Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Well-balanced, energy stable schemes for the shallow water equations with varying topography

U.S. Fjordholm, S. Mishra and E. Tadmor[*]

Research Report No. 2010-26
September 2010

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

[*] Dept. of Mathematics, University of Maryland, MD 20742-4015, USA.
The author's research was supported in part by ONR grant #N00014-09-10385 and NSF grant DMS10-08397.

WELL-BALANCED, ENERGY STABLE SCHEMES FOR THE SHALLOW WATER EQUATIONS WITH VARYING TOPOGRAPHY

U. S. FJORDHOLM, S. MISHRA, AND E. TADMOR

Abstract. We consider the shallow water equations with bottom topography. The smooth solutions of these equations are energy conservative, whereas weak solutions are energy stable. The equations possess interesting steady states like the lake at rest in both one and two space dimensions, as well as moving equilibrium states in one dimension. We design an energy conservative finite volume scheme that preserves the lake at rest as well as moving equilibrium states. Suitable energy stable numerical diffusion operators based on energy and equilibrium variables are designed to preserve the lake at rest and moving equilibrium states, respectively. Several numerical experiments illustrating the robustness of the energy preserving and energy stable well-balanced schemes are presented.

Contents

Date: July 12, 2010.

1991 *Mathematics Subject Classification.* 65M06,35L65.

Key words and phrases. Shallow water equations, Energy preserving schemes, Energy stable schemes, Eddy viscosity, Numerical diffusion.

Acknowledgment. The work on this paper was started when S. M. visited the Center of Scientific Computation and Mathematical Modeling (CSCAMM) and he thanks CSCAMM and all its members for the excellent hospitality and facilities. E. T. Research was supported in part by ONR grant #N00014-09-10385 and NSF grant DMS10-08397.

1. Introduction

Flows in lakes, rivers, irrigation channels and near-shore oceanic flows are of great interest in hydrology, oceanography and climate modeling. Common to all of these flows is the fact that vertical scales of motion are much smaller than the horizontal scales. By this and the assumption of hydrostatic balance (see [40]), the incompressible Navier-Stokes equations of fluid dynamics can be simplified and reduce to the so-called shallow water equations

$$\begin{aligned} h_t + (hu)_x + (hv)_y &= 0, \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x + (huv)_y &= -ghb_x, \\ (hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2}gh^2\right)_y &= -ghb_y. \end{aligned} \tag{1.1}$$

Here, h is the height of the fluid column and (u, v) is the velocity field. The constant g is the acceleration due to gravity and the function $b \equiv b(x, y)$ represents the bottom topography of the surface over which the fluid flows. In general, the bottom topography can be rather complicated and possibly discontinuous. We have neglected eddy viscosity in the above equation. When the variation of the unknowns in the y-direction are negligible, one may find the one-dimensional version of (1.1) by setting v and all the derivatives in the y-direction to zero, thus obtaining the system

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x &= -ghb_x. \end{aligned} \tag{1.2}$$

The shallow water system with topography (1.1) amounts to a system of *balance laws*,

$$U_t + f(U)_x + g(U)_y = -s(x, y, U), \tag{1.3}$$

where $U = [h, hu, hv]^\top$ is the vector of unknowns, $f = [hu,\ hu^2 + \frac{1}{2}gh^2,\ huv]^\top$ and $g = [hv,\ huv,\ hv^2 + \frac{1}{2}gh^2]^\top$ are the flux vectors, and $s = [0,\ ghb_x,\ ghb_y]^\top$ is the source vector.

If the bottom topography is flat, i.e. $b \equiv Const.$, then (1.1) is reduced to the standard shallow water equations without topography, which is a strictly hyperbolic system of conservation laws,

$$U_t + f(U)_x + g(U)_y = 0. \tag{1.4}$$

It is well-known that solutions of the conservation law (1.4), and likewise, solutions of the balance law (1.3), can develop shock discontinuities in a finite time, independent of whether the initial data is smooth or not. Hence, the solutions of balance laws (1.3) are considered in the weak sense and are well-defined as long as the source s remains uniformly bounded, [6]. In particular, weak solutions of (1.1) are well-defined under the assumption that the topography function b is in $W^{1,\infty}(\mathbb{R}^2)$. However, difficulties arise when the topography function is discontinuous: the action of the source term on the right of (1.1) can be interpreted as a non-conservative product (see [7]), or by a limiting smoothing process of b.

1.1. The entropy condition.

Weak solutions of conservation laws (1.4), and likewise, weak solutions of the balance law (1.3), need not be unique. Another aspect of non-uniqueness enters (1.1) through the action of the source term $s(x, y, U) = -gh\nabla b(x, y)$: its interpretation as a non-conservative product or using a limiting smoothing process depends on a non-unique choice of a path integral. To address this issue of non-uniqueness, an additional admissibility criterion is imposed, based on the so-called *entropy condition*. To this end, one assumes that the general system of balance laws (1.3) is equipped with a convex entropy function $E = E(U)$, associated entropy flux functions $H = H(U)$, $K = K(U)$ and $J = \big[J_1(x, y, U), J_2(x, y, U)\big]^\top$, such that the following compatibility relations, expressed in terms of the vector of *entropy variables* $V := \partial_U E$, hold:

$$\partial_U H = \langle V,\ \partial_U f(U)\rangle, \qquad \partial_U K = \langle V,\ \partial_U g(U)\rangle, \qquad \partial_x J_1 + \partial_y J_2 = \langle V,\ s\rangle. \tag{1.5a}$$

Multiplying (1.3) by $V = \partial_U E$, the compatibility relations (1.5a) imply that smooth solutions of (1.4) satisfy the conservation law

$$E(U)_t + \big(H(U) + J_1\big)_x + \big(K(U) + J_2\big)_y = 0. \tag{1.5b}$$

Conversely, if this additional conservation law holds for *all smooth* functions U, then E is an entropy function, i.e., (1.5a) holds with the entropy fluxes H, K and J. This balance between the entropy and entropy fluxes has to be modified to take into account the presence of possible discontinuities in (1.3): we postulate that the discontinuous solution U of the balance laws (1.3) can be realized by a vanishing viscosity limit, which in turn leads to the distributional entropy inequality

$$E(U)_t + \big(H(U) + J_1\big)_x + \big(K(U) + J_2\big)_y \leq 0. \tag{1.5c}$$

In the absence of a source term ($s \equiv 0$), (1.5c) amounts to the usual entropy condition for conservation laws [6]. Scalar conservation laws are equipped with an infinitely many entropy pairs — indeed, every convex function serves as a scalar entropy function, and this paves the way for a proof of existence, uniqueness and stability in the scalar framework. For general systems of conservation laws, however, the existence of entropy pairs places a compatibility restriction on the structure of the fluxes $f(\cdot)$ and $g(\cdot)$ which are not always met. Similarly, general systems of balance laws need not posses entropy functions, except for special systems which are endowed with at least one entropy function. Observe that in the particular case of balance laws, the source term, s also has to have a special structure for the entropy compatibility (1.5a) to hold.

An illustrative example is provided by the shallow water system with bottom topography (1.1). Here, the total energy

$$E(U) := \frac{1}{2}\left(hu^2 + hv^2 + gh^2 + ghb\right)$$

serves as an entropy function. The total energy $E(U)$ consists of the kinetic energy $h\big(u^2 + v^2\big)/2$ and the gravitational potential energy $gh(h+b)$, which involves the bottom topography b. A straightforward calculation reveals that if U is a smooth solution of (1.1) then

$$E(U)_t + \left(\frac{1}{2}\left(hu^3 + huv^2\right) + ghu(h+b)\right)_x + \left(\frac{1}{2}\left(hu^2v + hv^3\right) + ghv(h+b)\right)_y = 0. \tag{1.6}$$

Thus, $E(U)$ is an entropy function associated with entropy fluxes

$$H(U) := \frac{1}{2}\left(hu^3 + huv^2\right) + gh^2u, \qquad K(U) := \frac{1}{2}\left(hu^2v + hv^3\right) + gh^2v, \qquad J := ghb[u,v]^\top.$$

Integration of (1.6) yields that for smooth solutions of the balance law (1.1), energy is conserved, $\frac{d}{dt}\int_{\mathbb{R}^2} E \equiv 0$. However, energy should be dissipated across shock discontinuities, as dictated by the entropy dissipation postulate (1.5c)

$$E(U)_t + \left(\frac{1}{2}\left(hu^3 + huv^2\right) + ghu(h+b)\right)_x + \left(\frac{1}{2}\left(hu^2v + hv^3\right) + ghv(h+b)\right)_y \leq 0. \tag{1.7}$$

Note that the bottom topography plays a crucial role in the the entropy condition (1.7), whose weak formulation is *independent* of any specific realization (using a specific path-integral or a smoothing process) of the non-conservative product $gh\nabla b$.

1.2. **Numerical approximations.** In the absence of explicit solution formulas, numerical schemes are a key tool in the study of systems of balance laws like (1.3). Among the popular methods for discretizing conservation (balance) laws are the so-called finite volume (FV) methods [22]. For simplicity, we consider a uniform Cartesian mesh $\{(x_i, y_j)\}$ in $\mathbb{R}^2$ with a fixed mesh size $\Delta x := x_{i+1/2} - x_{i-1/2}$ and $\Delta y := y_{j+1/2} - y_{j-1/2}$, respectively. The domain is partitioned into rectangular cells $I_{i,j} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$. A standard cell-centered FV method consists of updating the cell averages

$$U_{i,j}(t) = \frac{1}{\Delta x \Delta y}\int_{I_{i,j}} U(x,y,t)dxdy$$

at each time level. For simplicity, we drop the time dependence of every quantity and write a standard finite volume scheme for (1.3) in the semi-discrete form as

$$\frac{d}{dt}U_{i,j} = -\frac{1}{\Delta x}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) - \frac{1}{\Delta y}\left(G_{i,j+1/2} - G_{i,j-1/2}\right) - S_{i,j}. \tag{1.8}$$

There are three main ingredients in the formulation of the FV schemes (1.8).

(i) $F_{i\pm1/2,j}$ and $G_{i,j\pm1/2}$ are numerical fluxes at the cell-edges consistent with the differential fluxes f and g, respectively. These numerical fluxes can be evaluated in terms of the Godunov, Roe or HLL fluxes [22]. Higher-order accuracy can be achieved by reconstruction of non-oscillatory numerical fluxes which can be chosen out of a large library of TVD or (W)ENO fluxes coupled with stencils of either upwind of central schemes [13, 14, 30, 31, 32, 24, 19, 35].

(ii) Discretization of the source terms is often performed with either a cell-centered evaluation of the source term or a fractional steps method [22]. For example, one may use

$$S_{i,j} = \left[0,\ gh_{i,j}\frac{b_{i+1,j} - b_{i-1,j}}{2\Delta x},\ gh_{i,j}\frac{b_{i,j+1} - b_{i,j-1}}{2\Delta y}\right]^{\top}. \tag{1.9}$$

Note that this discretization is consistent with the source in (1.1) — in fact, it is second-order accurate for smooth solutions.

(iii) Finally, time-integration employs strong stability preserving (SSP) Runge-Kutta methods. In this paper we use the second-order SSP Runge-Kutta method of [11]: given a solution $U_{i,j}^n$ at time step t_n, the solution $U_{i,j}^{n+1}$ is computed by

$$\begin{aligned} U_{i,j}^{*} &= U_{i,j}^{n} + \Delta t^n \mathcal{L}(U_{i,j}^{n}) \\ U_{i,j}^{**} &= U_{i,j}^{*} + \Delta t^n \mathcal{L}(U_{i,j}^{*}) \\ U_{i,j}^{n+1} &= \frac{1}{2}(U_{i,j}^{n} + U_{i,j}^{**}), \end{aligned} \tag{1.10}$$

where $\mathcal{L}$ is the right-hand side of (1.8). The time step Δt^n is determined by a standard CFL condition. In all simulations we use a CFL number of 0.45, unless otherwise is specified.

1.3. **Entropy stable schemes.**

Many of the above mentioned FV approximations of (1.1) perform well in practice, but the question of their stability remains open. In particular, these schemes do not necessarily respect the energy dissipation statement in (1.7), or they may be "overloaded" with an excessive amount of numerical dissipation near shocks, which in turn leads to large numerical errors, particularly for long time integration; see [1, 2, 3] for an extensive discussion of this issue. Hence, it is highly desirable to design a high-order entropy stable FV scheme which respects a "faithful" description of the energy balance of the shallow water system (1.7). In particular, they add a minimal amount of numerical dissipation which guarantees energy conservation in the smooth regime.

