F}_{i+1/2}$ and $\boldsymbol{S}_i$ to $\widehat{F}(\widehat{U})\big|_{x=x_{i+1/2}}$ and $\widehat{S}(\widehat{U})\big|_{x=x_i}$, respectively, where $\widehat{F}$ and $\widehat{S}$ are the exact flux and source functions in (2.13). Therefore we omit most details, pointing out only the following quantitative approximations in space (ignoring the time variable t):

$$\begin{aligned}
\overline{\boldsymbol{U}}_{i+1/2} &= \widehat{U}(x_{i+1/2}) + \mathcal{O}(\Delta x^2), & [\![\boldsymbol{U}]\!]_{i+1/2} &= \Delta x\, \widehat{U}_x(x_{i+1/2}) + \mathcal{O}(\Delta x^2)\\
\mathcal{P}(\overline{\boldsymbol{h}}_{i+1/2}) &= \mathcal{P}\Big(\widehat{h}\big(x_{i+1/2}\big)\Big) + \mathcal{O}(\Delta x^2), & \overline{\boldsymbol{u}}_{i+1/2} &= \widehat{u}\big(x_{i+1/2}\big) + \mathcal{O}(\Delta x^2).
\end{aligned}$$

Note that the implicit constants hidden in the asymptotic notation above depend on the maximum singular value of $\mathcal{P}(\widehat{h}_{xx}(x))$ and the minimum singular value of $\mathcal{P}(\widehat{h}(x_{i+1/2}))$.

The final result we need is a sufficient condition for a numerical flux to result in an EC scheme.

Lemma 4.4. *Let* $\boldsymbol{S}_i$ *be chosen as in* (4.12b). *Suppose that* $\mathcal{F}_{i+1/2}$ *satisfies*

$$[\![\boldsymbol{V}]\!]^\top_{i+\frac{1}{2}}\mathcal{F}_{i+\frac{1}{2}} = [\![\boldsymbol{\Psi}]\!]_{i+\frac{1}{2}} + g[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2}}\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)\overline{\boldsymbol{u}}_{i+\frac{1}{2}}. \tag{4.17}$$

Then the corresponding FV scheme (4.3) *is an EC scheme, i.e., satisfies* (4.10), *where the numerical energy flux is given by,*

$$\mathcal{H}_{i+\frac{1}{2}} := \overline{\boldsymbol{V}}^\top_{i+\frac{1}{2}}\mathcal{F}_{i+\frac{1}{2}} - \overline{\boldsymbol{\Psi}}_{i+\frac{1}{2}} - \frac{g}{4}[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2}}\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}}. \tag{4.18}$$

Proof. Multiplying (4.3) by $\boldsymbol{V}_i^\top$ and using the definition of $\boldsymbol{V}_i$ in (4.5b), we obtain,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}_i = -\frac{1}{\Delta x}\left(\underbrace{\boldsymbol{V}_i^\top\mathcal{F}_{i+\frac{1}{2}}}_{(A1)} - \underbrace{\boldsymbol{V}_i^\top\mathcal{F}_{i-\frac{1}{2}}}_{(A2)} - \underbrace{\Delta x\boldsymbol{V}_i^\top\overline{\boldsymbol{S}}_i}_{(B)}\right). \tag{4.19}$$

The first term, labeled (A1), can be expanded to,

$$\begin{aligned}
(\mathrm{A1}) &\overset{(4.7)}{=} \overline{\boldsymbol{V}}^\top_{i+1/2}\mathcal{F}_{i+1/2} - \frac{1}{2}[\![\boldsymbol{V}]\!]^\top_{i+1/2}\mathcal{F}_{i+1/2}\\
&\overset{(4.17)(4.18)}{=} \mathcal{H}_{i+\frac{1}{2}} + \overline{\boldsymbol{\Psi}}_{i+\frac{1}{2}} + \frac{g}{4}[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2}}\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}} - \frac{1}{2}[\![\boldsymbol{\Psi}]\!]_{i+\frac{1}{2}} - \frac{g}{2}[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2}}\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\\
&\overset{(4.7)}{=} \mathcal{H}_{i+\frac{1}{2}} + \boldsymbol{\Psi}_i - \frac{g}{2}[\![\boldsymbol{B}]\!]^\top_{i+\frac{1}{2}}\mathcal{P}\Big(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\Big)\boldsymbol{u}_i.
\end{aligned}$$

In an analogous computation, the term labeled (A2) is given by,

$$(\text{A2}) = \mathcal{H}_{i-\frac{1}{2}} + \boldsymbol{\Psi}_i + \frac{g}{2}[\![\boldsymbol{B}]\!]_{i-\frac{1}{2}}^{\top}\mathcal{P}\left(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\right)\boldsymbol{u}_i.$$

Finally, a direct computation shows that term (B) is,

$$(\text{B}) \overset{(4.12b),(4.5b)}{=} -\frac{g}{2}\boldsymbol{u}_i^{\top}\mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{B}]\!]_{i+\frac{1}{2}} - \frac{g}{2}\boldsymbol{u}_i^{\top}\mathcal{P}\left(\overline{\boldsymbol{h}}_{i-\frac{1}{2}}\right)[\![\boldsymbol{B}]\!]_{i-\frac{1}{2}}.$$

Using the expressions for terms (A1), (A2), and (B) derived above in (4.19) establishes that the scheme satisfies (4.10), *i.e.*, is an EC scheme. □

We now have all the ingredients necessary to prove Theorem 4.1.

Proof of Theorem 4.1. Lemmas 4.2 and 4.3 verify that the scheme is well-balanced and second-order. We therefore need only show that it is EC. To do this, we must verify the condition in Lemma 4.4. We accomplish this with direct computation:

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

which verifies (4.17), and hence Lemma 4.4 is applicable, showing that this is an EC scheme. □

4.3. A first-order ES scheme

The scheme determined by (4.12) numerically preserves the energy of the PDE system (1.1). However, it may lead to spurious oscillations since the energy should dissipate in the presence of shocks. The issue can be resolved by introducing appropriate numerical viscosity [16–18, 36, 37]. Our numerical diffusion operators are a straightforward stochastic extension of the energy-stable diffusion operators proposed in [16, 17].

For context of the approach, the introduction of a traditional Roe-type diffusion for a conservation law involves augmenting an EC flux as follows:

$$\mathcal{F}_{i+1/2}^{\text{RD}} \coloneqq \mathcal{F}_{i+1/2}^{\text{EC}} - \frac{1}{2}\boldsymbol{Q}_{i+1/2}^{\text{Roe}}[\![\boldsymbol{U}]\!]_{i+1/2},$$

where $\boldsymbol{Q}^{\text{Roe}}$ is a positive semi-definite matrix defined through a diagonalization of the interfacial flux Jacobian at a Roe-averaged state:

$$\boldsymbol{Q}_{i+\frac{1}{2}}^{\text{Roe}} \coloneqq \boldsymbol{T}^{\text{Roe}}\left|\boldsymbol{\Lambda}^{\text{Roe}}\right|\left(\boldsymbol{T}^{\text{Roe}}\right)^{-1}, \qquad \frac{\partial\widehat{F}}{\partial\widehat{U}}(\overline{\boldsymbol{U}}_{i+1/2}) = \boldsymbol{T}^{\text{Roe}}\boldsymbol{\Lambda}^{\text{Roe}}\left(\boldsymbol{T}^{\text{Roe}}\right)^{-1}. \tag{4.20}$$

Then the semi-discrete scheme (4.3) using the numerical flux $\mathcal{F}_{i+1/2} = \mathcal{F}^{\mathrm{RD}}_{i+1/2}$ would behave like,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}_i(t) &= -\frac{1}{\Delta x}\left(\mathcal{F}^{\mathrm{EC}}_{i+1/2} - \mathcal{F}^{\mathrm{EC}}_{i-1/2}\right) + \frac{1}{2\Delta x}\left(\boldsymbol{Q}^{\mathrm{Roe}}_{i+1/2}[\![\boldsymbol{U}]\!]_{i+1/2} - \boldsymbol{Q}^{\mathrm{Roe}}_{i-1/2}[\![\boldsymbol{U}]\!]_{i-1/2}\right) + \boldsymbol{S}_i \\ &\approx -\frac{1}{\Delta x}\left(\widehat{F}\left(\widehat{U}\right)\big|_{x=x_{i+1/2}} - \widehat{F}\left(\widehat{U}\right)\big|_{x=x_{i-1/2}}\right) + \Delta x\, \boldsymbol{Q}\widehat{U}_{xx}\big|_{x=x_i} + S\left(\widehat{U}\right)\big|_{x=x_i},\end{aligned}$$

where $\boldsymbol{Q}$ is a positive-definite matrix and $\widehat{U}_{xx}$ is the second spatial derivative of the PDE state variables introduced in equation (2.12), and hence this introduces diffusion into an EC scheme. While the above approach works in terms of adding a diffusion-like term, a convenient way to ensure energy stability is to employ a numerical diffusion term that operates on the entropic variables $\boldsymbol{V}$ instead of the conservative variables $\boldsymbol{U}$:

$$\mathcal{F}^{\mathrm{ES}}_{i+\frac{1}{2}} := \mathcal{F}^{\mathrm{EC}}_{i+\frac{1}{2}} - \frac{1}{2}\boldsymbol{Q}^{\mathrm{ES}}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}}, \tag{4.21}$$

where $\boldsymbol{Q}^{\mathrm{ES}}_{i+\frac{1}{2}}$ is a positive definite matrix that will be identified in a Roe-type way from the two adjacent states $\boldsymbol{U}_i$ and $\boldsymbol{U}_{i+1}$ at the cell interface $x = x_{i+\frac{1}{2}}$. The term $\boldsymbol{V}_i$ is as given in (4.5b), and is a second-order approximation to the cell-average of the entropy variable $\widehat{V}$. We are interested in the Roe-type energy-stable operator defined as,

$$\mathcal{Q}_{i+1/2}(\boldsymbol{U}_i, \boldsymbol{U}_{i+1}) := \boldsymbol{T}|\boldsymbol{\Lambda}|\boldsymbol{T}^\top \geq 0, \tag{4.22}$$

where the matrices $\boldsymbol{T}$ and $\boldsymbol{\Lambda}$ are matrices from the eigendecomposition of the flux Jacobian (2.14) evaluated at a Roe-type average state:

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

Note in particular that $\overline{\boldsymbol{q}}_{i+1/2} \neq \mathcal{P}(\overline{\boldsymbol{h}}_{i+1/2})\overline{\boldsymbol{u}}_{i+1/2}$, so that $\widetilde{\boldsymbol{U}}_{i+1/2} \neq \overline{\boldsymbol{U}}_{i+1/2}$. The focal scheme of this section uses the numerical flux (4.21), where $\boldsymbol{Q}$ is given by the Roe-type diffusion matrix introduced above,

$$\boldsymbol{Q}^{\mathrm{ES1}}_{i+\frac{1}{2}} := \mathcal{Q}_{i+1/2}(\boldsymbol{U}_i, \boldsymbol{U}_{i+1}) = \boldsymbol{T}|\boldsymbol{\Lambda}|\boldsymbol{T}^\top, \tag{4.24}$$

where we refer to this scheme as "ES1" because we will show it is first-order accurate. Our main result for this scheme is as follows.

Theorem 4.2 (ES1 scheme)**.** *Consider the finite volume scheme* (4.3) *with source term* (4.12b) *and diffusive numerical flux* (4.21)*, selecting the diffusion matrix as,*

$$\boldsymbol{Q}^{\mathrm{ES}}_{i+1/2} = \boldsymbol{Q}^{\mathrm{ES1}}_{i+1/2}. \tag{4.25}$$

The resulting scheme is a first-order, well-balanced ES scheme.

Proof. We omit some details that are similar to the proof of Theorem 4.1. We have already established in Theorem 4.1 that $\mathcal{F}^{\mathrm{EC}}_{i+1/2}$ is second-order accurate. That this ES1 scheme is first-order is direct from the definition of $\boldsymbol{V}_i$ in (4.5b), resulting in the approximation

$$[\![\boldsymbol{V}]\!]_{i+1/2} \approx \Delta x \widehat{V}_x(x_{i+1/2})$$

which implies that the diffusive augmentation in (4.21) commits a first-order local truncation error.

To establish that this scheme is well-balanced, we assume the stochastic lake-at-rest initial data (4.13), and this coupled with the definition of $\boldsymbol{V}_i$ in (4.5b) implies $[\![\boldsymbol{V}]\!]_{i+1/2} = 0$. Since the EC flux and source are well-balanced (Lem. 4.2), then this implies that this ES1 scheme is also well-balanced.

Finally, we seek to show the ES property. We define the ES1 energy flux,

$$\mathcal{H}^{\mathrm{ES1}}_{i+\frac{1}{2}} = \mathcal{H}_{i+\frac{1}{2}} - \frac{1}{2}\overline{\boldsymbol{V}}^{\top}_{i+\frac{1}{2}}\boldsymbol{Q}^{\mathrm{ES1}}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}}$$

with $\mathcal{H}_{i+1/2}$ as defined in (4.18). As in Lemma 4.4, we multiply (4.3) by $\boldsymbol{V}_i^{\top}$; after manipulations that are similar to those in the proof of Lemma 4.4, we have,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}_i(t) = -\frac{1}{\Delta x}\Big(\mathcal{H}^{\mathrm{ES1}}_{i+\frac{1}{2}} - \mathcal{H}^{\mathrm{ES1}}_{i-\frac{1}{2}}\Big) - \frac{1}{4\Delta x}\Big([\![\boldsymbol{V}]\!]^{\top}_{i+\frac{1}{2}}\boldsymbol{Q}^{\mathrm{ES1}}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}} + [\![\boldsymbol{V}]\!]^{\top}_{i-\frac{1}{2}}\boldsymbol{Q}^{\mathrm{ES1}}_{i-\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i-\frac{1}{2}}\Big).$$

Since $\boldsymbol{Q}^{\mathrm{ES1}}_{i+1/2}$ is positive semi-definite, then this scheme satisfies (4.11), and hence is an ES scheme. □

4.4. ES1 diffusion *vs.* Roe diffusion

We provide in this section a result that motivates and justifies our particular form of the ES1 diffusion modification defined in (4.24) and (4.25). This result states that if the bottom topography function vanishes (*i.e.*, we are in the specialized case of a conservation law), then our chosen Roe-type ES1 diffusion in (4.22) and (4.23) coincides with a standard Roe-type diffusion term. Hence, in specialized scenarios our diffusive augmentations using entropic variables are equivalent to more standard Roe-type diffusion.

Proposition 4.1. *Define the Roe diffusion matrix as in* (4.20)*, but using the flux Jacobian evaluated at* $\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}$*,*

$$\frac{\partial \widehat{F}}{\partial \widehat{U}}\Big(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}\Big) = \boldsymbol{T}^{\mathrm{Roe}}\boldsymbol{\Lambda}^{\mathrm{Roe}}\Big(\boldsymbol{T}^{\mathrm{Roe}}\Big)^{-1} \tag{4.26}$$

where we have evaluated the flux jacobian at $\widetilde{\boldsymbol{U}}_{i+1/2}$ *instead of at* $\overline{\boldsymbol{U}}_{i+1/2}$*. Assume* $\boldsymbol{B}_i = 0$ *for all* $i \in [M]$*. Then,*

$$\boldsymbol{Q}^{\mathrm{Roe}}_{i+\frac{1}{2}}[\![\boldsymbol{U}]\!]_{i+\frac{1}{2}} = \boldsymbol{Q}^{\mathrm{ES1}}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}}. \tag{4.27}$$

Proving this result requires some setup: Under the assumptions of Proposition 4.1 we consider the SG SWE (2.12) with flat bottom, *i.e.*, $\widehat{B} = \boldsymbol{0}$, together with entropy $E^{\mathrm{flat}}(\widehat{U}) = \frac{1}{2}(\widehat{q})^{\top}\widehat{u} + \frac{g}{2}\|\widehat{h}\|^2$ and the entropy variables,

$$\widehat{V}^{\mathrm{flat}} = \partial_{\widehat{U}}E = \begin{pmatrix} -\frac{1}{2}\mathcal{P}(\widehat{u})\widehat{u} + g\widehat{h} \\ \widehat{u} \end{pmatrix}. \tag{4.28}$$

Our main tool will be some results of the proof of Theorem 3.1 in [11]; in particular, while we have provided the flux Jacobian for this system in (2.14), we will need the explicit similarity transform that accomplishes its symmetrization.

Lemma 4.5 ([11], Thm. 3.1)**.** *Assume* $\mathcal{P}(\widehat{h}) > 0$*. Define* $G = \sqrt{g\mathcal{P}(\widehat{h})}$ *as the positive definite square root matrix of* $g\mathcal{P}(\widehat{h})$*. Then,*

$$\frac{\partial \widehat{F}}{\partial \widehat{U}}\Big(\widehat{U}\Big) = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1},$$

where $\boldsymbol{D}$ *is the symmetric matrix,*

$$\boldsymbol{D}\Big(\widehat{U}\Big) = \frac{1}{2}\begin{pmatrix} 2G + \mathcal{P}(\widehat{u}) + gG^{-1}\mathcal{P}(\widehat{q})G^{-1} & \mathcal{P}(\widehat{u}) - gG^{-1}\mathcal{P}(\widehat{q})G^{-1} \\ \mathcal{P}(\widehat{u}) - gG^{-1}\mathcal{P}(\widehat{q})G^{-1} & \mathcal{P}(\widehat{u}) + gG^{-1}\mathcal{P}(\widehat{q})G^{-1} - 2G \end{pmatrix}, \tag{4.29}$$

and

$$\boldsymbol{R}\Big(\widehat{U}\Big) = \frac{1}{\sqrt{2g}}\begin{pmatrix} I & I \\ \mathcal{P}(\widehat{u}) + \sqrt{g\mathcal{P}\Big(\widehat{h}\Big)} & \mathcal{P}(\widehat{u}) - \sqrt{g\mathcal{P}\Big(\widehat{h}\Big)} \end{pmatrix}. \tag{4.30}$$

The second lemma reveals the relation between the cell interface jump of $\boldsymbol{V}^{\text{flat}}$ (the spatial approximation corresponding to the cell-averaged entropy variable $\widehat{V}^{\text{flat}}$ in (4.28)) and $\boldsymbol{U}$.

Lemma 4.6. *Recall the definition of* $\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}$ *in* (4.23)*:*

$$\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} = \begin{pmatrix} \overline{\boldsymbol{h}}_{i+\frac{1}{2}} \\ \mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)\overline{\boldsymbol{u}}_{i+\frac{1}{2}} \end{pmatrix}$$

which is an intermediate state defined by the arithmetic average of $\boldsymbol{h}$ *and* $\boldsymbol{u}$ *across the cell interface* $x = x_{i+\frac{1}{2}}$*. Denote,* $\boldsymbol{V}^{\text{flat}}$ *to be the corresponding spatial approximation of the cell-averaged entropy variable defined in* (4.28)*. Then,*

(1) *The jump* $[\![\boldsymbol{U}]\!]_{i+\frac{1}{2}}$ *is a rescaling of the jump* $\left[\!\left[\boldsymbol{V}^{\text{flat}}\right]\!\right]_{i+\frac{1}{2}}$*, i.e.,*

$$[\![\boldsymbol{U}]\!]_{i+\frac{1}{2}} = \left(\boldsymbol{V}_{\boldsymbol{U}}^{\text{flat}}\right)_{i+\frac{1}{2}} \left[\!\left[\boldsymbol{V}^{\text{flat}}\right]\!\right]_{i+\frac{1}{2}}, \tag{4.31}$$

where

$$\left(\boldsymbol{V}_{\boldsymbol{U}}^{\text{flat}}\right)_{i+\frac{1}{2}} := \frac{1}{g}\begin{pmatrix} I & \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) \\ \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) & \mathcal{P}^2\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) + g\mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right) \end{pmatrix},$$

$$\left[\!\left[\boldsymbol{V}^{\text{flat}}\right]\!\right]_{i+1/2} \overset{(4.28)}{=} \begin{pmatrix} -\frac{1}{2}[\![\mathcal{P}(\boldsymbol{u})\boldsymbol{u}]\!]_{i+\frac{1}{2}} + g[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} \\ [\![\boldsymbol{u}]\!]_{i+\frac{1}{2}} \end{pmatrix}.$$

(2) *Let* $\boldsymbol{R}_{i+\frac{1}{2}}$ *denote the matrix that symmetrizes the flux Jacobian at the state* $\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}$*,*

$$\boldsymbol{R}_{i+\frac{1}{2}} := \boldsymbol{R}\left(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}\right) = \frac{1}{\sqrt{2g}}\begin{pmatrix} I & I \\ \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) + \sqrt{g\mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)} & \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) - \sqrt{g\mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)} \end{pmatrix},$$

cf. (4.30)*. Then,*

$$\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{R}_{i+\frac{1}{2}}^{\top} = \left(\boldsymbol{V}_{\boldsymbol{U}}^{\text{flat}}\right)_{i+\frac{1}{2}}. \tag{4.32}$$

Proof. Part (2), *i.e.*, (4.32), is a straightforward matrix algebra calculation that we omit. For part (1), we first recall that (4.8a) implies,

$$\frac{1}{2}[\![\mathcal{P}(\boldsymbol{u})\boldsymbol{u}]\!]_{i+\frac{1}{2}} = \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}}. \tag{4.33}$$

Second, we use the linearity of $\mathcal{P}(\cdot)$, the property (4.6) for arithmetic averages, and the commutation property (2.9), to conclude,

$$\mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} + \mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}} = [\![\mathcal{P}(\boldsymbol{h})\boldsymbol{u}]\!]_{i+\frac{1}{2}} = [\![\boldsymbol{q}]\!]_{i+\frac{1}{2}}. \tag{4.34}$$

Therefore,

$$\left(\boldsymbol{V}_{\boldsymbol{U}}^{\text{flat}}\right)_{i+\frac{1}{2}}\left[\!\left[\boldsymbol{V}^{\text{flat}}\right]\!\right]_{i+\frac{1}{2}} = \frac{1}{g}\begin{pmatrix} -\frac{1}{2}[\![\mathcal{P}(\boldsymbol{u})\boldsymbol{u}]\!]_{i+\frac{1}{2}} + g[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} + \mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}} \\ -\frac{1}{2}\mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right)[\![\mathcal{P}(\boldsymbol{u})\boldsymbol{u}]\!]_{i+\frac{1}{2}} + g\mathcal{P}\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right)[\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} + \left(\mathcal{P}^2\left(\overline{\boldsymbol{u}}_{i+\frac{1}{2}}\right) + g\mathcal{P}\left(\overline{\boldsymbol{h}}_{i+\frac{1}{2}}\right)\right)[\![\boldsymbol{u}]\!]_{i+\frac{1}{2}} \end{pmatrix}$$

$$\overset{(4.33)(4.34)}{=\!=} \begin{pmatrix} [\![\boldsymbol{h}]\!]_{i+\frac{1}{2}} \\ [\![\boldsymbol{q}]\!]_{i+\frac{1}{2}} \end{pmatrix} = [\![\boldsymbol{U}]\!]_{i+\frac{1}{2}}.$$

□

Now we are in position to show (4.27) in Proposition 4.1.

Proof of Proposition 4.1. Let $\boldsymbol{D}_{i+\frac{1}{2}}$ be the *symmetric* matrix defined in (4.29) evaluated at $\boldsymbol{U}_{i+\frac{1}{2}}$, and $\boldsymbol{D}_{i+\frac{1}{2}} = \boldsymbol{L}_{i+\frac{1}{2}}\boldsymbol{\Lambda}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}^\top$ be its eigenvalue decomposition. Then,

$$\frac{\partial \widehat{F}}{\partial \widehat{U}}\left(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}\right) = \boldsymbol{R}_{i+\frac{1}{2}}\left(\boldsymbol{L}_{i+\frac{1}{2}}\boldsymbol{\Lambda}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}^\top\right)\boldsymbol{R}_{i+\frac{1}{2}}^{-1} \tag{4.35a}$$

$$=: \boldsymbol{T}_{i+\frac{1}{2}}\boldsymbol{\Lambda}_{i+\frac{1}{2}}\boldsymbol{T}_{i+\frac{1}{2}}^{-1}, \tag{4.35b}$$

is an eigendecomposition of the Jacobian matrix $\frac{\partial \widehat{F}}{\partial \widehat{U}}(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}})$, where we have used the fact that $\boldsymbol{L}_{i+\frac{1}{2}}^{-1} = \boldsymbol{L}_{i+\frac{1}{2}}^\top$ due to the symmetry of $\boldsymbol{D}_{i+\frac{1}{2}}$. The Roe-diffusion operator evaluated at the location $\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}$ as indicated in (4.26) is then given by,

$$\boldsymbol{Q}_{i+\frac{1}{2}}^{\text{Roe}} \overset{(4.20)}{=} \boldsymbol{T}_{i+\frac{1}{2}}\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\boldsymbol{T}_{i+\frac{1}{2}}^{-1}.$$