The question of entropy stability for general systems of conservation laws of the form (1.4) was addressed in the pioneering papers [34, 36]. In [34], entropy stability was pursued by a *comparison* principle: a FV scheme was shown to be entropy stable if it contains more numerical diffusion than certain *entropy conservative schemes*, where "more" is interpreted in the sense of ordering that exists between symmetric matrices. Explicit expressions for entropy conservative schemes in terms of a novel pathwise decomposition was presented in [36]. Higher order entropy conservative schemes for systems of conservation laws were developed in [20, 21]. These entropy conservative schemes were used in [37, 38] for computing solutions of Euler and, respectively, the shallow water system with flat bottom topography. In a recent paper [8], we designed new explicit energy preserving FV schemes for the shallow water equations with flat bottom topography. These schemes were shown to be more computationally efficient than those proposed in [38], and novel, computationally efficient numerical diffusion operators were proposed to gain overall energy stability.

The first aim in this paper is to address the question of entropy stability for FV approximations of general balance laws (1.3). Specifically, we consider the shallow water system (1.1) where the presence of a bottom topography enters into a more involved entropy balance (1.7). In Section 2 we present a one-dimensional energy conservative scheme, satisfying the discrete analogue of the energy conservation statement (1.6), with which we are able to design a general class of energy stable approximations for (1.1). We discuss first- and second-order energy stable schemes in Sections 2.3 and 2.4 respectively. The energy conservative scheme presented here is an extension of the explicit energy conservative scheme for the shallow water system with a flat bottom topography, proposed in the recent paper [8]. The two-dimensional extension of energy stable schemes is presented in Section 4.

1.4. **Steady states and well-balanced schemes.** Another important issue which arises in connection with balance laws such as the shallow water system (1.1) is the simulation of their steady states. A *steady state* for (1.3) is a solution that is constant in time. We mention two prototypical examples.

(i) The most important example of a steady state for (1.1) is the so-called *lake at rest*, given by

$$u \equiv 0, \qquad v \equiv 0, \qquad h + b \equiv \text{constant}. \tag{1.11}$$

Many interesting applications involve computing perturbations of the lake at rest. Waves on a lake or tsunami waves in deep ocean (the amplitude of a typical tsunami wave is of the order of centimeters whereas the height of water in deep ocean is of the order of kilometers) are typical situations where the main interest is in computing perturbations of the "lake at rest" solutions.

(ii) In the one-dimensional equation (1.2), all steady states satisfy the algebraic relations

$$m \equiv \text{constant}, \qquad p \equiv \text{constant}, \tag{1.12a}$$

where the *equilibrium variables* m and p are defined as

$$m := hu, \qquad p := \frac{u^2}{2} + g(h+b). \tag{1.12b}$$

We note that the one-dimensional lake at rest (1.11) is a special case of (1.12a) corresponding to $u \equiv 0$. The conditions (1.12a) are nonlinear and possess a rich family of solutions. These *moving equilibrium states* are much more difficult to compute than the lake at rest. Recent results on well-balanced schemes with respect to these general moving steady states can be found at [26, 29], but this issue is still a work in progress.

Standard numerical schemes like (1.8) with naive discretizations of the source term like (1.9) do not preserve the lake at rest [22]. This implies that the scheme does not keep a discrete form of (1.11) stationary in time. The error can be at least of the order of truncation error for each time step and can lead to large deviations from the steady state for long time scales. Furthermore, computing small perturbations of (1.11) is not possible due to the lack of balancing. A numerical scheme which preserves a discrete version of a steady state like (1.11) is termed *well-balanced* with respect to the steady state. Well-balanced schemes are essential for computing perturbations of steady states.

Well-balanced schemes for the shallow water equations are still undergoing extensive development. The pioneering paper of LeVeque [23] was one of the first to propose a well-balanced scheme for the lake at rest. Many other well-balanced schemes for this state have been proposed in [4, 15, 16, 9, 5, 18, 25] and other references therein. The basic idea behind most of these papers is to modify the numerical fluxes by a hydrostatic reconstruction and introduce a source discretization to balance the flux difference. The design of well-balanced schemes for general steady states (1.12a) can be quite complicated. Their implementation is not necessarily efficient away from steady states (see [17]), and we refer to [4] as one of the few results on the stability of well-balanced schemes. Accordingly, more robust well-balanced schemes are sought.

The second aim in this paper is to address the question of a well-balanced simulation which preserves discrete versions of the steady states (1.11) and (1.12a). At first glance, the two aims of entropy stability and well-balancing may seem unrelated. To clarify this matter, assume that U is a steady state of the one-dimensional shallow water equation (1.2); the energy balance (1.6) then implies that $(H(U) + J_1)_x \equiv 0$. The flux term $H + J_1$ may be rewritten as

$$H(U) + J_1 = hu\left(\frac{u^2}{2} + g(h+b)\right) = mp,$$

where m and p are the equilibrium variables defined in (1.12b). Hence, in a one-dimensional steady state, the conservation of momentum and energy implies constancy of the equilibrium variable p, leading to the preservation of the steady state. This connection manifests itself at the discrete level, whence our energy preserving scheme also preserves a discrete version of the steady state (1.12a).

Energy conservative schemes produce oscillations at shocks. This is expected as energy needs to be dissipated at shocks. To obtain an energy stable scheme, suitable numerical diffusion operators have to be designed. In the first part of this paper, We combine the novel numerical diffusion operator of [8] together with the energy conservative fluxes and show that the resulting scheme is energy stable. Furthermore, this energy stable scheme also preserves the lake at rest. However, this choice of numerical diffusion operator may not preserve the general equilibrium state (1.12a), even though the energy preserving scheme preserves a discrete version of such steady states. Hence, we introduce *another* novel numerical diffusion operator, based on the equilibrium variables, that is well balanced with respect to discrete versions of the general equilibrium state (1.12a). Note that both these diffusion operators are added to the *same* energy preserving scheme.

The resulting schemes are extremely simple to code and computationally cheap: no algebraic equations are solved, and by non-oscillatory reconstructions we achieve second-order accuracy. Numerical experiments demonstrating the computational efficiency of the well-balanced energy preserving and energy stable schemes are presented in Section 2.5.

2. Well-balanced schemes for the one-dimensional problem

For simplicity, we start with the one-dimensional form of the shallow water equations (1.2). This system is an example of the general one-dimensional system of conservation laws

$$U_t + f(U)_x = -s(x,U), \tag{2.1}$$

with U the vector of unknowns, $f(U)$ the flux vector and $s(x,U)$ the source term.

Smooth solutions of (1.2) satisfy the energy equality

$$E(U)_t + \big(H(U) + J(U)\big)_x = 0, \tag{2.2}$$

where $E(U) = \frac{1}{2}\left(hu^2 + gh^2\right) + ghb$, $H(U) = \frac{1}{2}hu^3 + gh^2u$ and $J(U) = ghub$ are the energy and energy flux functions. We postulate that weak solutions satisfy a weak form of the corresponding inequality

$$\left(\frac{1}{2}\left(hu^2 + gh^2\right) + ghb\right)_t + \left(\frac{1}{2}hu^3 + ghh(h+b)\right)_x \leq 0.$$

2.1. **Energy conservative schemes.**

Our aim is to design a FV schemes for (1.2) which satisfy a discrete form of the energy conservation (2.2). We consider FV schemes on a uniform mesh $\{x_i\}_i$ in their semi-discrete form

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F_{i+1/2} - F_{i-1/2}\right) - S_i. \tag{2.3}$$

Here, U_i is the cell average on $I_i := [x_{i-1/2}, x_{i+1/2}]$, $F_{i+1/2}$ is the numerical flux at the interface $x_{i+1/2}$ and S_i is a suitable discretization of the source term in (1.2).

We begin with the following characterization of energy conservative schemes. These schemes will be characterized in terms of the *entropy variables* $V := \partial_U E(U)$. For the one-dimensional shallow water equations, we have

$$V = \begin{bmatrix} V^{(1)} \\ V^{(2)} \end{bmatrix} = \begin{bmatrix} g(h+b) - \frac{u^2}{2} \\ u \end{bmatrix}. \tag{2.4}$$

The *energy potential* is the function $\Psi := \langle V,\ f\rangle - H = \frac{1}{2}guh^2$. Throughout the paper, we use

$$[\![a]\!]_{i+1/2} := a_{i+1} - a_i, \qquad \overline{a}_{i+1/2} := \frac{1}{2}(a_i + a_{i+1}),$$

to denote the jump, and respectively, the average of a quantity a across the interface $x_{j+1/2}$.

Lemma 2.1. *A numerical flux $F_{i+1/2}$ is energy conservative if*

$$\langle [\![V_{i+1/2}]\!],\ F_{i+1/2}\rangle = [\![\Psi]\!]_{i+1/2} + g[\![b]\!]_{i+1/2}\overline{h}_{i+1/2}\overline{u}_{i+1/2} \tag{2.5}$$

The corresponding FV scheme then satisfies the energy conservation statement

$$\frac{d}{dt}E_i = -\frac{1}{\Delta x}\left(\widehat{H}_{i+1/2} - \widehat{H}_{i-1/2}\right), \tag{2.6a}$$

where the numerical energy flux $\widehat{H}$ *is given by*

$$\widehat{H}_{i+1/2} := \langle \overline{V}_{i+1/2},\ F_{i+1/2}\rangle - \overline{\Psi}_{i+1/2} - \frac{g}{4}\overline{h}_{i+1/2}\llbracket u\rrbracket_{i+1/2}\llbracket b\rrbracket_{i+1/2}. \tag{2.6b}$$

In particular, the total energy is preserved: $\displaystyle\sum_i E_i(t)\Delta x \equiv \sum_i E_i(0)\Delta x.$

Proof. The proof is a modification of the energy conserving statement in [8]. Taking the inner product of (2.3) with $V_i = \partial_U E(U_i)$ yields

$$\begin{aligned}
\frac{d}{dt}E_i =& -\frac{1}{\Delta x}\left(\langle V_i,\ F_{i+1/2}\rangle - \langle V_i,\ F_{i-1/2}\rangle\right) - \langle V_i,\ S_i\rangle \\
\overbrace{\equiv}^{(\#1)}& -\frac{1}{\Delta x}\left(\left(\langle \overline{V}_{i+1/2},\ F_{i+1/2}\rangle - \frac{1}{2}\langle \llbracket V_{i+1/2}\rrbracket,\ F_{i+1/2}\rangle\right) - \left(\langle \overline{V}_{i-1/2},\ F_{i-1/2}\rangle + \frac{1}{2}\langle \llbracket V_{i-1/2}\rrbracket,\ F_{i-1/2}\rangle\right)\right) \\
& - \frac{g}{2\Delta x}u_i\left(\overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2}\right) \\
\overbrace{=}^{(\#2)}& -\frac{1}{\Delta x}\Bigg(\left(\widehat{H}_{i+1/2} + \overline{\Psi}_{i+1/2} + \frac{g}{4}\overline{h}_{i+1/2}\llbracket u\rrbracket_{i+1/2}\llbracket b\rrbracket_{i+1/2} - \frac{1}{2}\llbracket \Psi\rrbracket_{i+1/2}\right) \\
& - \left(\widehat{H}_{i-1/2} + \overline{\Psi}_{i-1/2} + \frac{g}{4}\overline{h}_{i-1/2}\llbracket u\rrbracket_{i-1/2}\llbracket b\rrbracket_{i-1/2} + \frac{1}{2}\llbracket \Psi\rrbracket_{i-1/2}\right)\Bigg) \\
& - \frac{g}{2\Delta x}u_i\left(\overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2}\right) \\
\overbrace{=}^{(\#3)}& -\frac{1}{\Delta x}\left(\widehat{H}_{i+1/2} - \widehat{H}_{i-1/2}\right).
\end{aligned}$$

The first step #1 is a direct consequence of the identities $V_i \equiv \overline{V}_{i\pm1/2} \mp \frac{1}{2}\llbracket V_{i\pm1/2}\rrbracket$; step #2 follows from (2.5) and (2.6b) and step #3 is verified by cancellation of terms. □

Motivated by the energy preserving scheme for shallow water equations with flat bottom topography proposed in a recent paper [8], we propose the following numerical flux and source discretizations:

$$F^{\mathrm{EC}}_{i+1/2} = \begin{bmatrix} \overline{h}_{i+1/2}\overline{u}_{i+1/2} \\ \frac{g}{2}\overline{h^2}_{i+1/2} + \overline{h}_{i+1/2}\left(\overline{u}_{i+1/2}\right)^2 \end{bmatrix}, \quad S^{\mathrm{EC}}_i = \begin{bmatrix} 0 \\ \frac{g}{2\Delta x}\left(\overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2}\right) \end{bmatrix}. \tag{2.7}$$