Therefore,

$$\begin{aligned}
\boldsymbol{Q}_{i+\frac{1}{2}}^{\text{Roe}}[\![\boldsymbol{U}]\!]_{i+\frac{1}{2}} &= \boldsymbol{T}_{i+\frac{1}{2}}\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\boldsymbol{T}_{i+\frac{1}{2}}^{-1}[\![\boldsymbol{U}]\!]_{i+\frac{1}{2}}, \\
&\overset{(4.35),(4.31)}{=} \left(\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}\right)\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\left(\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}\right)^{-1}\left(\boldsymbol{V}_{\boldsymbol{U}}^{\text{flat}}\right)_{i+\frac{1}{2}}[\![\boldsymbol{V}^{\text{flat}}]\!]_{i+\frac{1}{2}}, \\
&\overset{(4.32)}{=} \left(\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}\right)\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\left(\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{L}_{i+\frac{1}{2}}\right)^{-1}\boldsymbol{R}_{i+\frac{1}{2}}\boldsymbol{R}_{i+\frac{1}{2}}^\top[\![\boldsymbol{V}^{\text{flat}}]\!]_{i+\frac{1}{2}}, \\
&= \boldsymbol{R}_{i+\frac{1}{2}}\left(\boldsymbol{L}_{i+\frac{1}{2}}\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\boldsymbol{L}_{i+\frac{1}{2}}^\top\right)\boldsymbol{R}_{i+\frac{1}{2}}^\top[\![\boldsymbol{V}^{\text{flat}}]\!]_{i+\frac{1}{2}}, \\
&= \boldsymbol{T}_{i+\frac{1}{2}}\left|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\right|\boldsymbol{T}_{i+\frac{1}{2}}^\top[\![\boldsymbol{V}^{\text{flat}}]\!]_{i+\frac{1}{2}} \\
&\overset{(4.24)}{=} \boldsymbol{Q}_{i+\frac{1}{2}}^{\text{ES1}}[\![\boldsymbol{V}^{\text{flat}}]\!]_{i+\frac{1}{2}}.
\end{aligned} \tag{4.36}$$

□

4.5. A second-order ES scheme

To develop a second-order accurate energy-stable scheme, we use jump operators with $O(\Delta x^2)$ accuracy. A natural choice is to use the jumps obtained by non-oscillatory second-order reconstructions of the entropy variable. However, attaining a provable energy-stable scheme requires the more subtle reconstruction procedure in [18] that we follow. The new idea for second-order diffusion is to use reconstructions in order to compute jumps. To that end, we let $\boldsymbol{V}_i^+$ and $\boldsymbol{V}_{i+1}^-$ be second-order reconstructions from the right and left, respectively, of the entropy variable $\boldsymbol{V}(x)$ at location $x = x_{i+1/2}$. We will describe later in this section how these reconstructions are computed.

Assuming we have these reconstructions in hand, we can compute second-order accurate jumps of the entropy variables:

$$\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2}} = \boldsymbol{V}_{i+1}^- - \boldsymbol{V}_i^+. \tag{4.37}$$

The overall scheme is similar as the previous section, but uses a second-order diffusive augmentation of a conservative flux,

$$\mathcal{F}_{i+\frac{1}{2}}^{\text{ES2}} := \mathcal{F}_{i+\frac{1}{2}}^{\text{EC}} - \frac{1}{2}\boldsymbol{Q}_{i+\frac{1}{2}}^{\text{ES2}}\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2}}. \tag{4.38}$$

We choose the matrix $\boldsymbol{Q}^{\mathrm{ES2}}$ as for the ES1 scheme,

$$\boldsymbol{Q}^{\mathrm{ES2}}_{i+1/2} = \boldsymbol{Q}^{\mathrm{ES1}}_{i+1/2} = \mathcal{Q}_{i+1/2}(\boldsymbol{U}_i, \boldsymbol{U}_{i+1}) = \boldsymbol{T}_{i+1/2}|\boldsymbol{\Lambda}_{i+1/2}|\boldsymbol{T}^{\top}_{i+1/2}, \tag{4.39}$$

where we recall that the eigendecomposition matrices $\boldsymbol{T}$, $\boldsymbol{\Lambda}$ are computed from the Roe-type average of the flux Jacobian, cf. (4.22), (4.23). One could alternatively select $\boldsymbol{Q}^{\mathrm{ES2}}$ by using second-order reconstructions of $\boldsymbol{U}$ as input to $\mathcal{Q}$, *e.g.*,

$$\boldsymbol{Q}^{\mathrm{ES2}}_{i+1/2} = \mathcal{Q}_{i+1/2}\big(\boldsymbol{U}^{-}_i, \boldsymbol{U}^{+}_i\big),$$

for some second-order reconstructions $\boldsymbol{U}^{\pm}_i$.

What remain is to describe how $\boldsymbol{V}^{\pm}_i$ are computed in a way that ensures the energy stable property. The main idea is to design $\boldsymbol{V}^{\pm}_i$ through a second-order reconstruction of *scaled* (transformed) versions of the entropy variables:

$$\boldsymbol{w}^{\pm}_i \coloneqq \boldsymbol{T}^{\top}_{i\pm1/2}\boldsymbol{V}_i, \tag{4.40}$$

where the matrices $\boldsymbol{T}_{i\pm1/2}$ are as in (4.39). Once these have been computed, we perform a second-order total variation-diminishing (TVD) reconstruction on the $\boldsymbol{w}$ variable at the interfaces:

$$\widetilde{\boldsymbol{w}}^{\pm}_i \coloneqq \boldsymbol{w}^{\pm}_i \pm \frac{1}{2}\phi\big(\boldsymbol{\theta}^{\pm}_i\big) \circ \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i\pm1/2}, \tag{4.41}$$

where $\circ$ is the Hadamard (elementwise) product on vectors, and $\boldsymbol{\theta}^{\pm}_i$ are difference quotients,

$$\boldsymbol{\theta}^{\pm}_i \coloneqq \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i\mp1/2} \oslash \langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i\pm1/2},$$

where $\oslash$ is the Hadamard (elementwise) division between vectors. We select the function ϕ to be the minmod limiter,

$$\phi(\theta) = \begin{cases} 0, & \text{if} \quad \theta < 0, \\ \theta. & \text{if} \quad 0 \le \theta \le 1, \\ 1, & \text{otherwise} \end{cases} \tag{4.42}$$

which operates elementwise on vector inputs. Note that other slope limiter functions ϕ may be selected, but minmod is the only valid limiter in this context that also satisfies the TVD property ([18], Sect. 3.4). Finally, the desired reconstructions for $\boldsymbol{V}^{\pm}_i$ are defined by inverting the $\boldsymbol{w}$-to-$\boldsymbol{V}$ map,

$$\boldsymbol{T}^{\top}_{i\pm1/2}\boldsymbol{V}^{\pm}_i \coloneqq \widetilde{\boldsymbol{w}}^{\pm}_i. \tag{4.43}$$

The full scheme has now been described, and satisfies the following properties.

Theorem 4.3 (ES2 scheme)**.** *The FV scheme* (4.3) *choosing the flux* $\mathcal{F}_{i+1/2} = \mathcal{F}^{\mathrm{ES2}}_{i+1/2}$ *defined in* (4.38) *is a second-order, well-balanced, ES scheme.*

We focus the remaining discussion in this section on sketching the proof of the above result. The second-order property results from the fact that the jumps are computed using second-order accurate reconstructions; the well-balanced property can be proven in exactly the same way as is done for the ES1 scheme in the proof of Theorem 4.2. To show the ES property, we exercise one of the major results in [18] that we reproduce below.

Lemma 4.7 ([18], Lem. 3.2)**.** *For each* i*, if there exists a positive diagonal matrix* $\boldsymbol{\Pi}_{i+1/2} \ge 0$ *such that the second-order jump satisfies,*

$$\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2}} = \Big(\boldsymbol{T}^{\top}_{i+\frac{1}{2}}\Big)^{-1}\boldsymbol{\Pi}_{i+\frac{1}{2}}\boldsymbol{T}^{\top}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}}, \tag{4.44}$$

then the scheme (4.3) *with flux term* $\mathcal{F}_{i+1/2} = \mathcal{F}^{\mathrm{ES2}}_{i+1/2}$ *is an ES scheme.*

Hence, showing the ES property for our scheme only requires us to establish (4.44). To accomplish this, note that the definition (4.41) implies,

$$\langle\!\langle \widetilde{\boldsymbol{w}} \rangle\!\rangle^{\ell}_{i+\frac{1}{2}} = \left(1 - \frac{1}{2}\phi\Big(\big(\boldsymbol{\theta}^{-}_{i+1}\big)^{\ell}\Big) - \frac{1}{2}\phi\Big(\big(\boldsymbol{\theta}^{+}_{i}\big)^{\ell}\Big)\right)\langle\!\langle \boldsymbol{w} \rangle\!\rangle^{\ell}_{i+\frac{1}{2}}. \tag{4.45}$$

I.e., we have,

$$\langle\!\langle \widetilde{\boldsymbol{w}} \rangle\!\rangle_{i+\frac{1}{2}} = \boldsymbol{\Pi}_{i+\frac{1}{2}}\langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i+\frac{1}{2}}, \qquad \big(\boldsymbol{\Pi}_{i+1/2}\big)_{\ell,\ell} \coloneqq \left(1 - \frac{1}{2}\phi\big(\big(\boldsymbol{\theta}^{-}_{i+1}\big)_{\ell}\big) - \frac{1}{2}\phi\big(\big(\boldsymbol{\theta}^{+}_{i}\big)_{\ell}\big)\right) \tag{4.46}$$

and in particular $\boldsymbol{\Pi}_{i+1/2}$ is a diagonal matrix and positive semi-definite since $0 \leq \phi(\theta) \leq 1$. Since the jump operators $\langle\!\langle\cdot\rangle\!\rangle$ and $[\![\cdot]\!]$ are linear in their arguments, then combining (4.45) with the relations (4.40) and (4.43) that connect $\boldsymbol{w}_i$ and $\widetilde{\boldsymbol{w}}_i$ to $\boldsymbol{V}_i$ and $\boldsymbol{V}^{\pm}_i$ yields the relation (4.44) with a positive-definite diagonal matrix $\boldsymbol{\Pi}_{i+1/2}$. Hence, this is an ES scheme, and completes the proof of Theorem 4.3.

Finally, we remark that the implementation of the diffusion term in the ES2 flux (4.38) does not require explicit construction of $\boldsymbol{V}^{\pm}_i$. *I.e.*, we have,

$$\begin{aligned}
\frac{1}{2}\boldsymbol{Q}^{\mathrm{ES2}}_{i+\frac{1}{2}}\langle\!\langle \boldsymbol{V} \rangle\!\rangle_{i+\frac{1}{2}} &\overset{(4.22),(4.44)}{=} \frac{1}{2}\boldsymbol{T}_{i+\frac{1}{2}}\Big|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\Big|\boldsymbol{T}^{\top}_{i+\frac{1}{2}}\Big(\boldsymbol{T}^{\top}_{i+\frac{1}{2}}\Big)^{-1}\boldsymbol{\Pi}_{i+\frac{1}{2}}\boldsymbol{T}^{\top}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}} \\
&= \frac{1}{2}\boldsymbol{T}_{i+\frac{1}{2}}\Big|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\Big|\boldsymbol{\Pi}_{i+\frac{1}{2}}\boldsymbol{T}^{\top}_{i+\frac{1}{2}}[\![\boldsymbol{V}]\!]_{i+\frac{1}{2}} \\
&\overset{(4.40)}{=} \frac{1}{2}\boldsymbol{T}_{i+\frac{1}{2}}\Big|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\Big|\boldsymbol{\Pi}_{i+\frac{1}{2}}\langle\!\langle \boldsymbol{w} \rangle\!\rangle_{i+\frac{1}{2}}, \\
&\overset{(4.46)}{=} \frac{1}{2}\boldsymbol{T}_{i+\frac{1}{2}}\Big|\boldsymbol{\Lambda}_{i+\frac{1}{2}}\Big|\langle\!\langle \widetilde{\boldsymbol{w}} \rangle\!\rangle_{i+\frac{1}{2}},
\end{aligned}$$

and hence one need only compute $\widetilde{\boldsymbol{w}}^{\pm}_i$ in order to directly evaluate the diffusion part of the ES2 flux.

4.6. Algorithmic details

Our overall scheme is the semi-discrete form (4.3), which we pair with a numerical time-stepping scheme. We provide pseudocode in this section that describes a fully discrete SG SWE time-stepping algorithm. This full pseudocode introduces some additional details for the scheme that were devised in [11], many of which are based on standard procedures used in schemes for deterministic SWE models [23]. We very briefly describe these additional details in the coming sections; more comprehensive discussion can be found in [11]. The full algorithmic pseudocode is given in Algorithm 1.

4.6.1. Velocity desingularization

Computing $\boldsymbol{u}_i$ requires inversion of the matrix $\mathcal{P}(\boldsymbol{h}_i)$, which is assumed (and enforced in the scheme) to be symmetric and positive-definite. However, this matrix may be ill-conditioned. To ameliorate numerical artifacts associated with this ill-conditioned operation, we employ a *desingularization* procedure, introduced for the deterministic SWE in [25]. We describe here the stochastic variant of the desingularization procedure, proposed in [11]. If $\mathcal{P}(\boldsymbol{h}_i)$ has the eigenvalue decomposition,

$$\mathcal{P}(\boldsymbol{h}_i) = \boldsymbol{Q}\boldsymbol{\Pi}\boldsymbol{Q}^{\top}, \qquad \boldsymbol{\Pi} = \mathrm{diag}(\pi_1, \ldots, \pi_K),$$

where $\pi_k > 0$ are the eigenvalues of $\mathcal{P}(\boldsymbol{h}_i)$, then the desingularization process approximates $\mathcal{P}(\boldsymbol{h}_i)^{-1}\boldsymbol{q}_i$ by regularizing the matrix inverse procedure:

$$\boldsymbol{u}_i = \boldsymbol{Q}\widetilde{\boldsymbol{\Pi}}^{-1}\boldsymbol{Q}^{T}\boldsymbol{q}_i, \qquad \widetilde{\boldsymbol{\Pi}} = \mathrm{diag}(\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K), \qquad \widetilde{\pi}_k = \frac{\sqrt{\pi_k^4 + \max\{\pi_k^4, \epsilon^4\}}}{\sqrt{2}\pi_k}, \tag{4.47}$$

where $\epsilon > 0$ is a small constant; we choose it to be $\epsilon = \Delta x$. Note that if $\pi_k \geq \epsilon^{1/4}$, then $\widetilde{\pi}_k = \pi_k$, and hence regularization is performed only in the presence of small eigenvalues. Compared to (4.4). This procedure to compute $\boldsymbol{u}_i$ is a stabilized way to compute velocities.

For scheme consistency, if the desingularization above is activated, then we recompute the discharge variable:

$$\boldsymbol{q}_i \leftarrow \mathcal{P}(\boldsymbol{h}_i)\boldsymbol{u}_i.$$

This update for $\boldsymbol{q}_i$ may affect global conservation of $\boldsymbol{q}$ through a $\mathcal{O}(\epsilon^2/\pi_k) \sim \mathcal{O}(\Delta x^2/\pi_k)$ augmentation; however, this desingularization process is constructed to minimize the numerical effects of such desingularization while simultaneously restoring well-behaved and stable computation. Without such a desingularization procedure, the scheme can quickly become numerically unstable for small water heights.

4.6.2. Hyperbolicity preservation

The SG SWE PDE (2.12) is hyperbolic and has an entropy pair if $\mathcal{P}(\widehat{h}) > 0$, *i.e.*, Theorems 2.1 and 3.1, respectively. To ensure this holds at the discrete level, we require the condition $\mathcal{P}(\boldsymbol{h}_i) > 0$ for every cell i. To enforce this, we employ ([11], Thm. 3.4, Cor. 3.5), which state that a sufficient condition for $\mathcal{P}(\boldsymbol{h}_i) > 0$ is that for every $m = 1, \ldots, M$,

$$\widehat{h}_i(\xi_m) > 0, \qquad \widehat{h}_i(\xi) \coloneqq \sum_{k=1}^{K} \boldsymbol{h}_{i,k}\phi_k(\xi_m), \qquad \boldsymbol{h}_i = (\boldsymbol{h}_{i,1}, \ldots, \boldsymbol{h}_{i,K})^\top, \tag{4.48}$$

where $\{\xi_m\}_{m=1}^M$ is a nodal set in $\mathbb{R}^d$ for a positive-weight quadrature rule having sufficient accuracy relative to the ξ-polynomial space P defined in (2.4). The functions ϕ_k are the basis of P in (2.4) for which $\boldsymbol{h}_i$ are coordinates. The function $\widehat{h}_i(\xi)$ is the SG SWE approximation to the $\mathcal{I}_i$-cell average of $\widehat{h}(x, t, \xi)$ at the current time. Hence, the computational vehicle we use to enforce hyperbolicity of the underlying PDE in our scheme is to enforce the above positivity-type condition on the $\boldsymbol{h}_i$ variable.

4.6.3. Positivity-preservation

We enforce the positivity condition (4.48) by restricting the timestep size. We assume that the current time value of $\boldsymbol{h}_i$ satisfies (4.48). If forward Euler with a stepsize Δt is used to discretize (4.3), then (4.48) is true at the next time step if,

$$\Delta t < \lambda \coloneqq \min_i \min_{m=1,\ldots,M} \left| \frac{\Delta x\, \widehat{h}_i(\xi_m)}{\widehat{\mathcal{F}}^h_{i+1/2}(\xi_m) - \widehat{\mathcal{F}}^h_{i-1/2}(\xi_m)} \right|, \tag{4.49}$$

where $\widehat{\mathcal{F}}^h_{i+1/2}(\cdot)$ is the SG approximation of the $\widehat{h}$-variable flux:

$$\widehat{\mathcal{F}}^h_{i+1/2}(\xi) \coloneqq \sum_{k=1}^{K} \mathcal{F}^h_{i+1/2,k}\phi_k(\xi), \qquad \mathcal{F}_{i+1/2} = \left(\left(\mathcal{F}^h_{i+1/2}\right)^\top, \left(\mathcal{F}^q_{i+1/2}\right)^\top \right)^\top \in \mathbb{R}^{2K}.$$

Hence, we enforce positivity preservation by ensuring a small enough timestep so that the positivity condition (4.48) is respected globally over all spatial cells. We must also restrict Δt to satisfy the wave speed CFL condition; see equation (4.16) of [11].

4.6.4. Adaptive time-stepping

The time step restriction (4.49) works for forward Euler time-stepping. To extend this to a higher-order temporal scheme, we employ a third-order strong stability-preserving scheme, which is a convex combination of forward Euler steps [21]. However, the intermediate stages of a(ny) time-stepping scheme need not obey the positivity-preserving property, even if Δt is chosen to obey the condition (4.49) determined at the initial step.

To address this issue, we employ the *adaptive* time-stepping strategy proposed in Remark 3.6 of [6]. We refer the reader to that reference for details, and present here only a high-level description of the procedure: λ is initialized as the initial stage value of λ, as shown in (4.49). At intermediate stages, new intermediate values of λ are computed. If an intermediate-stage value of λ is smaller than the current value of λ, then we restart the entire time-step using the new, smaller-λ restriction on Δt.

Algorithm 1. The fully discrete SG SWE schemes proposed in this paper; we ignore specifying the handling of boundary conditions.

Input scheme type: `scheme` = EC, ES1, or ES2
Input: Bottom topography B, initial data $U(t=0)$, polynomial index set Λ
Input: Terminal time T
Initialize: $\boldsymbol{U}_i$, $t = 0$
repeat
  Compute $\boldsymbol{B}_i$ from B for all i
  Compute $\boldsymbol{u}_i$ in (4.47) for all i
  Compute $\mathcal{F}^{\mathrm{EC}}_{i+1/2}$ for all i, given by (4.12)
  if `scheme` is EC **then** for all i:
    Set $\mathcal{F}_{i+1/2} \leftarrow \mathcal{F}^{\mathrm{EC}}_{i+1/2}$.
  else for all i:
    Compute entropy variable $\boldsymbol{V}_i$ using (4.5b).
    Compute $\boldsymbol{T}_{i+1/2}$, $\boldsymbol{\Lambda}_{i+1/2}$ through (4.23).
    if `scheme` is ES1 **then**:
      Compute $\boldsymbol{Q}^{\mathrm{ES1}}_{i+1/2}$ using (4.23),(4.24) with $\boldsymbol{T}_{i+1/2}$, $\boldsymbol{\Lambda}_{i+1/2}$.
      Compute $\mathcal{F}_{i+1/2} \leftarrow \mathcal{F}^{\mathrm{ES}}_{i+1/2}$ in (4.21) using $\mathcal{F}^{\mathrm{EC}}_{i+1/2}$, $\boldsymbol{V}_i$, and $\boldsymbol{Q}^{\mathrm{ES}}_{i+1/2} \leftarrow \boldsymbol{Q}^{\mathrm{ES1}}_{i+1/2}$.
    else if `scheme` is ES2 **then**:
      Construct $\boldsymbol{Q}^{\mathrm{ES2}}_{i+1/2}$ as in (4.39) with $\boldsymbol{T}_{i+1/2}$, $\boldsymbol{\Lambda}_{i+1/2}$.
      Construct $\boldsymbol{V}^{\pm}_i$ through (4.40), (4.41), and (4.43).
      Compute $\mathcal{F}_{i+1/2} \leftarrow \mathcal{F}^{\mathrm{ES2}}_{i+1/2}$ in (4.38) and (4.37) using $\boldsymbol{Q}^{\mathrm{ES2}}_{i+1/2}$, $\boldsymbol{V}^{\pm}_i$, and $\mathcal{F}^{\mathrm{EC}}_{i+1/2}$.
    end if
  end if
  Initialize λ and Δt as shown in (4.49).
  Adaptively determine Δt using the procedure discussed in Section 4.6.4.
  Use a third-order SSP method to take a time step of size Δt, updating $\boldsymbol{h}_i$ and $\boldsymbol{q}_i$.
  Set $t \leftarrow t + \Delta t$.
until $t \geq T$

5. Numerical experiments

Below we present several numerical examples to illustrate properties of the developed schemes. We refer to the second order energy-conservative scheme, the first order energy-stable scheme, and the second order energy-stable scheme as the EC, ES1, and ES2 schemes, respectively. We introduce the relative change in energy quantity,

$$\text{relative energy} = \frac{E(t) - E(0)}{E(t)}, \tag{5.1}$$

where $E(t)$ is computed as $\sum_i \Delta x \boldsymbol{E}_i(t)$. This provides a way to visualize the relative change in the discrete energy for different numerical schemes, namely for the EC, ES1 and ES2. In the examples of Section 5.1 through Section 5.4, we take the random variable ξ as a scalar parameter, uniformly distributed on $[-1, 1]$. We generally use $K = 9\,\mathrm{PC}$ terms (except in some accuracy tests) with the polynomial space P spanned by orthonormal

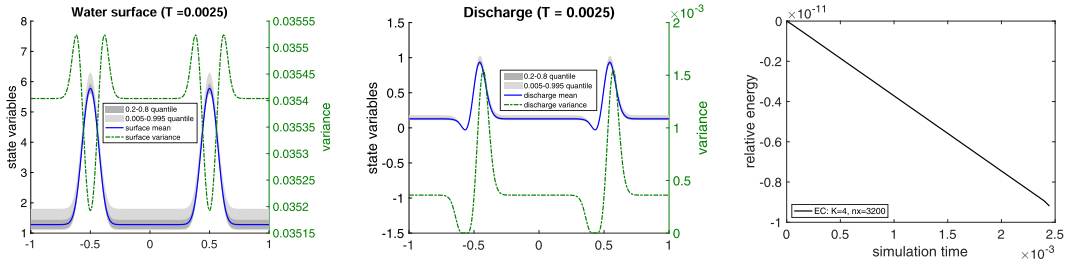

FIGURE 1. Results for Section 5.1. *Left*: water surface and *middle*: discharge. *Right*: relative energy change in EC over time. Mesh size $n_x = 3200$ and PC basis functions $K = 4$.

Legendre polynomials. Our final example in Section 5.5 employs a two-dimensional vector ξ, and the details of that experiment are provided in that section.

Instead of visualizing the conservative variable h corresponding to water height, we will plot the *water surface* w, defined as $w = h + B$, with the bottom topography B superimposed on the same graph; plots of (w, B) are more physically interpretable than directly plotting the water height h.

5.1. Accuracy test for the EC scheme on smooth initial data

We start by examining the accuracy of the proposed EC scheme (4.3), (4.12a), (4.12b) for the SG system (2.12) on a test with a smooth initial data and a smooth solution (we select a final time to ensure that the solution is smooth). On a smooth solution, the system satisfies the energy equality (3.5) and the EC scheme is constructed to satisfy the discrete version of the energy equality. We consider a smooth stochastic water surface,

$$h(x,0,\xi) + B(x,0,\xi) = w(x,0,\xi) = 1.1 + 0.1e^{-2\xi} + 0.001e^{-10\sin(\cos(2\pi x))}, \quad u(x,0,\xi) = 0.1, \tag{5.2}$$

with a flat bottom $B(x,t,\xi) \equiv 0$. For all tests in this section we compute the solution to a final time $t = 0.0025$ over the physical domain $x \in \mathcal{D} = [-1,1]$. The profile of the water surface and discharge computed by the EC scheme at a final time on a mesh of $n_x = 3200$ elements with $K = 4$ is shown in Figure 1 (water surface (left) and discharge (middle)). The right plot in Figure 1 illustrates that the EC scheme numerically preserves energy on this test – a very small numerical error of $O(10^{-12})$ in the energy is due to time discretization.