The numerical flux $F^{\mathrm{EC}}_{i+1/2}$ is exactly the same as the energy conserving scheme proposed in [8] in connection with the shallow water equations with flat bottom. It is the discretization of the source which is different from the standard one in (1.9), which enables us to obtain the desired property of energy conservation in the presence of varying bottom topography. The FV scheme (2.3) with the EC flux and source in (2.7) amount to

$$\begin{aligned}
\frac{d}{dt}h_i =& -\frac{1}{\Delta x}\left(\overline{h}_{i+1/2}\overline{u}_{i+1/2} - \overline{h}_{i-1/2}\overline{u}_{i-1/2}\right) \\
\frac{d}{dt}(h_i u_i) =& -\frac{1}{\Delta x}\left(\overline{h}_{i+1/2}\left(\overline{u}_{i+1/2}\right)^2 + \frac{g}{2}\overline{h^2}_{i+1/2} - \overline{h}_{i-1/2}\left(\overline{u}_{i-1/2}\right)^2 - \frac{g}{2}\overline{h^2}_{i-1/2}\right) \\
& - \frac{g}{2\Delta x}\left(\overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2}\right).
\end{aligned} \tag{2.8}$$

We refer to (2.8) as the *energy conservative* (EC) scheme, analogous to the nomenclature in [8]. Our next theorem shows that the EC scheme (2.8) does both: it is energy conservative and it is well-balanced in the sense of preserving a discrete form of the lake at rest (2.9). Recall that the lake at rest steady state (1.11) in the one-dimensional case is given by

$$u \equiv 0, \qquad h + b \equiv \text{constant}. \tag{2.9}$$

Theorem 2.2. *The EC scheme* (2.8) *satisfies the following properties.*

(i) Accuracy: *It is a second-order accurate approximation of the one-dimensional shallow water system* (1.2)*.*

(ii) Energy conservation: *It is an energy conserving scheme, i.e.,* (2.6) *holds.*

(iii) Well-balanced: *It preserves the lake at rest – given initial data*

$$u_i \equiv 0, \qquad h_i + b_i \equiv \text{constant} \qquad \forall\ i, \tag{2.10a}$$

then the solution computed by (2.8) *satisfies*

$$\frac{d}{dt}h_i \equiv 0, \qquad \frac{d}{dt}(h_i u_i) \equiv 0 \qquad \forall\ i. \tag{2.10b}$$

Proof. A straightforward truncation error analysis shows that the local truncation error is $O(\Delta x^2)$ which confirms (i). The energy conservation (ii) follows by verifying that the numerical flux (2.8) satisfies (2.5). We remark that both the special form of the fluxes in (2.7) and the specific structure of the source term in (2.7) are crucial for obtaining the discrete energy identity. Finally, to prove (iii), we employ the identity

$$\overline{h^2}_{i+1/2} - \overline{h^2}_{i-1/2} \equiv \overline{h}_{i+1/2}[\![h]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![h]\!]_{i-1/2}. \tag{2.11}$$

Assume that h_i, b_i, u_i are such that the discrete lake at rest condition (2.10a) is satisfied. Then $\overline{u}_{i+1/2} \equiv 0$ for all i. Plugging this into the first equation of (2.8), we see that the fluxes are zero and

$$\frac{d}{dt}h_i \equiv 0 \qquad \forall\ i,$$

thus proving the first assertion in (2.10b). Using $\overline{u}_{i+1/2} \equiv 0$ in the second equation of (2.8), we obtain

$$\frac{d}{dt}(h_i u_i) = -\frac{g}{\Delta x}\left(\overline{h^2}_{i+1/2} - \overline{h^2}_{i-1/2} + \overline{h}_{i+1/2}[\![b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![b]\!]_{i-1/2}\right).$$

Using (2.11), this expression reduces to

$$\frac{d}{dt}(h_i u_i) = -\frac{g}{\Delta x}\left(\overline{h}_{i+1/2}[\![h+b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![h+b]\!]_{i-1/2}\right).$$

As the data satisfies the discrete lake at rest (2.10a), we have $[\![h+b]\!] \equiv 0$, and so the above equation reduces to

$$\frac{d}{dt}(h_i u_i) \equiv 0.$$

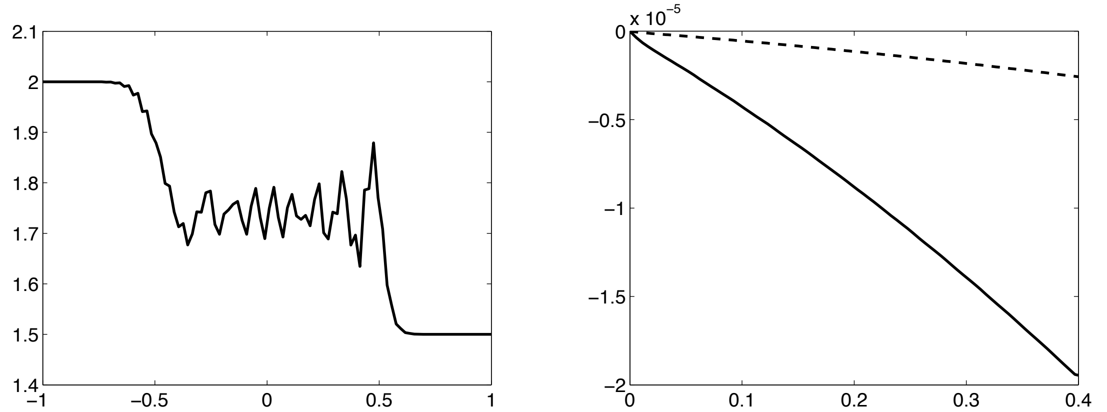
□

This theorem establishes that the EC scheme (2.8) conserves energy and preserves a discrete version of the lake at rest. Furthermore, it is very easy to implement and computationally cheap (the computational cost is similar to evaluating the fluxes and the source in (1.2)). This should be contrasted with other well-balanced schemes in literature like those in [4, 25] where the scheme is more complicated in the design and implementation.

2.2. **Numerical experiments.** We test the EC scheme on some numerical experiments in order to ascertain its numerical performance. To begin with, we simulate (1.2) with a flat bottom topography (i.e b is constant) and consider a dam-break problem with the initial data

$$h(x,0) = \begin{cases} 2 & \text{if} \quad x < 0 \\ 1.5 & \text{if} \quad x > 0 \end{cases} \qquad u(x,0) \equiv 0. \tag{2.12}$$

The computational domain is $[-1,1]$ and the exact solution consists of a left-going rarefaction and a right-going shock. We present the solution computed with the EC scheme and 100 mesh points in Figure 1. The figure shows that the EC scheme computes the rarefaction and the shock quite accurately, but at the expense of large post-shock oscillations. These oscillations are to be expected as energy must be dissipated across the shock, although the energy identity (2.6a) forces the scheme to preserve energy in each cell. Thus, the inertial term



(a) Height at $t = 0.4$

(b) Relative energy over time. Solid line uses CFL=0.4, dotted line CFL=0.2

FIGURE 1. The EC scheme computes a dambreak problem

in (1.2) transfers energy to lowest resolved scale (i.e mesh size) in the form of oscillations. These oscillations have been studied extensively (see [10]) and are described in detail in [8]. The numerical energy conservation is demonstrated on the right panel of Figure 1, where we plot the total energy over time. As shown in the figure, the time stepping produces small energy dissipation errors. These errors are reduced considerably by decreasing the CFL number, and hence the time step. This example is reproduced from [8] and serves to illustrate some features of the EC scheme for a flat bottom topography.

2.2.1. *Lake at rest.* Next, we present a standard numerical experiment first considered in [12] and used in numerous papers [23, 4] and other references therein. The bottom topography is a parabolic "bump" in the middle of the domain $[0, 20]$,

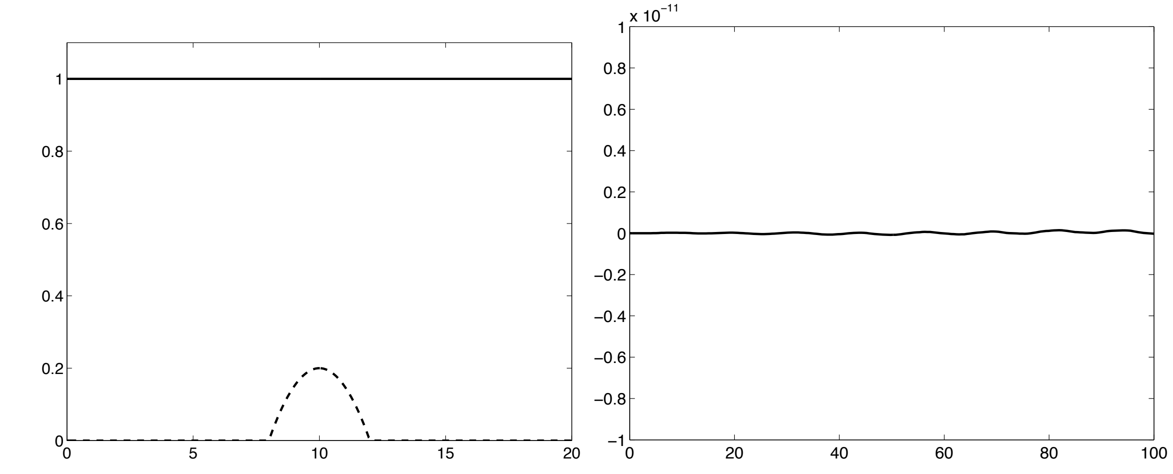
$$b(x) = \begin{cases} \frac{4-(x-10)^2}{20} & \text{if } |x-10| < 2 \\ 0 & \text{else.} \end{cases} \tag{2.13}$$

We impose the lake at rest initial condition $u_i \equiv 0$, $h_i + b_i \equiv 1$. The gravitational constant is set to $g = 9.812$, and we impose Neumann ("open") boundary conditions. The scheme is run till time $T = 100$ and the resulting states are shown in Figure 2. As shown in this figure, the steady state is preserved exactly, even at this large time. This is a consequence of Theorem 2.2 establishing that the EC scheme preserves the lake at rest (2.10a). Furthermore, the energy vs. time graph in Figure 2 shows that the energy errors are very small (of the order of 10^{-12}). These errors are due to the discretization in the time stepping. Thus, the EC scheme preserves the steady state as well as energy.

2.2.2. *Perturbations of lake at rest.* The main interest in the design of well-balanced schemes is to employ them in computing perturbations of interesting steady states. A steady state like the lake at rest is known a priori and is not interesting to compute by itself. We perturb the lake at rest in the previous numerical experiment by letting

$$h(x, 0) = \begin{cases} 1.01 - b(x) & \text{if } |x-6| < 1/4, \\ 1 - b(x) & \text{else} \end{cases} \tag{2.14}$$

and u and b as above. Hence, the perturbation is a very small disturbance of the lake at rest and we seek to study how this disturbance propagates in time. The results are computed with the EC scheme with 200 mesh points. The resulting height and the deviation from the steady state are shown in Figure 3. The deviation from the steady state shown in the right panel of Figure 3 clearly shows that the EC scheme is able to approximate

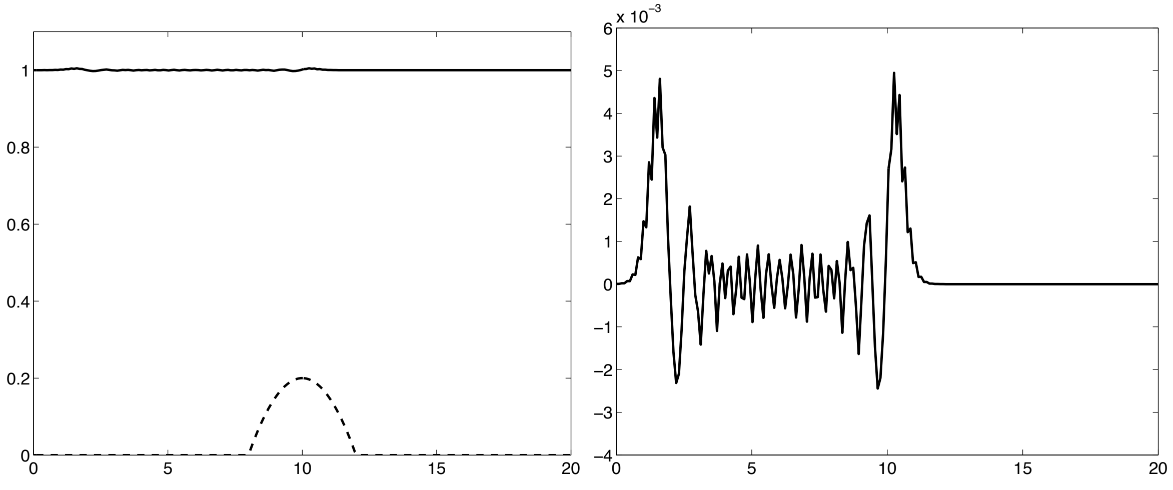


(a) Water level $h + b$ (solid line) and bottom topography (dotted line)

(b) Relative change in energy over time

Figure 2. Lake at rest at $t = 100$ using 200 mesh points

both waves. This is a consequence of its ability to preserve the steady state. There are small oscillations trailing the right going wave; again, this is to be expected, as the EC scheme preserves energy even across a shock.