We next test spatial accuracy of the EC scheme against a reference solution computed by the EC scheme on mesh of size $n_x = 3200$ with $K = 4$, Figure 1. We illustrate convergence of the scheme using the water height h by computing the error between the reference solution and the numerical solution using an L^1 norm in physical space and an L^2 norm in parameter space,

$$\text{Error}(h_{n_x}) = \|h_{n_x}(x,t,\xi) - h_{\text{ref}}(x,t,\xi)\|_{L^1(\mathcal{D};L^2_\rho(\mathbb{R}^d))}, \tag{5.3a}$$

$$\|h(x,t,\xi)\|_{L^1(\mathcal{D};L^2_\rho(\mathbb{R}^d))} := \int_{\mathcal{D}} \|h(x,t,\xi)\|_{L^2_\rho}\,\mathrm{d}x \tag{5.3b}$$

where we recall that $\mathcal{D}$ is the physical domain and $\mathbb{R}^d$ is the stochastic domain, h_{n_x} is the numerical water height PC solution to the SG SWE equation on a mesh of size n_x, and h_{ref} is the K-term (fixed $K = 4$) PC reference solution. The results for this spatial convergence test are presented in Table 2a and show second order convergence for the smooth solution, as expected.

Our next test investigates convergence with respect to the number of PC terms K for one-dimensional parameter space, $d = 1$. For infinitely smooth solutions, we expect spectral (often exponential) convergence. For this test, the reference solution corresponds to $n_x = 6400$ with $K = 25$, and again illustrate convergence of the scheme using water height h by computing the error between the reference solution and the numerical solution

TABLE 2. Results for Section 5.1: EC order of convergence in physical space (A) for the problem (5.2) with smooth initial data. Error is computed using (5.3a). Order of convergence in stochastic space is presented in (B). Error is computed using (5.4).

(a) Spatial convergence

Mesh (n_x)	Error (EC)	Order
100	5.1347e−04	–
200	1.3184e−04	1.9615
400	3.3784e−05	1.9644
800	8.1163e−06	2.0574

(b) Stochastic convergence

K (Even)	Error (EC)	Order	K (Odd)	Error (EC)	Order
2	1.7406e−01	–	3	5.1643e−02	–
4	1.2391e−02	3.8122	5	2.4103e−03	5.9993
6	3.9375e−04	8.5063	7	5.5405e−05	11.2129
8	6.8442e−06	14.0861	9	7.5328e−07	17.1020
10	7.4743e−08	20.2430	11	6.7508e−09	23.4951

using,

$$\text{Error}(h_K) = \|h_K(x,t,\xi) - h_{\text{ref}}(x,t,\xi)\|_{L^1(\mathcal{D};L^2_\rho(\mathbb{R}^d))}, \tag{5.4}$$

where h_K is the water height corresponding to a solution using K PC terms. We observe spectral convergence of EC scheme in Table 2b, faster than any fixed polynomial order, suggesting this solution has very high regularity, and confirming the expected numerical convergence rate in K.

5.2. Flat-bottom dam break

In the next experiment, we consider a stochastic water surface,

$$h(x,0,\xi) + B(x,0,\xi) = w(x,0,\xi) = \begin{cases} 2.0 + 0.1\xi & x < 0 \\ 1.5 + 0.1\xi & x > 0 \end{cases}, \quad q(x,0,\xi) = 0,$$

with a flat bottom $B(x,t,\xi) \equiv 0$. This is a stochastic modification of the deterministic "dam break test" problem from [17]. In Figure 2, we use a uniform grid size $n_x = 400$ over the physical domain $x \in [-1,1]$, and compute up to time $t = 0.4$. We test the example using the numerical methods EC, ES1, and ES2 developed in Section 4.

From Figure 2, similar to the results presented in Figures 1 and 4 of [17], we observe that the water surface with uncertainty develops a leftward-going rarefaction wave and a rightward-going shock. Similar to [17], EC computes such solutions accurately, but at the expense of large post-shock oscillations as observed on Figure 2 (right plot). These oscillations are expected since the EC scheme preserves energy on such solutions, and hence energy is not dissipated across the shock as it should. We also demonstrate in Figure 3 (middle and right plots) numerical energy conservation for the EC scheme. We note that the energy conservation errors due to time discretization are reduced significantly by decreasing the time step/CFL constant (right figure), similar to the results reported in Figure 1 in [17]. The presented results in Figures 2 and 3 (left figure) also illustrate that ES2 produces less smearing than ES1 at both the rarefaction and the shock waves. The schemes ES1 and ES2 are both designed to dissipate energy which is also confirmed by the numerical results in Figure 3 (middle plot), with the energy dissipation in ES2 being lower than with the ES1 scheme. In addition, the numerical results seem to indicate that the ES2 scheme is better able to capture large variance spikes compared to the ES1 scheme. Finally, the employment of the the numerical diffusion operators in ES1 and ES2 schemes removes oscillations present in the numerical solution using the EC scheme. The observed results are also in agreement with deterministic model numerical results reported in [17].

5.3. Stochastic bottom topography

Next, we consider the shallow water system with deterministic initial conditions,

$$w(x,0) = \begin{cases} 1 & x < 0 \\ 0.5 & x > 0 \end{cases}, \quad q(x,0) = 0,$$

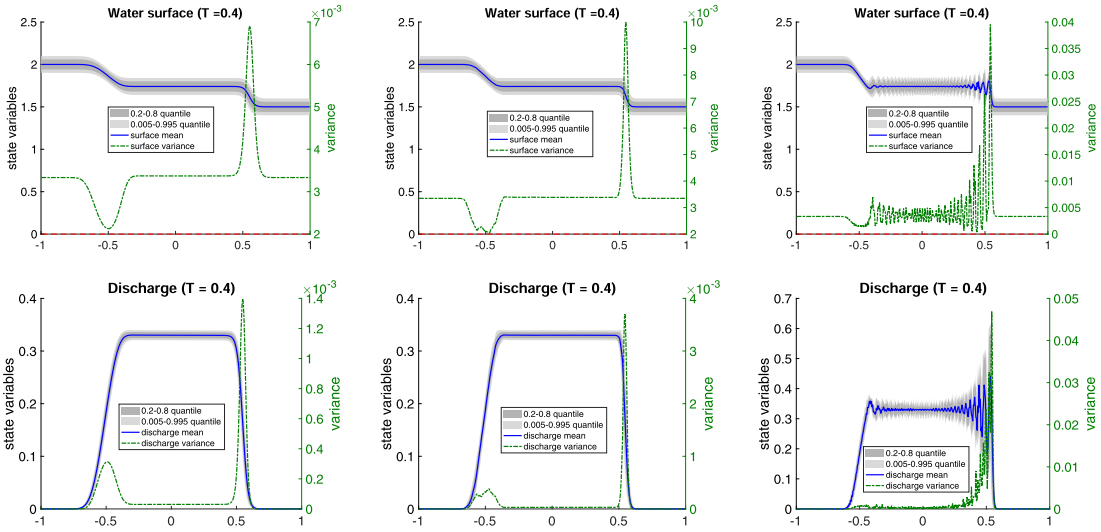

FIGURE 2. Results for Section 5.2. *Top*: water surface, *bottom*: discharge: *left*: ES1. *Middle*: ES2. *Right*: EC. Mesh $n_x = 400$ and PC basis functions $K = 9$.

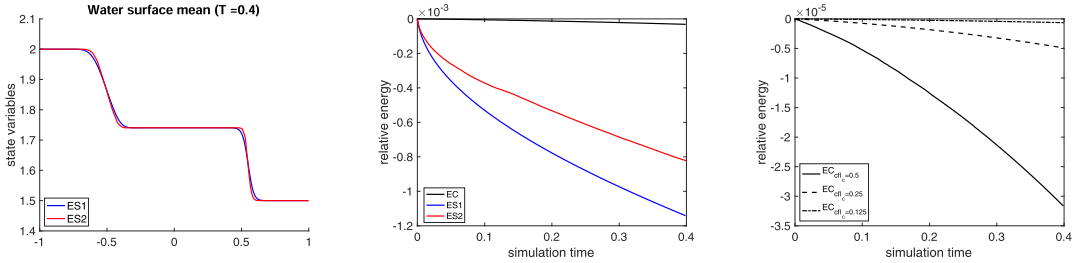

FIGURE 3. Results for Section 5.2. Comparison: *Left*: water surface mean ES1 *vs.* ES2. *Middle*: relative energy change in EC, ES1, ES2. *Right*: relative energy change in EC under different time step/CFL constant. Mesh $n_x = 400$ and PC basis functions $K = 9$.

and with a stochastic bottom topography,

$$B(x,\xi) = \begin{cases} 0.125(\cos(5\pi x) + 2) + 0.125\xi, & |x| < 0.2 \\ 0.125 + 0.125\xi, & \text{otherwise.} \end{cases} \tag{5.5}$$

This test example was presented previously in [11]. Initially, the highest possible bottom barely touches the initial water surface at $x = 0$, see Figures 4–6. We use a uniform grid size $n_x = 400, 800, 1600$ over the physical domain $x \in [-1, 1]$, and compute up to time $t = 0.0995$. (Immediately after this time, the EC scheme fails for $n_x = 400$ due to spurious oscillations near sharp gradients of the solution.) In Figures 5 and 6 we compare only performance of ES1 and ES2 at $t = 0.0995$ since EC fails on those meshes even earlier. Again, the numerical results indicate that the ES2 scheme can more easily resolve large, spatially-concentrated variance values compared to the ES1 scheme, but under mesh refinement both schemes converge to similar numerical solutions. In Figure 7, we show the numerical solution obtained using ES1 and ES2 at the final time $t = 0.8$ and on a mesh of size $n_x = 800$. For both schemes, the 99% confidence region of the water surface stays above the 99% confidence region of the bottom function in Figure 7, and both methods produce similar numerical solutions. The presented results are comparable to the results in Section 5.1 of [11]. In Figure 8, we again observe

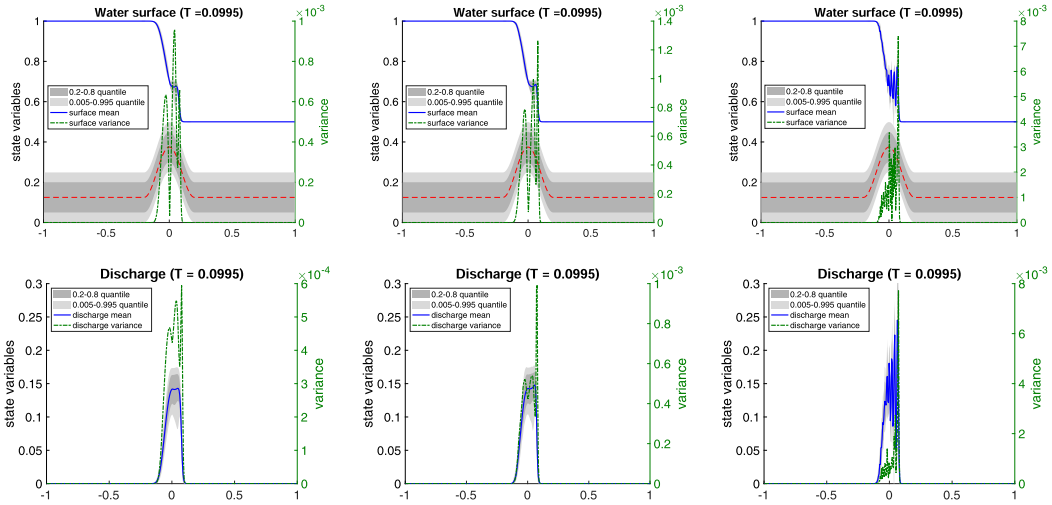

FIGURE 4. Comparison of the results for Section 5.3 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, *middle*: ES2, *right*: EC. Mesh $n_x = 400$ with $K = 9$ at earlier time $T = 0.0995$.

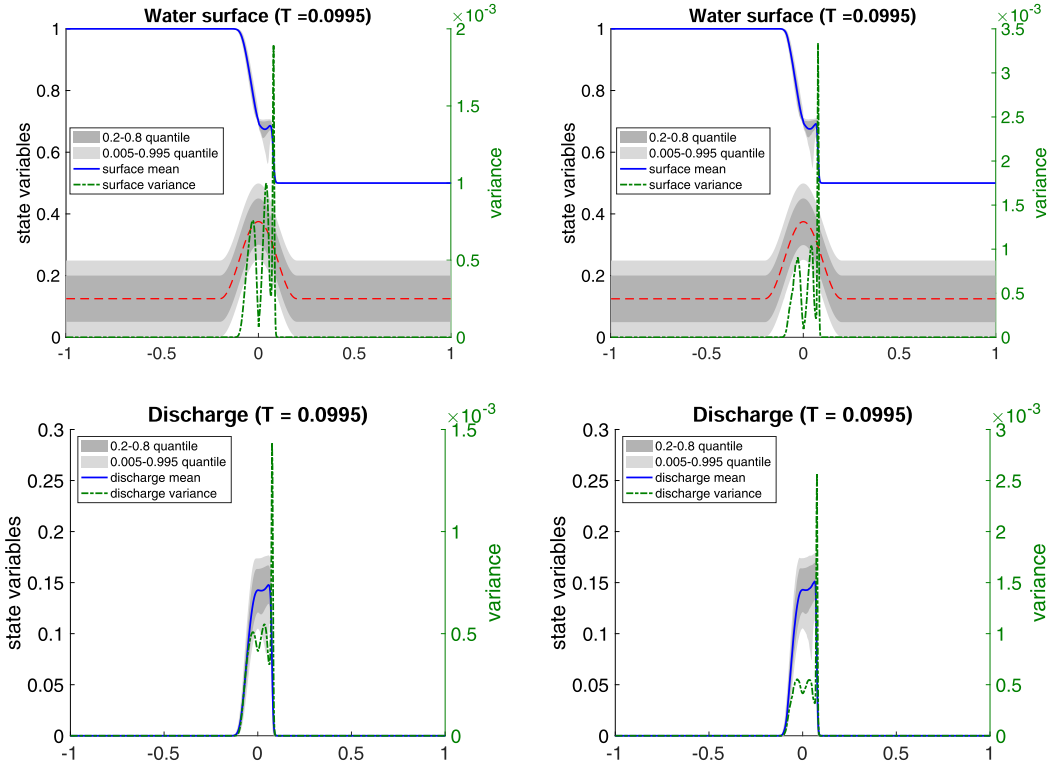

FIGURE 5. Comparison of the results for Section 5.3 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, *right*: ES2. Mesh $n_x = 800$ with $K = 9$ at earlier time $T = 0.0995$.

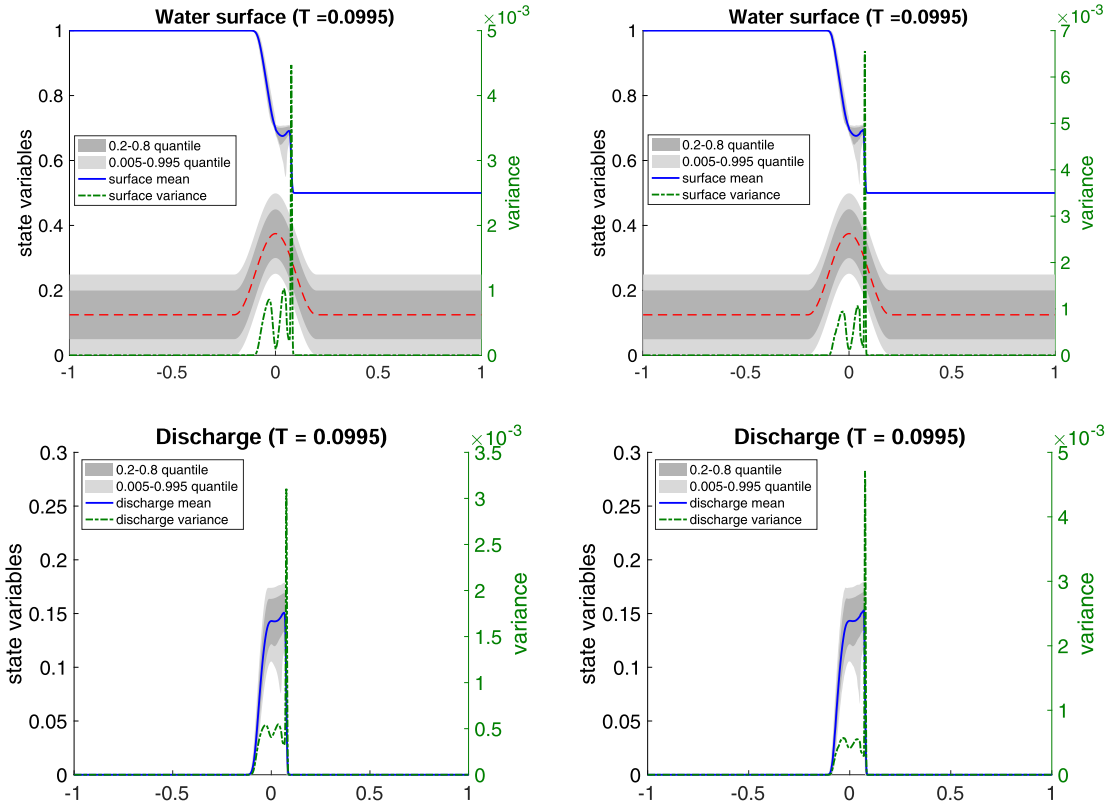

FIGURE 6. Comparison of the results for Section 5.3 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, *right*: ES2. Mesh $n_x = 1600$ with $K = 9$ at earlier time $T = 0.0995$.

as expected that the EC scheme numerically conserves energy, while ES1 and ES2 dissipate energy with larger dissipation produced by ES1 method.

5.4. Perturbation to lake at rest

For the next example, we consider the shallow water system with stochastic water surface,

$$w(x,0,\xi) = \begin{cases} 1 + 0.001(\xi + 1) & |x| \le 0.05 \\ 1 & \text{otherwise} \end{cases}, \quad q(x,0,\xi) = 0, \tag{5.6}$$

and with a deterministic bottom topography

$$B(x) = \begin{cases} 0.25(\cos(5\pi(x + 0.35)) + 1), & -0.55 < x < -0.15 \\ 0.125(\cos(10\pi(x - 0.35)) + 1), & 0.25 < x < 0.45 \\ 0, & \text{otherwise.} \end{cases} \tag{5.7}$$

The test is from [7] and is similar to the deterministic tests of the perturbation of lake at rest solution, for example to the one presented in [17].

We start by illustrating the accuracy of the ES1 and ES2 schemes on this problem with discontinuous initial data and nonsmooth solution, in which case we expect the system to dissipate energy. We compute the numerical

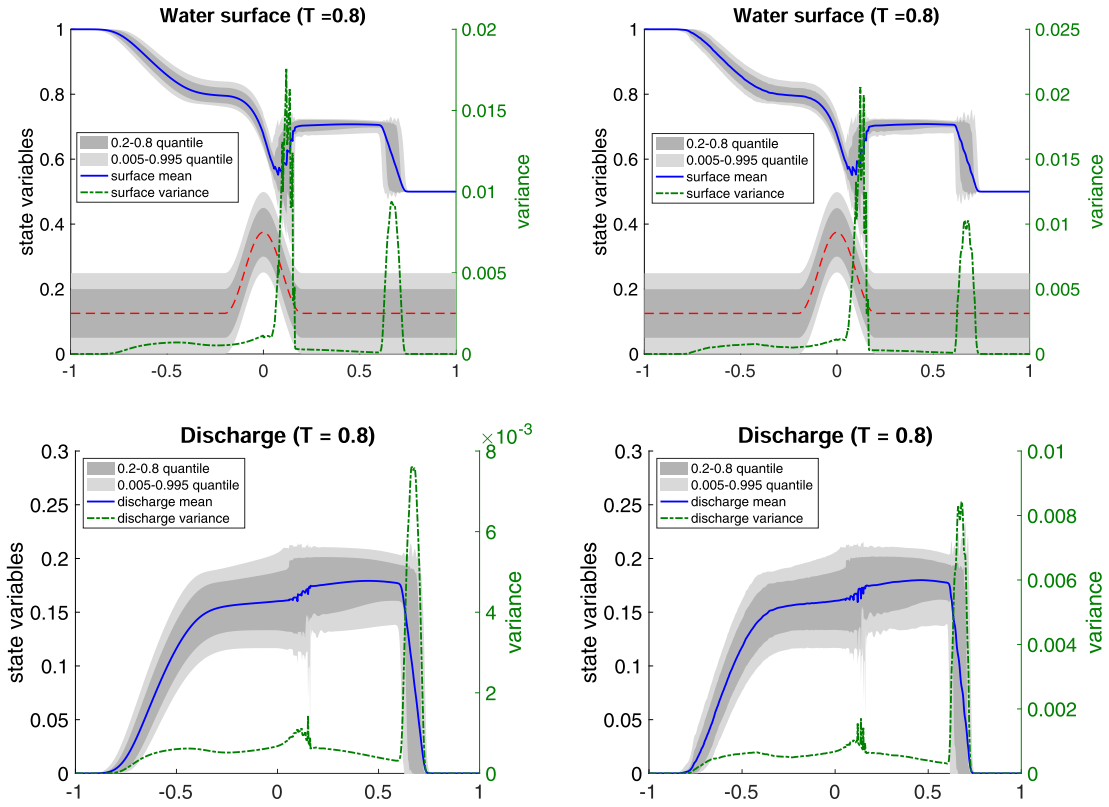

FIGURE 7. Comparison of the results for Section 5.3 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, and *right*: ES2. Mesh $n_x = 800$ with $K = 9$ at the final time $T = 0.8$.

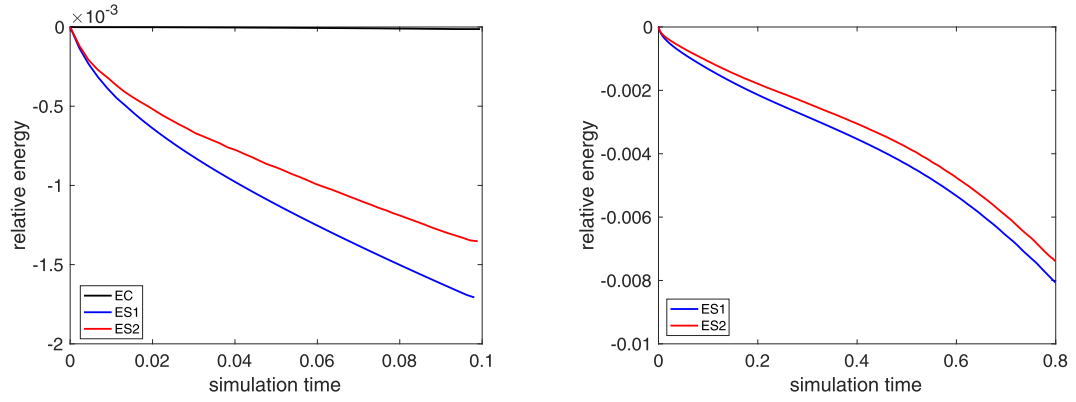

FIGURE 8. Comparison of the results for Section 5.3 using different schemes. Relative energy change: *Left*: EC *vs.* ES1 *vs.* ES2 on mesh $n_x = 400$ on time interval $[0, 0.0995]$. *Right*: ES1 *vs.* ES2 on mesh $n_x = 800$ on time interval $[0, 0.8]$.

TABLE 3. Results for Section 5.4: ES1 and ES2 order of the convergence in space for nonsmooth solution (5.6) and (5.7). Error is computed using (5.3a).