(a) Water level $h + b$ (solid line) and bottom topography (dotted line)

(b) Deviation from the steady state

Figure 3. Lake at rest with perturbation at $t = 1.5$

2.3. **Energy stable scheme — first-order diffusion.** The numerical examples above show that the EC scheme preserves energy and the lake at rest steady state. Hence, it can compute small perturbations of the steady state. However, the scheme will lead to non-physical oscillations due to the lack of energy dissipation at shocks. This problem can be tackled by using efficient numerical diffusion operators [33, 36]. Our aim is to design a numerical diffusion operator that dissipates energy (and hence is energy stable) and preserves the lake at rest steady state. A novel strategy for designing numerical diffusion operators for the shallow water equations with flat bottom topography was presented in a recent paper [8]. We omit details of how this numerical diffusion operator can be derived and give the explicit expression of this operator below. The interested reader can consult [8, Lemma 4.3]. Given the left and right states, $U_i = [h_i, (hu)_i]^\top$ and $U_{i+1} = [h_{i+1}, (hu)_{i+1}]^\top$, we let $R_{i+1/2}$ and $\Lambda_{i+1/2}$ denote the eigenvector and eigenvalue matrices associated with the Roe decomposition [27] of the left- and right-side pair (U_i, U_{i+1}),

$$R_{i+1/2} = \frac{1}{\sqrt{2g}}\begin{bmatrix} 1 & 1 \\ \lambda_- & \lambda_+ \end{bmatrix}, \qquad \lambda_\pm := \overline{u}_{i+1/2} \pm \sqrt{g\overline{h}_{i+1/2}}, \tag{2.15a}$$

and

$$|\Lambda_{i+1/2}| = \begin{bmatrix} |\lambda_-| & 0 \\ 0 & |\lambda_+| \end{bmatrix}. \tag{2.15b}$$

The numerical diffusion coefficient matrix $D^{\mathrm{ES1}}_{i+1/2} \equiv D^{\mathrm{ES1}}(U_i, U_{i+1})$ is then given by

$$D^{\mathrm{ES1}}_{i+1/2} := R_{i+1/2}|\Lambda_{i+1/2}|R^\top_{i+1/2}. \tag{2.16a}$$

Note that the diffusion matrix in (2.16a) is positive definite. It generalizes the diffusion operator proposed in [8] for the case of a flat bottom topography. The resulting FV flux is

$$F^{\mathrm{ES1}}_{i+1/2} = F^{\mathrm{EC}}_{i+1/2} - \frac{1}{2}D^{\mathrm{ES1}}_{i+1/2}[\![V]\!]_{i+1/2}, \tag{2.16b}$$

where $F^{\mathrm{EC}}_{i+1/2}$ is the energy conservative flux (2.7) and $V = [g(h+b) - \frac{u^2}{2},\ u]^\top$ is the vector of energy variables. We remark that the above flux differs from the standard Roe flux [27] in two essential aspects: (i) the standard central average flux is replaced by an energy conserving flux and (ii) the numerical diffusion matrix acts on the jump in entropy variables rather than the conservative ones. The resulting scheme reads as

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F^{\mathrm{ES1}}_{i+1/2} - F^{\mathrm{ES1}}_{i-1/2}\right) - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![b]\!]_{i-1/2} \end{bmatrix} \tag{2.16c}$$

This scheme will be termed as the *first-order energy stable* (ES1) scheme in the remainder of the paper. Its main properties are summarized below.

Theorem 2.3. *The ES1 scheme* (2.16) *satisfies the following.*

(i) Accuracy: *It is a first-order accurate approximation of the one-dimensional shallow water equations* (1.2)*.*

(ii) Stability: *It satisfies the discrete energy identity*

$$\begin{aligned} \frac{d}{dt}E_i = &-\frac{1}{\Delta x}\left(\widetilde{H}_{i+1/2} - \widetilde{H}_{i-1/2}\right) \\ &- \frac{1}{4\Delta x}\Big(\langle[\![V_{i+1/2}]\!],\ D^{\mathrm{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle + \langle[\![V_{i-1/2}]\!],\ D^{\mathrm{ES1}}_{i-1/2}[\![V_{i-1/2}]\!]\rangle\Big), \end{aligned} \tag{2.17a}$$

where the energy dissipative numerical flux, $\widetilde{H}$ is given by

$$\widetilde{H}_{i+1/2} = \widehat{H}_{i+1/2} + \frac{1}{2}\langle\overline{V}_{i+1/2},\ D^{\mathrm{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle. \tag{2.17b}$$

Summing (2.17a) *we obtain*

$$\frac{d}{dt}\sum_i E_i\Delta x = -\frac{1}{2}\sum_i\langle[\![V_{i+1/2}]\!],\ D^{\mathrm{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle \leq 0.$$

which quantifies the precise energy dissipation of the our ES1 scheme (2.16)*.*

(iii) Well-balanced: *It preserves the discrete lake at rest* (2.10)*.*

Proof. The proof of (i) is straightforward. The proof of (ii) follows the proof of (2.6a) and we omit the details. As noted in [36, Corollary 5.1], it is essential that we use here a positive numerical diffusion matrix which acts on the jump in entropy variables. To prove (iii), we assume that the data satisfy (2.10a). Then we have

$$[\![u]\!]_{i+1/2} \equiv 0 \qquad \text{and} \qquad [\![h+b]\!]_{i+1/2} \equiv 0.$$

Consequently, by the definition of the energy variables, $[\![V]\!]_{i+1/2} \equiv 0$. Hence the diffusion operator (2.16a) drops out, and the scheme reduces to the EC scheme. Thus, by Theorem 2.2(iii), we have

$$\frac{d}{dt}h_i \equiv 0 \qquad \text{and} \qquad \frac{d}{dt}(h_i u_i) \equiv 0,$$

and the discrete lake at rest is preserved by the ES1 scheme. □

2.4. **Energy-stable scheme — second-order diffusion.** The ES1 scheme is restricted to first-order accuracy and will lead to smeared solutions. Higher order of accuracy can be recovered by using suitable piecewise polynomial reconstructions. The aim is to replace the piecewise constant cell averages U_i in (2.3) with a non-oscillatory piecewise linear reconstruction as in [19].

We will carry out the reconstruction in terms of the energy variables and when needed, convert them to the conservative variables. Define the numerical derivative of the energy variables V_i as

$$V_i' = \text{minmod}\left(\frac{V_{i+1}-V_i}{\Delta x}, \ \frac{V_i - V_{i-1}}{\Delta x}\right), \tag{2.18}$$

where the minmod function is defined as

$$\text{minmod}(a,b) = \begin{cases} \text{sign}\,(a)\min\{|a|,|b|\} & \text{if sign}\,(a) = \text{sign}\,(b) \\ 0 & \text{otherwise.} \end{cases}$$

(2.18) is evaluated component-wise. We now consider the piecewise linear reconstruction of the energy variables V in cell I_i:

$$\widetilde{V}_i(x) = V_i + V_i'(x - x_i) \qquad x \in I_i.$$

The *reconstructed pointvalues* along the edges of this cell are given by $V_i^r := \widetilde{V}_i(x_{i+1/2})$ and $V_{i+1}^\ell := \widetilde{V}_{i+1}(x_{i+1/2})$. The second-order version of the ES1 flux diffusion (2.16a) is defined in terms of these reconstructed point values,

$$D_{i+1/2}^{\text{ES2}} := D^{\text{ES1}}\big(V_i^r, V_{i+1}^\ell\big), \tag{2.19a}$$

where $D^{\text{ES1}} = D^{\text{ES1}}(\cdot,\cdot)$ is the first-order diffusion matrix in (2.16a). Thus, the matrices R and $|\Lambda|$ are now defined in terms of differences and averages of V_i^r and V_{i+1}^ℓ, and the resulting flux amounts to

$$F_{i+1/2}^{\text{ES2}} = F_{i+1/2}^{\text{EC}} - \frac{1}{2}D_{i+1/2}^{\text{ES2}}\big(V_{i+1}^\ell - V_i^r\big), \tag{2.19b}$$

where $F_{i+1/2}^{\text{EC}}$ is the energy conservative flux in (2.7). The resulting second-order scheme reads

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F_{i+1/2}^{\text{ES2}} - F_{i-1/2}^{\text{ES2}}\right) - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![b]\!]_{i-1/2} \end{bmatrix} \tag{2.19c}$$

This scheme will be termed as the *second-order energy stable* (ES2) scheme in the remaining part of the paper. Properties of this scheme are summarized below.

Theorem 2.4. *The ES2 scheme* (2.19) *is a second-order accurate approximation of the one-dimensional shallow water system* (1.2) *and it preserves the discrete lake at rest* (2.10a)*.*

Proof. The second-order accuracy follows by noting that the energy conservative flux is second-order accurate and the jump in the reconstructed values is of order $\mathcal{O}\big(|[\![V_{i+1/2}]\!]|^2\big)$.

To prove that the ES2 scheme (2.19c) preserves the lake at rest, observe that when the data satisfies (2.10a), we have

$$u_i \equiv 0 \qquad \text{and} \qquad [\![h+b]\!]_{i+1/2} \equiv 0;$$

hence $[\![V]\!]_{i+1/2} \equiv 0$. Therefore, by the definition of the slope in (2.18), we obtain $V_i' \equiv 0$, so

$$V_i^r = V_{i+1}^\ell \equiv \text{Constant}.$$

Consequently, the jump in energy variables, $\big(V^\ell_{i+1} - V^r_i\big)$ in (2.19a) vanishes, and we follow the same argument as in the proof of Theorem 2.3(iii) to conclude that

$$\frac{d}{dt}h_i \equiv 0 \qquad \text{and} \qquad \frac{d}{dt}(h_i u_i) \equiv 0.$$

Hence, the discrete lake at rest is preserved. Note that the key point is the use of energy variables in the reconstruction step which allows us to balance the reconstruction at the steady state. □

Remark 2.5. Second-order accuracy is maintained without updating the diffusion matrix, $D^{\text{ES1}}_{i+1/2} \mapsto D^{\text{ES2}}_{i+1/2}$; thus, Theorem 2.4 applies if one replaces (2.19b) with

$$F^{\text{ES2}}_{i+1/2} = F^{\text{EC}}_{i+1/2} - \frac{1}{2}D^{\text{ES1}}_{i+1/2}\big(V^\ell_{i+1} - V^r_i\big).$$

The resulting scheme reads

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F^{\text{ES2}}_{i+1/2} - F^{\text{ES2}}_{i-1/2}\right) - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![b]\!]_{i-1/2}\end{bmatrix} \tag{2.20}$$

Remark 2.6. Let us try to verify the energy stability of (2.20): we take the inner product of (2.20) against V_i; arguing along the lines of Lemma 2.1 (summation by parts) now yields

$$\frac{d}{dt}\sum_i E_i \Delta x = -\frac{1}{2}\sum_i \langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}\big(V^\ell_{i+1} - V^r_i\big)\rangle. \tag{2.21}$$

Thus, the main point is to show that the jump in the *reconstructed* energy variables,

$$V^\ell_{i+1} - V^r_i = [\![V_{i+1/2}]\!] - \big(V'_{i+1} + V'_i\big)\frac{\Delta x}{2} \equiv [\![V_{i+1/2}]\!] - \overline{V'}_{i+1/2}\Delta x.$$

is dominated by the jump in the original cell averages, $[\![V_{i+1/2}]\!]$, and to this end, it suffices to show that the energy dissipation terms on the right of (2.21)

$$\langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}\big(V^\ell_{i+1} - V^r_i\big)\rangle = \langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle - \langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}\overline{V'}_{i+1/2}\Delta x\rangle,$$

are positive. Cauchy-Schwarz inequality induced by the positive D implies

$$\langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}\overline{V'}_{i+1/2}\Delta x\rangle \le \langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle^{1/2} \cdot \langle \overline{V'}_{i+1/2}\Delta x,\ D^{\text{ES1}}_{i+1/2}\overline{V'}_{i+1/2}\Delta x\rangle^{1/2},$$

and thus it remains to show that

$$\langle \overline{V'}_{i+1/2}\Delta x,\ D^{\text{ES1}}_{i+1/2}\overline{V'}_{i+1/2}\Delta x\rangle \le \langle [\![V_{i+1/2}]\!],\ D^{\text{ES1}}_{i+1/2}[\![V_{i+1/2}]\!]\rangle. \tag{2.22}$$

By properties of the minmod limiter, there exists a diagonal matrix Θ such that

$$\overline{V'}_{i+1/2}\Delta x = \Theta[\![V_{i+1/2}]\!], \qquad \Theta = \begin{bmatrix}\theta_1 & 0\\ 0 & \theta_2\end{bmatrix}, \quad |\theta_1|, |\theta_2| \le 1. \tag{2.23}$$

Thus, the desired inequality (2.22) follows provided $D - \Theta D \Theta \ge 0$. In general, this inequality fails since we are limiting the energy variables. This procedure shows that one needs to work with the Roe variables instead: we set the limited slopes

$$W'_i := \text{minmod}\left(\frac{R^\top_{i+1/2}(V_{i+1} - V_i)}{\Delta x},\ \frac{R^\top_{i-1/2}(V_i - V_{i-1})}{\Delta x}\right),$$

and the reconstructed values to be

$$V^r_i := V_i + \left(R^\top_{i+1/2}\right)^{-1} W'_i\frac{\Delta x}{2}, \qquad V^\ell_{i+1} := V_{i+1} - \left(R^\top_{i+1/2}\right)^{-1} W'_{i+1}\frac{\Delta x}{2}.$$

Now (2.23) is replaced by

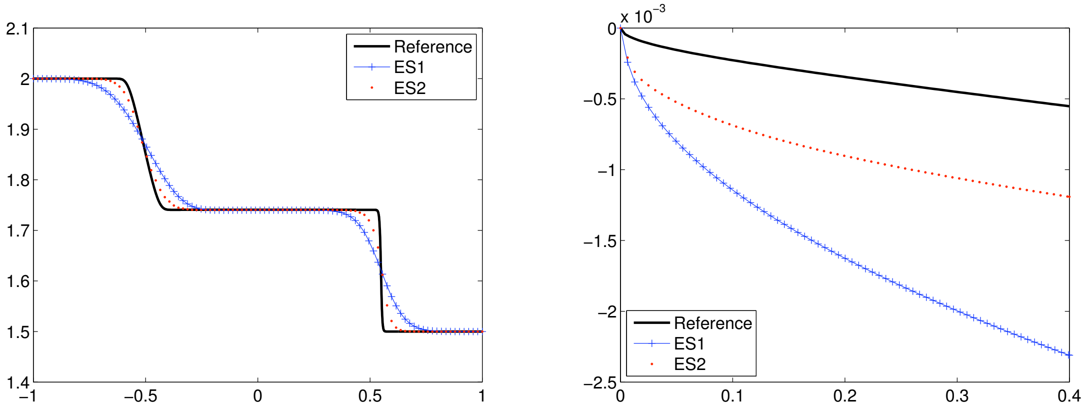
$$\overline{V'}_{i+1/2}\Delta x = \Sigma[\![V_{i+1/2}]\!], \qquad \Sigma := \left(R^\top_{i+1/2}\right)^{-1}\begin{bmatrix}\theta_1 & 0\\ 0 & \theta_2\end{bmatrix}\left(R^\top_{i+1/2}\right), \quad |\theta_1|, |\theta_2| \le 1,$$

and positivity follows since

$$D - \Sigma^\top D \Sigma = R|\Lambda|R^\top - R\Theta|\Lambda|\Theta R^\top = R|\Lambda|\left(I - \Theta^2\right)R^\top \ge 0.$$

2.5. **Numerical experiments.**

2.5.1. *Dambreak problem.* We repeat the numerical experiment of section 2.2 with the ES1 and ES2 schemes, and we present the results in Figure 4. The figure shows that the first-order ES1 scheme computes the solution with some smearing at both the rarefaction and the shock wave. The accuracy is increased considerably by using the second-order ES2 scheme. Both schemes dissipate energy, with the energy dissipation in ES2 being much lower than the ES1 scheme. Observe that using the numerical diffusion operators eliminates the post shock oscillations with the EC scheme observed in Figure 1.



(a) Height

(b) Energy vs. time

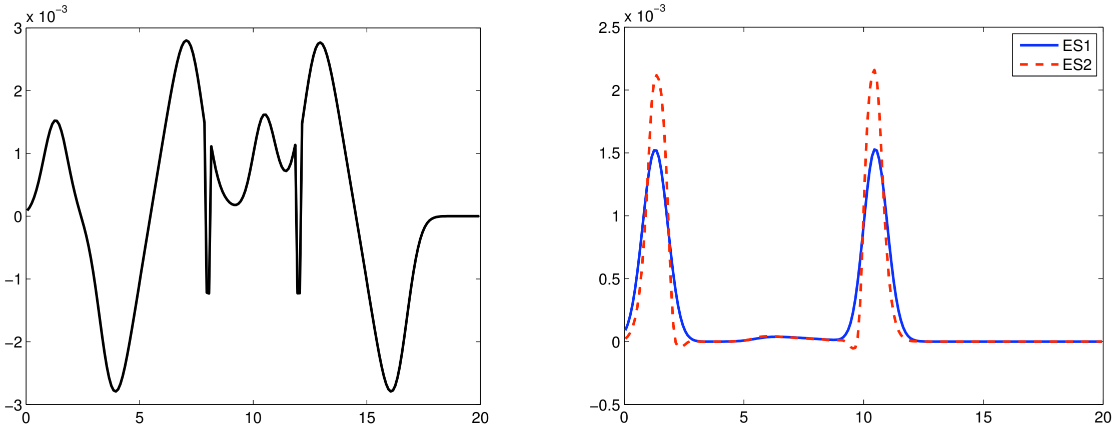
FIGURE 4. Solutions computed with the first and second-order versions of the energy stable scheme, ES1 and ES2 with 100 mesh points.

2.5.2. *Lake at rest.* Next, we use the ES1 and ES2 schemes to compute the lake at rest described in section 2.2.2. The bottom topography is given in (2.13) and the data satisfy $u_i \equiv 0$ and $h_i + b_i \equiv 1$. We compute both the ES1 and ES2 schemes on a sequence of meshes for this steady state and present the results in Table 1. In this table, we compute the L^1 error in the height at time $t = 10$ on a sequence of meshes. For the sake of comparison, we also present results with the EC scheme (2.8) and the standard Roe scheme [27]. As shown in the table, the EC, ES1 and ES2 schemes are well-balanced and preserve the lake at rest up to machine precision. On the other hand, the standard Roe scheme is not well-balanced and leads to errors of the order of truncation error. These errors might accumulate in time and lead to large discrepancies when perturbations of the steady state are computed.

N	Roe	EC	ES1	ES2
50	2.76e-2	6.27e-14	1.92e-18	3.17e-16
100	7.60e-3	1.62e-13	2.14e-18	4.48e-17
200	2.02e-3	6.74e-13	3.35e-18	2.34e-16
400	5.15e-4	1.76e-12	2.22e-17	1.04e-15

TABLE 1. The L^1 error in height for the lake at rest with different schemes on a sequence of meshes at time $t = 10$.

2.5.3. *Perturbed lake at rest.* We consider a small perturbation of the lake at rest given by (2.14). Since the perturbations are very small, they will not be clearly visible in a plot showing both the height and the bottom topography. In order to compare different schemes, we show the deviation from the steady state in Figure 5 for the standard Roe scheme and the ES1 and ES2 schemes. The figure clearly shows that the Roe scheme computes an incorrect solution; the exact solution should consist of a left and a right going wave. On the other hand, both the ES1 and ES2 schemes compute the perturbation quite well. The first-order ES1 scheme dissipates both the left and the right going waves somewhat, but accuracy is recovered with the second-order ES2 scheme. Still, the wave heights are lower than those computed with the EC scheme (Figure 3). The results are comparable to those obtained in [4] and other similar references.



(a) Steady state deviation, Roe.

(b) Steady state deviation, ES1, ES2.

Figure 5. Lake at rest with perturbation at $t = 1.5$ on a mesh with 200 mesh points

The above experiments show that the EC scheme is energy preserving and preserves the lake at rest. It can be used to compute perturbations of the lake at rest and approximates the wave forms quite well. However, there are unphysical oscillations due to lack of energy dissipation at shocks. These oscillations can be eliminated by using the first-order ES1 scheme. This scheme dissipates energy and preserves the lake at rest. However, it leads to smearing and loss of accuracy. Second-order accuracy is recovered using the ES2 scheme. This scheme preserves the steady state exactly and is quite robust in computing perturbations of steady states.

3. Well-balanced schemes with moving equilibrium states

The lake at rest (2.9) is a very important steady state but there are other interesting steady states of (1.2). By asserting $h_t = (hu)_t = 0$ in (1.2), one finds that any steady state must satisfy

$$m \equiv \text{constant}, \qquad p \equiv \text{constant}, \tag{3.1}$$

where m and p are defined in (1.12b). The values $P := [m,\ p]^\top$ are called the *equilibrium variables*; steady states are exactly those in which the equilibrium variables are constant in space. Note that the lake at rest (2.9) is a special case of (3.1) with $m \equiv 0$.

We begin with the *classification of equilibrium states*. Following [26], we can classify all steady states based on properties of the vector of equilibrium variables. Note that the condition (3.1) does not easily translate into conditions on the vector of conservative variables $U = [h,\ m]^\top$, as the condition

$$p(h, m, b) \equiv C \tag{3.2}$$

in (3.1) is nonlinear in both h and m. Fixing m and b and viewing p as a function of h, simple calculations show that the function $p(h)$ is convex and attains its unique minimum at the point

$$h_0 = \frac{m^{\frac{2}{3}}}{g^{\frac{1}{3}}}.$$

This point is exactly the point at which the Froude number $Fr := \frac{|u|}{\sqrt{gh}}$ is equal to unity. A typical example of the function $p(h)$ for fixed values of m and b is shown in Figure 6.

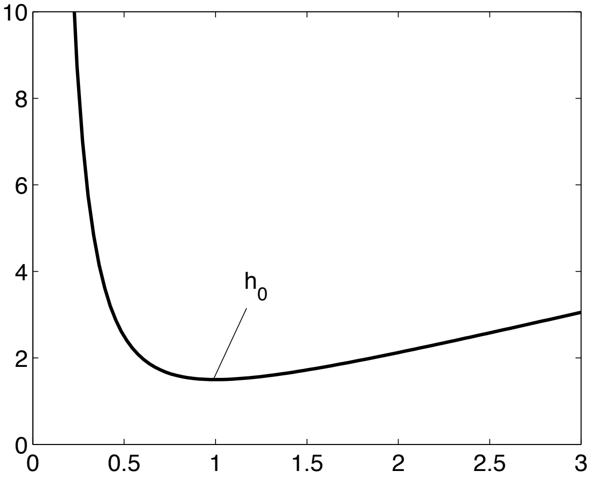


FIGURE 6. The equilibrium variable p as a function of h for fixed values of b and m.

Denote $p_0 := p(h_0)$. Given any pair $P = [m,\ p]^\top$, there are three possible cases:

1. If $p < p_0$, then there are no solutions of (3.2) and the given state is unphysical.
2. If $p = p_0$, then there is a unique solution of (3.2) corresponding to h_0 with Froude number equal to unity.
3. If $p > p_0$, then are two possible solutions of (3.2). One state corresponds to a subsonic steady state and the other to a supersonic steady state.

Since (3.2) is satisfied at every point in space, it also depends on the bottom topography b (which varies in space). We can have a steady of state of (1.2) which is entirely subsonic or supersonic. One can also obtain a steady state which is subsonic in one part of the domain and supersonic in another. Such steady states are termed *transsonic*. Hence, steady states of (1.2) show a rich variety, making numerical computations harder.

3.1. **Energy conservative scheme.** While many different numerical schemes have been designed for preserving the lake at rest (2.9), much less attention has been paid to designing schemes that preserve moving equilibrium states like (3.1). Recent papers like [26, 29] explore this problem and design numerical schemes preserving this rich hierarchy of steady states. It is natural to inquire how the schemes of the previous sections perform in this case. We start with the energy preserving EC scheme (2.8). It turns out that the EC scheme actually preserves a discrete form of the equilibrium state (3.1).

Lemma 3.1. *Define*

$$M_{i+1/2} = \overline{h}_{i+1/2}\overline{u}_{i+1/2} \qquad \textit{and} \qquad p_i = \frac{u_i^2}{2} + g(h_i + b_i). \tag{3.3}$$

The EC scheme preserves the state

$$M_{i+1/2} \equiv C_1, \qquad p_i \equiv C_2 \qquad \forall\ i \tag{3.4}$$

for constants C_1 and C_2.