Mesh (n_x)	Error (ES1)	Order	Error (ES2)	Order
100	1.6891e−03	–	1.6872e−03	–
200	4.9033e−04	1.7844	4.6473e−04	1.8602
400	1.6874e−04	1.5390	1.3640e−04	1.7685
800	6.6192e−05	1.3501	4.5776e−05	1.5752

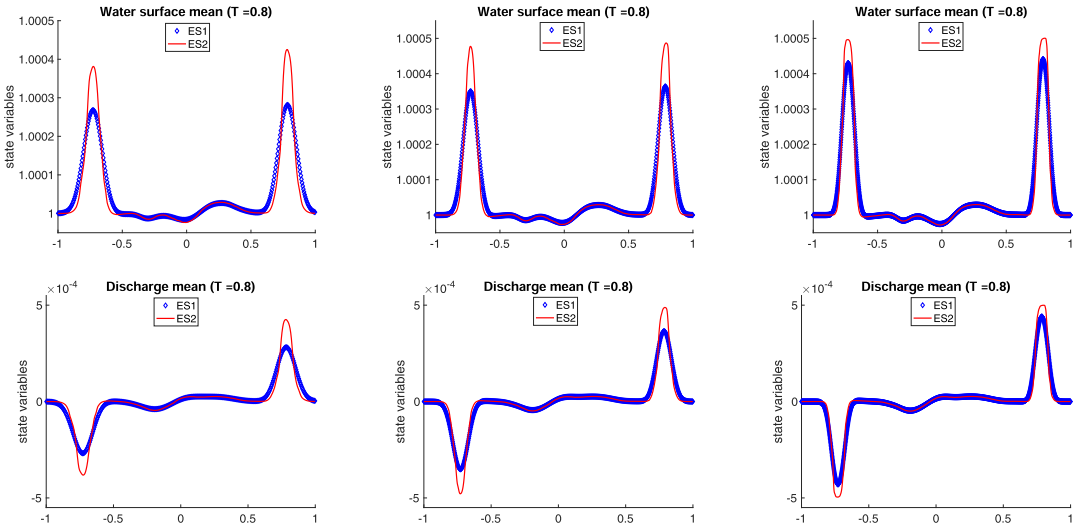

FIGURE 9. Results for Section 5.4. *Top*: water surface mean, *bottom*: discharge mean: *Left*: ES1 *vs.* ES2 on mesh $n_x = 400$. *Middle*: ES1 *vs.* ES2 on mesh $n_x = 800$. *Right*: ES1 *vs.* ES2 on mesh $n_x = 1600$. PC basis functions $K = 9$.

solution to a final time $t = 0.8$. We test the spatial accuracy of ES1 and ES2 schemes against reference solutions computed by ES1 and ES2 schemes, respectively, computed on a mesh of size $n_x = 3200$ with $K = 2$. As before, we illustrate convergence of the scheme using the water height h by computing the error between the reference solutions and the numerical solutions using (5.3a).

From Table 3 we observe that both ES1 and ES2 schemes perform well on problems with nonsmooth solutions (the order of convergence is limited by the regularity of the solution rather than the formal order of the scheme). ES1 produces somewhat higher than the guaranteed first order of convergence, while ES2 produces slightly lower than second order of the convergence, which is expected on such nonsmooth problems. The ES2 scheme delivers overall better resolution and accuracy than the ES1 scheme, as can be seen from the table and the plots of the solutions at different times in Figures 9 and 11.

From the results in Figures 9–11, we make conclusions similar to previous sections: both ES1 and ES2 capture small stochastic perturbations of the lake at rest solution quite well (with both leftward- and rightward- going waves present in the numerical solutions). The first order ES1 scheme exhibits much more dissipation in the left- and the right-going waves than the ES2 scheme, which produces a more accurate solution, as shown in Figures 9, 10 (left and middle figures), and 11. The results of the EC scheme is also shown in Figure 10 (right figure). The EC scheme resolves the left- and the right-going water waves with heights higher than in both ES1 and ES2 methods, but again there are oscillations present near both waves in the EC numerical solution since

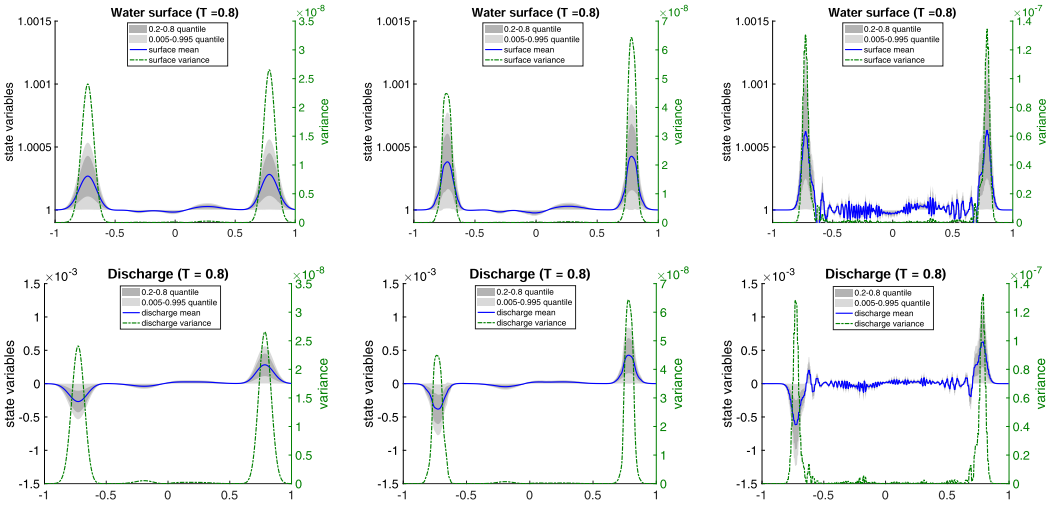

FIGURE 10. Comparison of the results for Section 5.4 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, *middle*: ES2, *right*: EC. Mesh $n_x = 400$ with $K = 9$ at time $T = 0.8$.

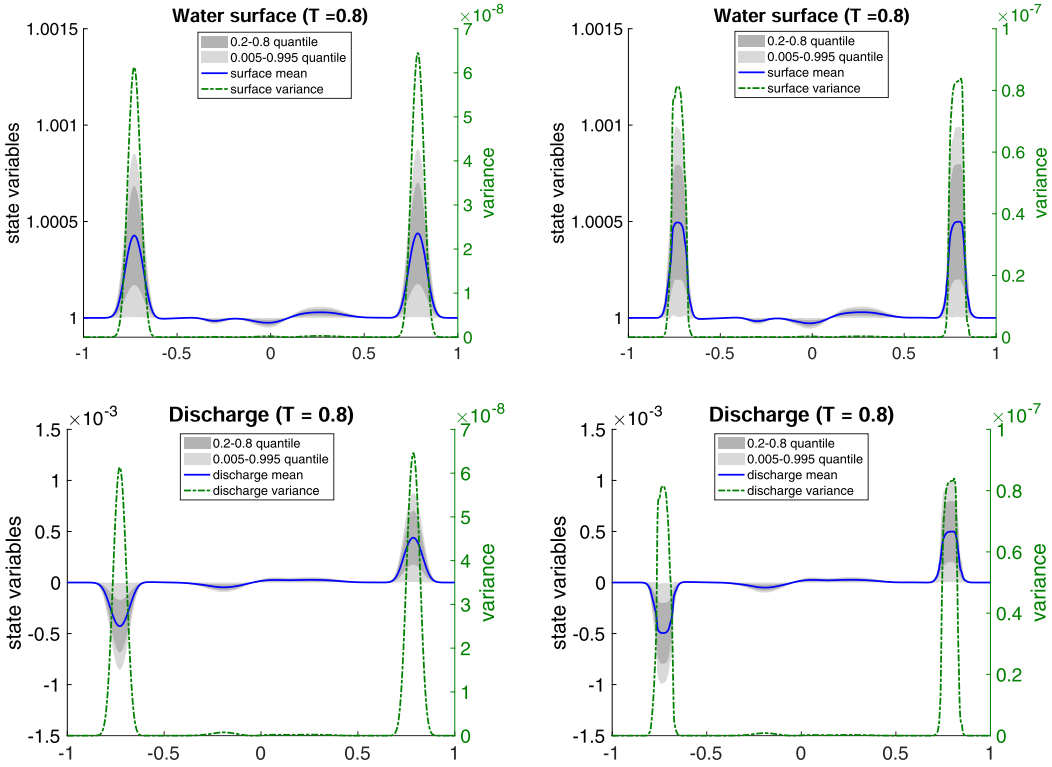

FIGURE 11. Comparison of the results for Section 5.4 using different schemes. *Top*: water surface. *Bottom*: discharge. *Left*: ES1, *right*: ES2. Mesh $n_x = 1600$ with $K = 9$ at time $T = 0.8$.

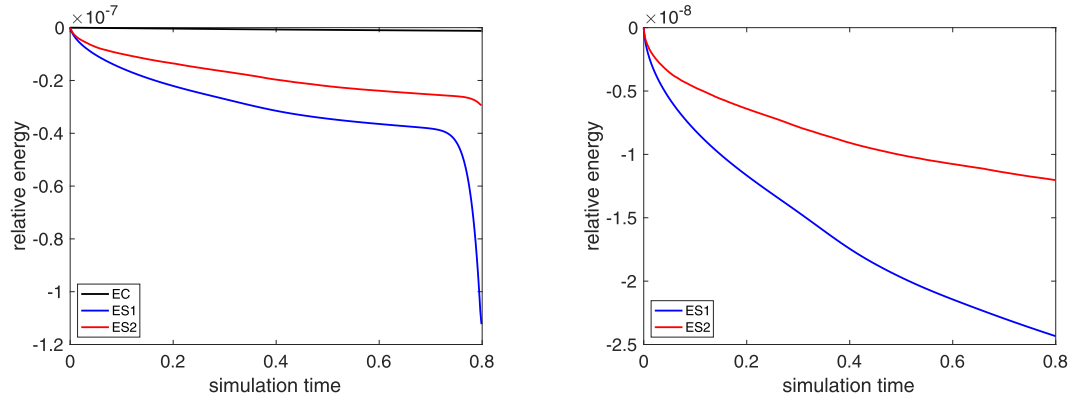

FIGURE 12. Comparison of the results for Section 5.4 using different schemes. Relative energy change: *Left*: EC *vs.* ES1 *vs.* ES2 on mesh $n_x = 400$. *Right*: ES1 *vs.* ES2 on mesh $n_x = 1600$.

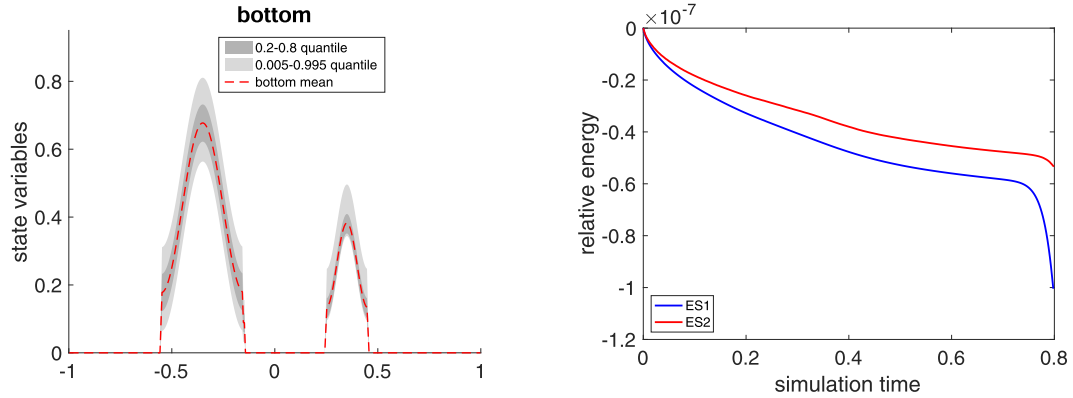

FIGURE 13. Results for Section 5.5. *Left*: bottom mean (5.9). *Right*: relative energy change: ES1 *vs.* ES2 on mesh $n_x = 400$.

EC does not dissipates energy across shocks. The relative energy change for this example produced by EC, ES1 and ES2 methods is illustrated in Figure 12. The presented results are also comparable to the results reported in [17] and in [7].