Proof. We rewrite the EC scheme (2.8) as

$$\begin{aligned}\frac{d}{dt}(h_i) &= -\frac{1}{\Delta x}\left(M_{i+1/2} - M_{i-1/2}\right)\\ \frac{d}{dt}(h_i u_i) &= -\frac{1}{\Delta x}\left(\frac{1}{2}\left(\overline{h}_{i+1/2}[\![p]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![p]\!]_{i-1/2}\right) + u_i\left(M_{i+1/2} - M_{i-1/2}\right)\right).\end{aligned} \tag{3.5}$$

Plugging in the condition (3.4) clearly implies that $M_{i+1/2} = M_{i-1/2}$ and $[\![p]\!]_{i+1/2} \equiv 0$. Hence, the right hand side of (3.5) is zero and we obtain that

$$\frac{d}{dt}h_i \equiv 0 \quad \text{and} \quad \frac{d}{dt}(h_i u_i) \equiv 0,$$

thus proving the lemma. □

Remark 3.2. The quantity $M_{i+1/2}$ in (3.3) is termed the *staggered momentum*. We note that the requirement $M_{i+1/2} \equiv C$ is slightly different from demanding that $m_i \equiv C$. The difference is of the order of Δx and one must think of (3.4) as a discrete form of (3.1).

The above lemma establishes that the energy preserving EC scheme (2.8) preserves not only the discrete lake at rest (2.10a) but also a discrete form of the most general steady state (3.1). Note that we are not adding any special modifications to the EC scheme. The structure of the scheme is so robust that it preserves *any* discrete steady state.

3.2. **First-order numerical diffusion.** The EC scheme (2.8) produces oscillations near discontinuities. As seen before, we need to design suitable numerical diffusion operators like (2.16a) to eliminate oscillations and still preserve discrete steady states. However, the ES1 scheme (2.16c) and its second order version (2.19c) do not necessarily preserve the general moving equilibrium state (3.1). We need to design a special diffusion operator that preserves such steady states.

The starting point of the design is the relationship between the conservative variables U and the equilibrium variables P. The change of variable matrix is given by

$$U_P := \partial_P U = \begin{bmatrix} 1/\alpha & -u/\alpha \\ 0 & 1 \end{bmatrix}, \qquad \alpha := g - \frac{u^2}{h}.$$

The state $\alpha = 0$ corresponds to a transonic point.

The standard Roe-type numerical diffusion in a FV scheme (2.3) acts on the jump in the conservative variables

$$D_{i+1/2}[\![U]\!]_{i+1/2} = R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}[\![U]\!]_{i+1/2}.$$

It can be converted to act on the equilibrium variables, $[\![U]\!]_{i+1/2} \approx (U_P)_{i+1/2}[\![P]\!]_{i+1/2}$,

$$D_{i+1/2}[\![U]\!]_{i+1/2} \approx R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(\widetilde{U}_P)_{i+1/2}[\![P]\!]_{i+1/2}.$$

Here $R_{i+1/2}$ and $\Lambda_{i+1/2}$ are defined in (2.15) and $(\widetilde{U}_P)_{i+1/2}$ is set to be

$$\widetilde{U}_P := \begin{bmatrix} 1/\widetilde{\alpha} & -u/\widetilde{\alpha} \\ 0 & 1 \end{bmatrix}, \qquad \widetilde{\alpha} := \max\{|\alpha|, \epsilon\},$$

where ϵ is a small tolerance which handles the problem of a singularity at a sonic point. Another simple modification is required for the discrete steady states to be preserved: to this end we observe that the discrete steady state (3.4) imposes a condition on the staggered momentum rather than on the momentum. Hence we work with *averaged* equilibrium variables, $\widetilde{P}_i := [\frac{1}{2}\left(M_{i+1/2} + M_{i-1/2}\right), p_i]^\top$. In summary, we use the diffusion matrix

$$D^{\text{WB1}}_{i+1/2} = R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(\widetilde{U}_P)_{i+1/2}. \tag{3.6a}$$

The corresponding flux is then given by

$$F^{\text{WB1}}_{i+1/2} = F^{\text{EC}}_{i+1/2} - \frac{1}{2}D^{\text{WB1}}_{i+1/2}[\![\widetilde{P}]\!]_{i+1/2}, \tag{3.6b}$$

where $F^{\mathrm{EC}}_{i+1/2}$ is the energy conservative flux. The resulting FV scheme amounts to

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F^{\mathrm{WB1}}_{i+1/2} - F^{\mathrm{WB1}}_{i-1/2}\right) - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2}\end{bmatrix} \tag{3.6c}$$

This scheme is termed as the *first-order well-balanced* (WB1) scheme in the remaining part of this paper.

Lemma 3.3. *The WB1 scheme* (3.6) *is a first-order approximation of the shallow water system* (1.2) *and it preserves the discrete steady state* (3.4)*.*

Proof. The first-order accuracy of (3.6c) is easily verified. Since (3.4) imply that $\llbracket \widetilde{P}\rrbracket_{i+1/2} \equiv 0$ for all i, the diffusion operator (3.6a) drops out, and we continue as in the proof of Theorem 2.3(iii) to find that

$$\frac{d}{dt}h_i \equiv 0, \qquad \frac{d}{dt}(h_i u_i) \equiv 0.$$

□

Remark 3.4. While defining the ES1 scheme (2.16c), we used a diffusion operator defined in terms of the energy variables V. The resulting scheme was energy stable and preserved the lake at rest (2.10a). In order to preserve the more general discrete steady states (3.4), we need to use a diffusion operator (3.6a) defined in terms of the equilibrium variables P. The resulting scheme preserves the steady state (3.4). However, it might not be energy stable.

3.3. **Second-order numerical diffusion.** The WB1 scheme (3.6c) is first-order accurate. To achieve second-order accuracy, one needs to invoke a reconstruction procedure like the one described in Section 2.5 for the ES2 scheme. The reconstruction must be performed in the equilibrium variables P in order to preserve the discrete steady states (3.4). A minmod limiter similar to (2.18) is applied to the staggered equilibrium variables $\widetilde{P}$ to obtain reconstructed values $\widetilde{P}^r_i, \widetilde{P}^\ell_i$. We omit the details as they are exactly the same as in Section 2.5. The resulting diffusion matrix is

$$D^{\mathrm{WB2}}_{i+1/2} = R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(\widetilde{U}_P)_{i+1/2}, \tag{3.7a}$$

where R, Λ and $\widetilde{U}_P$ are as before. The resulting scheme is

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F^{\mathrm{WB2}}_{i+1/2} - F^{\mathrm{WB2}}_{i-1/2}\right) - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}\llbracket b\rrbracket_{i+1/2} + \overline{h}_{i-1/2}\llbracket b\rrbracket_{i-1/2},\end{bmatrix} \tag{3.7b}$$

where the numerical flux is

$$F^{\mathrm{WB2}}_{i+1/2} = F^{\mathrm{EC}}_{i+1/2} - \frac{1}{2}D^{\mathrm{WB2}}_{i+1/2}\left(\widetilde{P}^\ell_{i+1} - \widetilde{P}^r_i\right), \tag{3.7c}$$

and $F^{\mathrm{EC}}_{i+1/2}$ is the energy conservative flux defined in (2.7). This scheme is termed as the WB2 scheme.

Lemma 3.5. *The WB2 scheme is second-order accurate and preserves the discrete steady state* (3.4)*.*

We omit the proof as it is very similar to the proof of Lemma 3.3. The key fact used in the proof is that the reconstruction is done with the equilibrium variables.

3.4. **Numerical experiments with moving equilibrium states.** We consider a series of numerical experiments proposed in [39] and reported in [26].

3.4.1. *Subsonic steady state.* The domain is $[0, 20]$ and the bottom topography is given by (2.13). The initial conditions are

$$p_i \equiv 22.07, \qquad M_{i+1/2} \equiv 4.42 \qquad \forall\ i.$$

We use $g = 9.812$. The resulting states are subsonic for the whole domain. The configuration of this problem is given in Figure 7. The algebraic relation (3.2) is solved using a Newton solver. We compute solutions with the EC scheme (2.8), the WB1 scheme (3.6c) and the second-order WB2 scheme (3.7b) and present the L^1 errors in height at time $t = 1.5$ on a sequence of meshes in Table 2. For the sake of comparison, the results obtained with a standard Roe scheme are also presented. The table clearly shows that the EC, WB1 and WB2 schemes

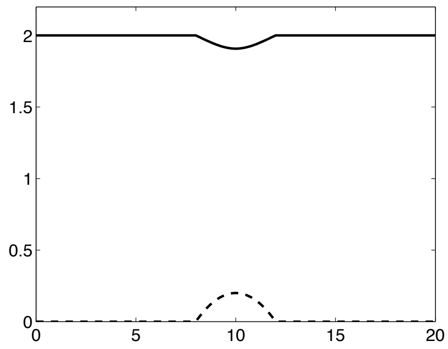


Figure 7. Initial surface level for the subsonic steady state.

are well-balanced and preserve the subsonic state to machine precision. The Roe scheme produces large errors (note that these errors are much larger than those obtained for the lake at rest in Table 1).

N	Roe	EC	WB1	WB2
50	1.42e-1	1.77e-14	1.71e-15	1.62e-15
100	7.65e-2	1.31e-14	5.32e-16	3.55e-16
200	4.07e-2	2.82e-14	3.77e-16	3.55e-17
400	2.10e-2	6.68e-14	4.88e-16	5.66e-16

Table 2. The L^1 error in height for the subsonic steady state with different schemes on a sequence of N mesh points at time $t = 1.5$.

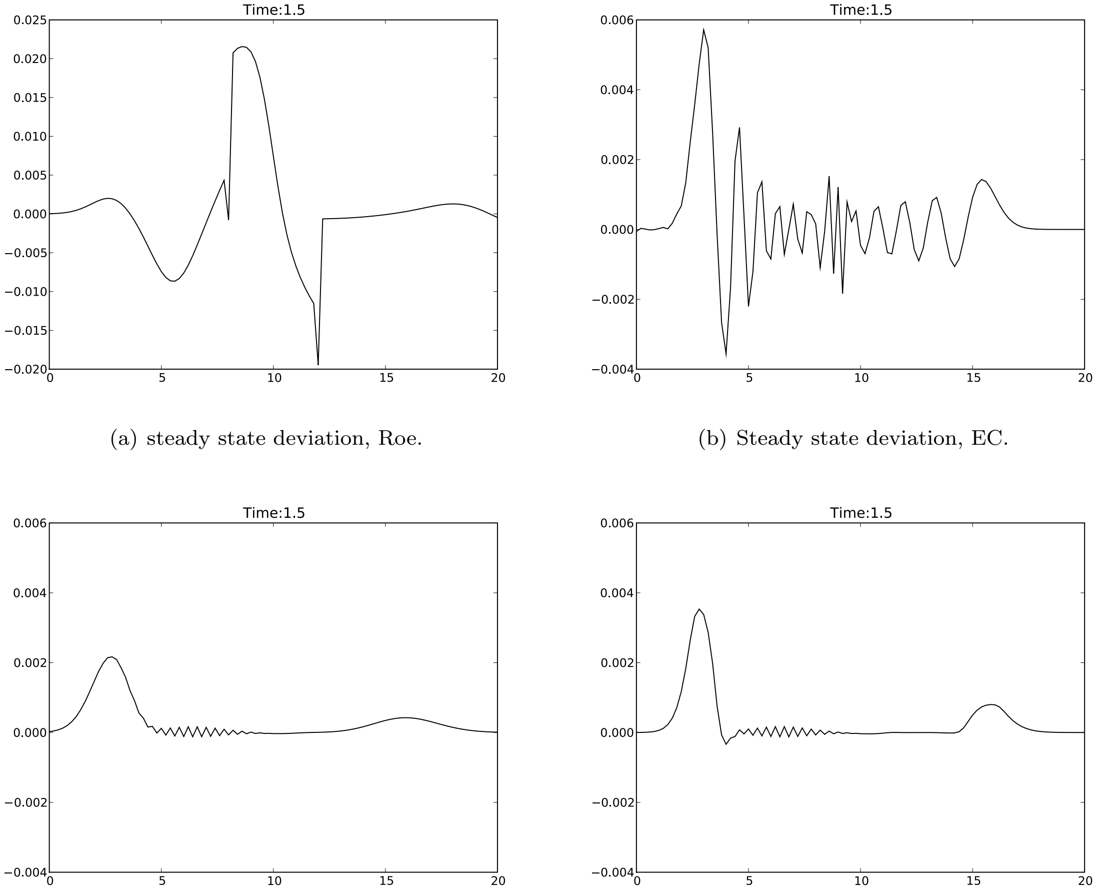
3.4.2. *Perturbed subsonic steady state.* As for the lake at rest, we will study the efficiency of the well-balanced schemes by perturbing the subsonic steady state. The initial conditions are the same as the previous experiment but with a perturbation of the height by a magnitude of $+0.01$ in the region $|x - 6| < 1/4$. The perturbation is similar to the one considered in (2.14). The solutions computed at time $t = 1.5$ with 100 mesh points are shown in Figure 8. For clarity, we present the deviations from the subsonic steady state. The exact solution breaks into two waves, one moving to the left and the other to the right. The Figure 8 shows that the standard Roe scheme fails to resolve the solution correctly and creates spurious waves as well as oscillations. Furthermore, these errors are an order of magnitude greater than the strength of the perturbation. This is to be expected as this scheme is not well-balanced.