5.5. Perturbation to lake at rest with two-dimensional random variable

As a final example, we demonstrate perfomance of ES1 and ES2 methods on a model with two random parameters that inject uncertainty in to the model. We consider the shallow water system with stochastic water surface,

$$w(x,0,\xi) = \begin{cases} 1 + 0.001(\xi_1 + 1) & |x| \le 0.05 \\ 1 & \text{otherwise} \end{cases}, \quad q(x,0,\xi) = 0, \tag{5.8}$$

and with a stochastic bottom topography

$$B(x,\xi) = \begin{cases} 0.25(\cos(5\pi(x+0.35)) + 1) + 0.12e^{\xi_2}, & -0.55 < x < -0.15 \\ 0.125(\cos(10\pi(x-0.35)) + 1) + 0.1(1+\xi_1), & 0.25 < x < 0.45 \\ 0, & \text{otherwise.} \end{cases} \tag{5.9}$$

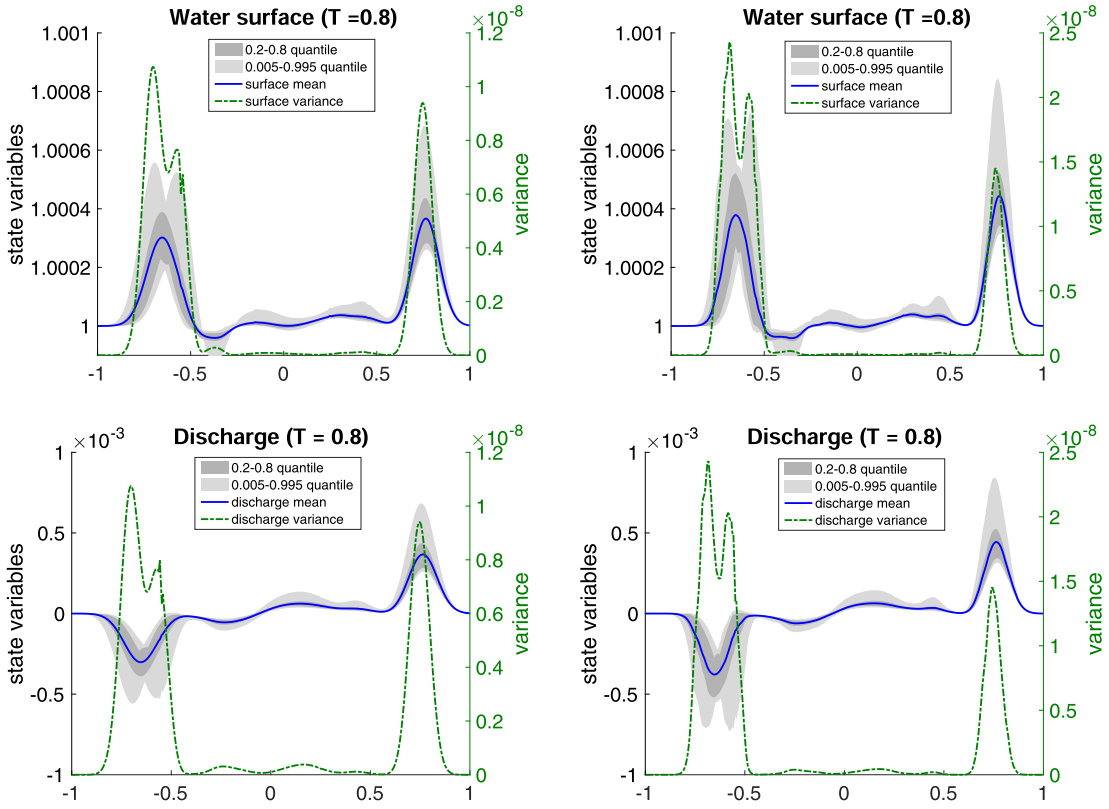

FIGURE 14. Results for Section 5.5. *Top*: water surface, *bottom*: discharge: *Left*: ES1. *Right*: ES2. Mesh $n_x = 400$. PC basis functions $(K_1 = 3, K_2 = 5)$ for (ξ_1, ξ_2).

In this system we consider the two-dimensional random variable $\xi = (\xi_1, \xi_2)$, where ξ_1 and ξ_2 are modeled as independent and identically distributed random variables over $[-1, 1]$ having a Beta distribution with parameters $(\alpha, \beta) = (1, 3)$. The plot of the bottom mean and quantiles as defined by B in (5.9) is presented in Figure 13 (left). To simulate the SG SWE system, we use $K_1 = 3$ and $K_2 = 5$ polynomial terms for parameters ξ_1 and ξ_2, respectively, corresponding to a tensor-product space of polynomials having dimension $K = \dim P = K_1 K_2 = 15$. We simulate up to time $T = 0.8$ with $n_x = 400$ elements for $x \in [-1, 1]$.

The numerical results are plotted in Figure 14, showing the water surface and the discharge develop more complex structures than in the previous example (5.6) and (5.7) in Section 5.4, and the relative energy change shown in Figure 13 (right) is of a comparable order of magnitude. In addition, we draw conclusions similar to previous sections: both ES1 and ES2 capture small stochastic perturbations of the lake at rest solution quite well (with both leftward- and rightward- going waves present in the numerical solutions). The first order ES1 scheme exhibits more dissipation in the left- and the right-going waves than the ES2 scheme, which produces a more accurate solution, as shown in Figure 14.

6. Conclusion

In this work we derived an entropy-entropy flux pair for the spatially one-dimensional hyperbolicity-preserving, positivity-preserving SG SWE system developed in [11]. Such entropy-entropy flux pairs are the theoretical starting point for proposing entropy admissibility criteria to resolve non-uniqueness of weak solutions. Using the proposed entropy-entropy flux pair, we designed second-order energy conservative, and first- and

second-order energy stable finite volume schemes for the SG SWE. The proposed schemes are also well-balanced. We provided several numerical experiments to illustrate performance of the methods. As part of future research, we plan to extend such methods to models in two spatial dimensions, to explore alternative constructions of diffusion operators, and to investigate other reconstruction approaches for the entropy variables.

Acknowledgements

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

References

[1] A.-J.-C. Barré de Saint-Venant, Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* (1871).

[2] S. Benzoni-Gavage and D. Serre, Multi-dimensional Hyperbolic Partial Differential Equations: First-order Systems and Applications. Oxford University Press on Demand (2007).

[3] A. Bermudez and M.E. Vazquez, Upwind methods for hyperbolic conservation laws with source terms. *Comput. Fluids* **23** (1994) 1049–1071.

[4] S. Bryson, Y. Epshteyn, A. Kurganov and G. Petrova, Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM: Math. Modell. Numer. Anal.* **45** (2011) 423–446.

[5] C. Chen, C. Dawson and E. Valseth, Cross-mode stabilized stochastic shallow water systems using stochastic finite element methods. *Comput. Methods Appl. Mech. Eng.* **405** (2023) 115873.

[6] A. Chertock, S. Cui, A. Kurganov and T. Wu, Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *Int. J. Numer. Methods Fluids* **78** (2015) 355–383.

[7] A. Chertock, S. Jin and A. Kurganov, A well-balanced operator splitting based stochastic Galerkin method for the one-dimensional Saint-Venant system with uncertainty. Preprint https://chertock.wordpress.ncsu.edu/files/2019/10/CJK2.pdf (2015).

[8] A. Cohen and R. DeVore, Approximation of high-dimensional parametric PDEs. *Acta Numer.* **24** (2015) 1–159.

[9] N. Črnjarić-Žic, S. Vuković and L. Sopta, Balanced finite volume WENO and central WENO schemes for the shallow water and the open-channel flow equations. *J. Comput. Phys.* **200** (2004) 512–548.

[10] C.M. Dafermos, Hyperbolic Conservation Laws in Continuum Physics. Springer (2016).

[11] D. Dai, Y. Epshteyn and A. Narayan, Hyperbolicity-preserving and well-balanced stochastic Galerkin method for shallow water equations. *SIAM J. Sci. Comput.* **43** (2021) A929–A952.

[12] D. Dai, Y. Epshteyn and A. Narayan, Hyperbolicity-preserving and well-balanced stochastic Galerkin method for two-dimensional shallow water equations. *J. Comput. Phys.* **452** (2022) 110901.

[13] B.J. Debusschere, H.N. Najm, P.P. Pébay, O.M. Knio, R.G. Ghanem and O.P. Le Maître, Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM J. Sci. Comput.* **26** (2004) 698–719.

[14] Y. Epshteyn and T. Nguyen, Adaptive central-upwind scheme on triangular grids for the Saint-Venant system. *Commun. Math. Sci.* **21** (2023) 671–708.

[15] O.G. Ernst, A. Mugler, H.-J. Starkloff and E. Ullmann, On the convergence of generalized polynomial chaos expansions. *ESAIM: Math. Modell. Numer. Anal.* **46** (2012) 317–339.

[16] U. Fjordholm, S. Mishra and E. Tadmor, Energy Preserving and Energy Stable Schemes for the Shallow Water Equations. *London Mathematical Society Lecture Note Series*. Cambridge University Press (2009) 93–139.

[17] U.S. Fjordholm, S. Mishra and E. Tadmor, Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *J. Comput. Phys.* **230** (2011) 5587–5609.

[18] U.S. Fjordholm, S. Mishra and E. Tadmor, Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. *SIAM J. Numer. Anal.* **50** (2012) 544–573.

[19] S. Gerster and M. Herty, Entropies and symmetrization of hyperbolic stochastic Galerkin formulations. *Commun. Comput. Phys.* **27** (2020) 639–671.

[20] S. Gerster, A. Sikstel and G. Visconti, Haar-type stochastic Galerkin formulations for hyperbolic systems with Lipschitz continuous flux function. Preprint arXiv:2203.11718 (2022).

[21] S. Gottlieb, C.-W. Shu and E. Tadmor, Strong stability-preserving high-order time discretization methods. *SIAM Rev.* **43** (2001) 89–112.

[22] S. Jin, D. Xiu and X. Zhu, A well-balanced stochastic galerkin method for scalar hyperbolic balance laws with random inputs. *J. Sci. Comput.* **67** (2016) 1198–1218.

[23] A. Kurganov, Finite-volume schemes for shallow-water equations. *Acta Numer.* **27** (2018) 289–351.

[24] A. Kurganov and D. Levy, Central-upwind schemes for the Saint-Venant system. *ESAIM: Math. Modell. Numer. Anal.* **36** (2002) 397–425.

[25] A. Kurganov and G. Petrova, A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Commun. Math. Sci.* **5** (2007) 133–160.

[26] J. Kusch, R.G. McClarren and M. Frank, Filtered stochastic Galerkin methods for hyperbolic equations. *J. Comput. Phys.* **403** (2020) 109073.
[27] R.J. LeVeque, Numerical Methods for Conservation Laws. Springer (1992).
[28] R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems. Cambridge University Press (2002).
[29] X. Liu, J. Albright, Y. Epshteyn and A. Kurganov, Well-balanced positivity preserving central-upwind scheme with a novel wet/dry reconstruction on triangular grids for the Saint-Venant system. *J. Comput. Phys.* **374** (2018) 213–236.
[30] G. Poëtte, B. Després, and D. Lucor, Uncertainty quantification for systems of conservation laws. *J. Comput. Phys.* **228** (2009) 2443–2467.
[31] R. Pulch and D. Xiu, Generalised polynomial chaos for a class of linear conservation laws. *J. Sci. Comput.* **51** (2012) 293–312.
[32] B.D. Rogers, A.G.L. Borthwick and P.H. Taylor, Mathematical balancing of flux gradient and source terms prior to using Roe's approximate Riemann solver. *J. Comput. Phys.* **192** (2003) 422–451.
[33] L. Schlachter and F. Schneider, A hyperbolicity-preserving stochastic Galerkin approximation for uncertain hyperbolic systems of equations. *J. Comput. Phys.* **375** (2018) 80–98.
[34] R.C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications. SIAM-Society for Industrial and Applied Mathematics, Philadelphia (2013).
[35] T.J. Sullivan. Introduction to Uncertainty Quantification. Vol. 63. Springer (2015).
[36] E. Tadmor, The numerical viscosity of entropy stable schemes for systems of conservation laws. I. *Math. Comput.* **49** (1987) 91–103.
[37] E. Tadmor, Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numer.* **12** (2003) 451–512.
[38] K. Wu, H. Tang and D. Xiu, A stochastic Galerkin method for first-order quasilinear hyperbolic systems with uncertainty. *J. Comput. Phys.* **345** (2017) 224–244.
[39] Y. Xing, High order finite volume WENO schemes for the shallow water flows through channels with irregular geometry. *J. Comput. Appl. Math.* **299** (2016) 229–244.
[40] Y. Xing, Chapter 13 – numerical methods for the nonlinear shallow water equations, in Handbook of Numerical Analysis. Vol. 18 of *Handbook of Numerical Methods for Hyperbolic Problems*, edited by R. Abgrall and C.-W. Shu. (2017) 361–384.
[41] Y. Xing and C.-W. Shu, High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *J. Comput. Phys.* **208** (2005) 206–227.
[42] Y. Xing and C.-W. Shu, High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *J. Comput. Phys.* **214** (2006) 567–598.
[43] Y. Xing and C.-W. 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. *Commun. Comput. Phys.* **1** (2006) 100–134.
[44] D. Xiu, Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press (2010).
[45] X. Zhong and C.-W. Shu, Entropy stable Galerkin methods with suitable quadrature rules for hyperbolic systems with random inputs. *J. Sci. Comput.* **92** (2022) 14.
[46] T. Zhou and T. Tang, Galerkin methods for stochastic hyperbolic problems using bi-orthogonal polynomials. *J. Sci. Comput.* **51** (2012) 274–292.
[47] J.G. Zhou, D.M. Causon, C.G. Mingham and D.M. Ingram, The surface gradient method for the treatment of source terms in the shallow-water equations. *J. Comput. Phys.* **168** (2001) 1–25.