The EC scheme captures the waves quite sharply but with unacceptably large post-shock oscillations. The oscillations are dampened considerably (but not entirely, with some very small residual oscillations) in the WB1 scheme, but the waves are smeared. The WB2 scheme increases the accuracy quite a bit and gives the best numerical results.

3.4.3. *Transonic steady state.* Next, we consider the same domain and bottom topography as in the previous experiment and the initial conditions

$$p_i \equiv \frac{3}{2}(mg)^{2/3} + \frac{g}{5}, \qquad M_{i+1/2} \equiv m \qquad \forall\ i,$$

with $m = 1.53$ and $g = 9.812$. The solution is a steady state that is part subsonic (on the left of the domain) and part supersonic (on the right) with a smooth transition in the middle of the domain (see Figure 9). This steady state is hence transsonic. We compute with the Roe, EC, WB1 and WB2 schemes up to time $t = 1.5$ and present L^1 errors in height in Table 3. As expected, the EC, WB1 and WB2 are all well-balanced and lead to very small errors, whereas the Roe scheme leads to unacceptably large errors.



(a) steady state deviation, Roe.

(b) Steady state deviation, EC.

(c) Steady state deviation, WB1.

(d) Steady state deviation, WB2.

FIGURE 8. Perturbed subsonic moving steady state at $t = 1.5$

N	Roe	EC	WB1	WB2
50	1.42e-1	3.29e-15	3.51e-15	3.02e-15
100	7.41e-2	3.63e-14	1.63e-14	9.17e-15
200	3.79e-2	2.92e-14	2.16e-14	1.60e-14
400	1.92e-2	3.32e-14	2.43e-14	9.00e-15

TABLE 3. The L^1 error in height for the transsonic steady state with different schemes on a sequence of N mesh points at time $t = 1.5$

3.4.4. *Perturbed transonic steady state.* We perturb the above transsonic steady state by adding +0.01 to height in the region $|x - 6| < 1/4$. All the other conditions are identical to the previous experiment. The results with the Roe, EC, WB1 and WB2 schemes are shown in Figure 10. We show the deviation from the transsonic steady state. We see that the Roe scheme produces spurious solutions. The EC scheme captures the small perturbations quite well, but with oscillations. The oscillations are reduced considerably with the first-order WB1 scheme but the waves are smeared. The high diffusion is demonstrated in the reduction of maximum wave height as compared to the EC scheme. Furthermore, there are small amplitude oscillations even with the WB1 scheme in this case. The WB2 scheme increase the sharpness and the wave height. Thus, the best numerical

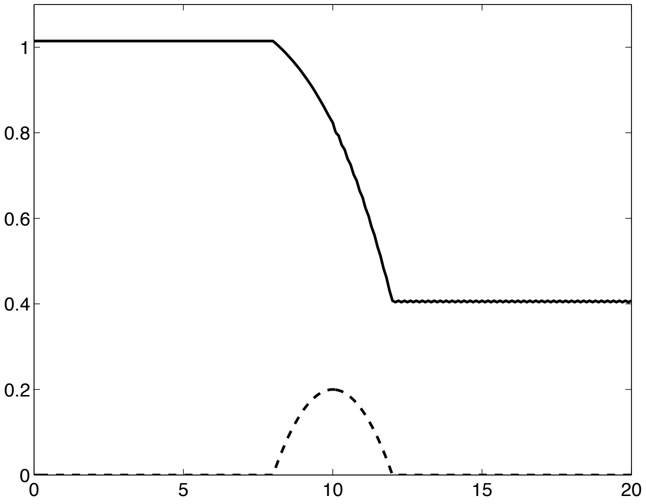


Figure 9. Initial surface level of the transsonic steady state.

results at this resolution are obtained with the WB2 scheme. The results are very similar to those obtained with the subsonic steady state.

Summing up, the results obtained by the WB1 and WB2 schemes are comparable to those shown in [26] after taking into account the fact that the schemes of [26] are higher than second-order accurate.

4. The two-dimensional problem

We consider the shallow water equations in two space dimensions given by (1.1). The energy preservation is given by the identity (1.6). The most interesting steady state in two space dimensions is the lake at rest given by (1.11). Our aim is to design numerical schemes that are energy preserving (energy stable) and preserve a discrete version of the lake at rest (1.11).

4.1. **Energy stable schemes.**

First, we extend the one-dimensional EC scheme (2.8) to two space dimensions. The extension is quite straightforward and follows the approach of [8]. The following notation is used:

$$\overline{a}_{i+1/2,j} = \frac{a_{i,j} + a_{i+1,j}}{2}, \qquad \overline{a}_{i,j+1/2} = \frac{a_{i,j} + a_{i,j+1}}{2},$$
$$[\![a]\!]_{i+1/2,j} = a_{i+1,j} - a_{i,j}, \qquad [\![a]\!]_{i,j+1/2} = a_{i,j+1} - a_{i,j}.$$

We define the following fluxes and sources, which are straightforward generalizations of their one-dimensional counterparts (2.7)

(4.1)

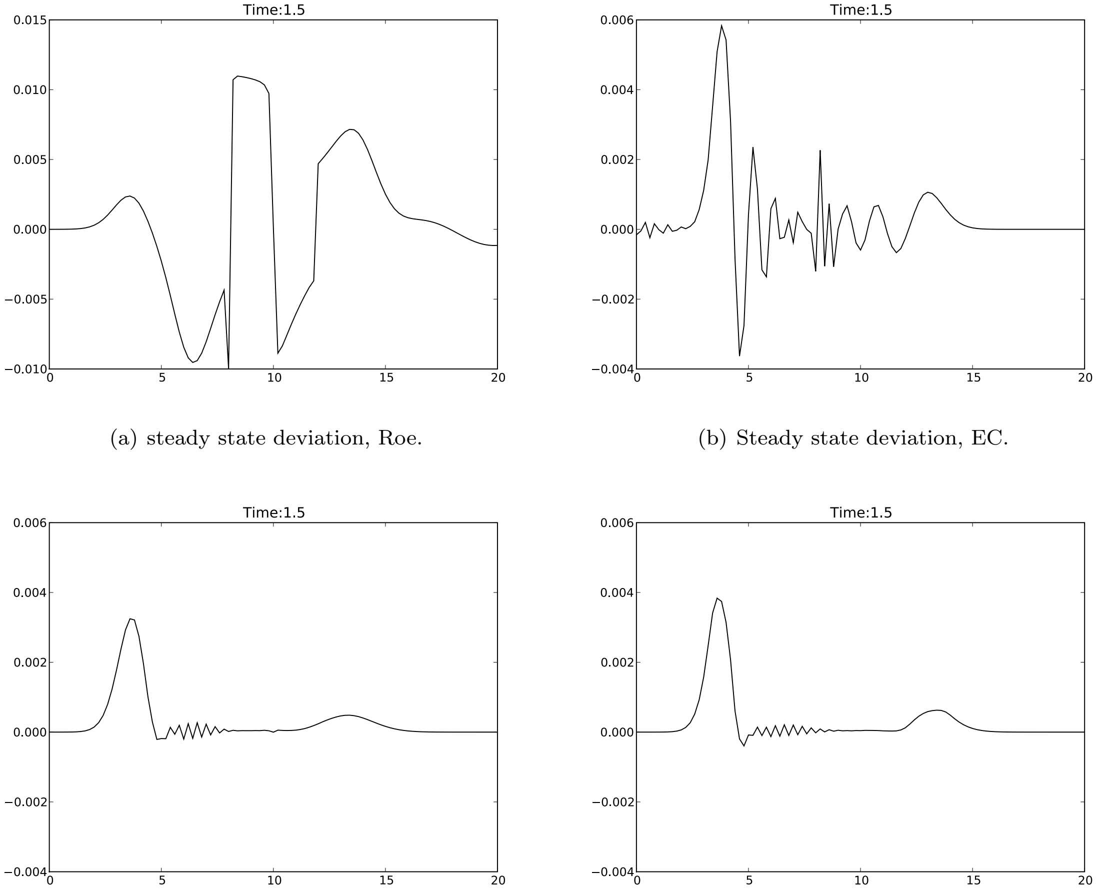
$$F^{\mathrm{EC}} i+1/2,j = \begin{bmatrix} \overline{h}_{i+1/2,j}\overline{u}_{i+1/2,j} \\ \overline{h}_{i+1/2,j}\left(\overline{u}_{i+1/2,j}\right)^2 + \frac{g}{2}(\overline{h^2})_{i+1/2,j} \\ \overline{h}_{i+1/2,j}\overline{u}_{i+1/2,j}\overline{v}_{i+1/2,j} \end{bmatrix}, \quad G^{\mathrm{EC}}_{i,j+1/2} = \begin{bmatrix} \overline{h}_{i,j+1/2}\overline{v}_{i,j+1/2} \\ \overline{h}_{i,j+1/2}\overline{u}_{i,j+1/2}\overline{v}_{i,j+1/2} \\ \overline{h}_{i,j+1/2}\left(\overline{v}_{i,j+1/2}\right)^2 + \frac{g}{2}(\overline{h^2})_{i,j+1/2} \end{bmatrix},$$

$$S^{\mathrm{EC}}_{i,j} = \begin{bmatrix} 0 \\ \frac{1}{2\Delta x}\left(\overline{h}_{i+1/2,j}[\![b]\!]_{i+1/2,j} + \overline{h}_{i-1/2,j}[\![b]\!]_{i-1/2,j}\right) \\ \frac{1}{2\Delta y}\left(\overline{h}_{i,j+1/2}[\![b]\!]_{i,j+1/2} + \overline{h}_{i,j-1/2}[\![b]\!]_{i,j-1/2}\right) \end{bmatrix},$$

The resulting two-dimensional scheme is then

$$\frac{d}{dt}U_{i,j} = -\frac{1}{\Delta x}\left(F^{\mathrm{EC}}_{i+1/2,j} - F^{\mathrm{EC}}_{i-1/2,j}\right) - \frac{1}{\Delta y}\left(G^{\mathrm{EC}}_{i,j+1/2} - G^{\mathrm{EC}}_{i,j-1/2}\right) - S^{\mathrm{EC}}_{i,j}. \tag{4.2}$$

We denote this scheme as the two-dimensional EC scheme. The properties of this scheme are summarized below.



(a) steady state deviation, Roe.

(b) Steady state deviation, EC.

(c) Steady state deviation, WB1.

(d) Steady state deviation, WB2.

FIGURE 10. Perturbed subsonic moving steady state at $t = 1.5$ using 100 mesh points.

Theorem 4.1. *The EC scheme (4.2) satisfies the following.*

(i) Accuracy: *It is a second-order accurate approximation of the two-dimensional shallow water equations* (1.1).

(ii) Energy conservation: *It is energy conservative, satisfying the discrete energy identity*

$$\frac{d}{dt}E_{i,j} + \frac{1}{\Delta x}\left(\widehat{H}_{i+1/2,j} - \widehat{H}_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(\widehat{K}_{i,j+1/2} - \widehat{K}_{i,j-1/2}\right) = 0,$$

where the numerical energy fluxes are

$$\widehat{H}_{i+1/2,j} = \langle \overline{V}_{i+1/2,j}, F_{i+1/2,j}\rangle - \overline{\Psi}_{i+1/2,j} - g\overline{h}_{i+1/2,j}[\![u]\!]_{i+1/2,j}[\![b]\!]_{i+1/2,j},$$
$$\widehat{K}_{i,j+1/2} = \langle \overline{V}_{i,j+1/2}, G_{i,j+1/2}\rangle - \overline{\Phi}_{i,j+1/2} - g\overline{h}_{i,j+1/2}[\![v]\!]_{i,j+1/2}[\![b]\!]_{i,j+1/2},$$

where

$$V_{i,j} = \begin{bmatrix} g(h_{i,j} + b_{i,j}) - \frac{u_{i,j}^2 + v_{i,j}^2}{2} \\ u_{i,j} \\ v_{i,j} \end{bmatrix}, \qquad \Psi_{i,j} = \frac{1}{2}gu_{i,j}h_{i,j}^2, \qquad \Phi_{i,j} = \frac{1}{2}gv_{i,j}h_{i,j}^2.$$

(iii) Well-balanced: *It preserves the discrete lake at rest steady state*

$$u_{i,j} \equiv 0, \qquad v_{i,j} \equiv 0, \qquad h_{i,j} + b_{i,j} \equiv \text{Constant}. \tag{4.3}$$

The proof of the above theorem is similar to the proof of Theorem 2.2 and we omit it here. The structure of the energy preserving fluxes and the source in (4.1) is essential in the proof.

As observed in the one dimensional case, the energy preserving EC scheme needs to be combined with suitable numerical diffusion operators to dampen oscillations and maintain energy stability. Furthermore, the energy stable scheme should preserve a discrete version of the lake at rest (4.3). We extend the numerical diffusion operator (2.16a) to two space dimensions to construct such a scheme. The extension follows the approach of [8] and involves the following matrices,

$$\begin{aligned} R^x_{i+1/2,j} &= \frac{1}{\sqrt{2g}} \begin{bmatrix} 1 & 0 & 1 \\ \overline{u}_{i+1/2,j} - \sqrt{g\overline{h}_{i+1/2,j}} & 0 & \overline{u}_{i+1/2,j} + \sqrt{g\overline{h}_{i+1/2,j}} \\ \overline{v}_{i+1/2,j} & \sqrt{g\overline{h}_{i+1/2,j}} & \overline{v}_{i+1/2,j} \end{bmatrix}, \\ R^y_{i,j+1/2} &= \frac{1}{\sqrt{2g}} \begin{bmatrix} 1 & 0 & 1 \\ \overline{u}_{i,j+1/2} & -\sqrt{g\overline{h}_{i,j+1/2}} & \overline{u}_{i,j+1/2} \\ \overline{v}_{i,j+1/2} - \sqrt{g\overline{h}_{i,j+1/2}} & 0 & \overline{v}_{i,j+1/2} + \sqrt{g\overline{h}_{i,j+1/2}} \end{bmatrix}, \end{aligned} \tag{4.4}$$

and

$$|\Lambda^x_{i+1/2,j}| = \text{diag}\left(|\overline{u}_{i+1/2,j} - \sqrt{g\overline{h}_{i+1/2,j}}|, |\overline{u}_{i+1/2,j}|, |\overline{u}_{i+1/2,j} + \sqrt{g\overline{h}_{i+1/2,j}}|\right),$$

$$|\Lambda^y_{i,j+1/2}| = \text{diag}\left(|\overline{v}_{i,j+1/2} - \sqrt{g\overline{h}_{i,j+1/2}}|, |\overline{v}_{i,j+1/2}|, |\overline{v}_{i,j+1/2} + \sqrt{g\overline{h}_{i,j+1/2}}|\right).$$

The numerical fluxes are given by

$$\begin{aligned} F^{\text{ES1}}_{i+1/2,j} &= F^{\text{EC}}_{i+1/2,j} - \frac{1}{2} R^x_{i+1/2,j} |\Lambda^x_{i+1/2,j}| (R^x_{i+1/2,j})^\top [\![V]\!]_{i+1/2,j} \\ G^{\text{ES1}}_{i,j+1/2} &= G^{\text{EC}}_{i,j+1/2} - \frac{1}{2} R^y_{i,j+1/2} |\Lambda^y_{i,j+1/2}| (R^y_{i,j+1/2})^\top [\![V]\!]_{i,j+1/2} \end{aligned} \tag{4.5}$$

where $F^{\text{EC}}_{i+1/2,j}$ and $G^{\text{EC}}_{i,j+1/2}$ are defined in (4.1). The resulting scheme is given by

$$\frac{d}{dt} U_{i,j} = -\frac{1}{\Delta x}\left(F^{\text{ES1}}_{i+1/2.j} - F^{\text{ES1}}_{i-1/2,j}\right) - \frac{1}{\Delta y}\left(G^{\text{ES1}}_{i,j+1/2} - G^{\text{ES1}}_{i,j-1/2}\right) - S^{\text{EC}}_{i,j}, \tag{4.6}$$

where $S^{\text{EC}}_{i,j}$ is the discretized source in (4.1). We refer to this scheme as the two-dimensional ES1 scheme. It is first-order accurate, consistent and energy stable. Furthermore, it preserves the discrete lake at rest (4.3). The proof of the above assertions follow in a similar way as the proof of Theorem 2.3, and so we omit the details. The well-balanced nature of the ES1 scheme (4.6) is a consequence of the fact that the diffusion is in terms of the energy variables which are constant in space for the lake at rest (4.3).

The two-dimensional ES1 scheme can be extended to second-order accuracy by using the approach of reconstructing in terms of the energy variables, as described in Section 2. This approach leads to a second-order accurate scheme that preserves the discrete lake at rest. We denote this second-order scheme as the two-dimensional ES2 scheme.

4.2. **Numerical experiments.**

4.2.1. *Two-dimensional lake at rest.* We consider the configuration used in [23, 25] among others and set the bottom topography to be

$$b(x,y) = 0.8 \exp\left(-5(x-0.9)^2 - 50(y-0.5)^2\right)$$

in the domain $(x,y) \in [0,2] \times [0,1]$. We use the lake at rest initial condition

$$h + b \equiv 1, \qquad u \equiv v \equiv 0.$$

The gravitational constant is set to $g = 9.812$. The configuration is shown in Figure 11. We compute with standard Roe, EC, ES1 and ES2 schemes on a sequence of meshes up to $t = 1$ and show the L^1 errors in height in Table 4. The table clearly shows that the EC, ES1 and ES2 preserve the steady states quite close to machine precision, whereas the standard Roe scheme produces large errors.

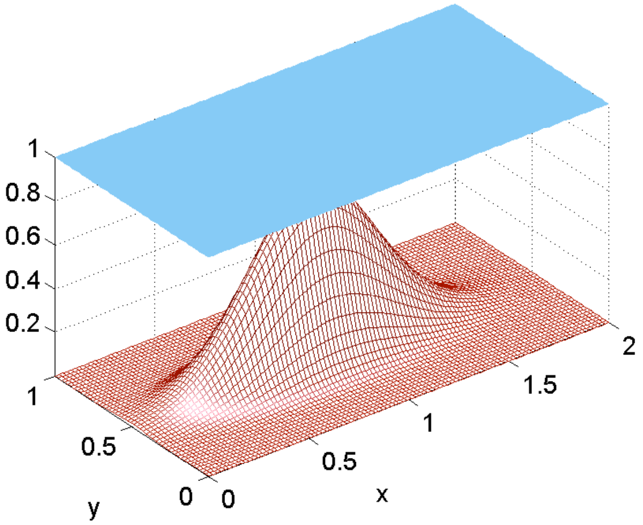


FIGURE 11. Water level and bottom topography for the two-dimensional lake at rest.

N	Roe	EC	WB1	WB2
50	1.71e-1	2.30e-15	2.95e-15	3.53e-15
100	8.73e-2	3.50e-14	3.48e-15	5.76e-15
200	5.81e-2	2.06e-11	3.95e-15	4.70e-15

TABLE 4. The L^1 error in height for the two-dimensional lake at rest with different schemes on a sequence of $2N \times N$ meshes at time $t = 1$

4.2.2. *Perturbed two-dimensional lake at rest.* Next, we consider a small perturbation to the above lake at rest by perturbing the height by $+0.01$ in the region $x \in [0.1, 0.2]$. The solutions computed by the ES1 scheme and ES2 scheme on a 600×300 mesh are shown in Figure 12. The solution exhibits complex features: It consists of both left- and right-going waves. As the right-going wave moves over the hump in the bottom, the middle part of the wave slows down and rises. The resulting wave patterns are quite intricate and consists of waves of different magnitudes. The left going wave hits the boundary at time $t = 0.03$ and we use Neumann type boundary conditions to ensure that the wave leaves the domain without numerical reflections. The figure shows that the first-order ES1 scheme captures the complex solution features qualitatively but smears them considerably. The second-order ES2 scheme is much more accurate and approximates the solution quite well. The results are comparable to those obtained in [23, 25] and other references therein.

5. CONCLUSIONS

The shallow water equations with bottom topography are considered in both one and two spatial dimensions. The smooth (weak) solutions of the equations are energy conservative (dissipating). Furthermore, the equations posses interesting steady states like the lake at rest (1.11) in both one and two space dimensions as well as general moving equilibrium states in one space dimension. Standard finite volume schemes for the shallow water equations are not energy conservative (energy stable), nor do they preserve discrete versions of interesting steady states. As a result, computations involving long time scales and perturbations of steady states are challenging.

We design a simple finite volume scheme termed the EC scheme (2.8) ((4.2) in two dimensions). This scheme is second-order accurate and conserves energy. Furthermore, it preserves discrete versions of the lake at rest in both one and two space dimensions. It also preserves a discrete version of the more general moving equilibrium state (3.4) in one space dimension. However, the scheme induces unphysical oscillations near shocks due to energy conservation. Shocks lead to energy dissipation in the continuous problem.

The EC scheme can be used as a basis to construct non-oscillatory energy stable schemes that preserve interesting steady states. Novel diffusion operators based on energy variables lead to energy stable schemes. Both the first- and second-order accurate versions of these schemes preserve the lake at rest (in both one and two space dimensions). Constructing a suitable numerical diffusion that preserves general moving equilibrium states (3.4) in one space dimension is trickier. We propose a diffusion operator based on equilibrium variables. Combined with the EC scheme, this diffusion operator leads to first- and second-order accurate schemes that preserve moving equilibrium states.

All the schemes designed in this paper are very simple to implement and computationally cheap. They require no special design features like hydrostatic reconstructions or solving nonlinear algebraic equations at each time step. They are natural extensions of the class of schemes proposed in [8] to the case of shallow water equations with topography. Numerical experiments demonstrating the robustness of the schemes in different configurations are presented and illustrate their computational efficiency. Given their simplicity of design and implementation, energy stability and low computational cost, the schemes of this paper appear to be attractive alternatives for computing flows involving the shallow water equations with realistic bottom topography.

We plan to extend the energy conservative and energy stable schemes to higher than second order of accuracy in a forthcoming paper. The approach of this paper will be extended to more complicated models like the multi-layer shallow water equations, the Euler equations for gas flows in nozzles and MHD equations for stratified magneto-atmospheres in the future.

References

[1] A. Arakawa. Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. *J. Comput. Phys.,* 1 (1), 1966, 119 - 143.

[2] A. Arakawa and V. R. Lamb. Computational design of the basic dynamical process of the UCLA general circulation model. *Meth. Comput. Phys.,* 17, 1977, 173-265.

[3] A. Arakawa and V. R. Lamb. A potential enstropy and energy conserving scheme for the shallow water equations. *Mont. Weat. Rev.,* 109, 1981, 18-36.

[4] E. Audusse, F. Bouchut, M. O. Bristeau, R. Klien and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM. Jl. Sci. Comp,* 25 (6), 2004, 2050 - 2065.

[5] M. Castro, J. M. Gallardo, C Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with non-conservative products. *Math. Comp.,* 75, 2006, 1103-1134.

[6] C. Dafermos. Hyperbolic conservation laws in continuum physics. Springer, Berlin, 2000.

[7] G. DalMaso, P. LeFloch and F. Murat. Definition and weak stability of nonconservative products. *J. Math. Pures. Appl.,* 74, 1995, 483-548.

[8] U. S. Fjordholm, S. Mishra and E. Tadmor. Energy preserving and energy stable schemes for the shallow water equations. *"Foundations of Computational Mathematics"*, Proc. FoCM held in Hong Kong 2008 (F. Cucker, A. Pinkus and M. Todd, eds), London Math. Soc. Lecture Notes Ser. 363, pp. 93-139, 2009.

[9] J. M. Greenberg and A. Y. LeRoux. A well-balanced scheme for numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.,* 33, 1996, 1-16.

[10] J. Goodman and P. D. Lax. On Dispersive Difference Schemes. I *Comm. Pure. Appl. Math.,* 41 (5), 1988, 591-613.

[11] S. Gottlieb, C. W. Shu and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM. Review,* 43, 2001, 89 - 112.

[12] N. Goutal and F. Maurel. Proceedings of the 2nd Workshop on Dam-Break Wave Simulation. *Technical Report HE-43/97/016/A, Electricit de France, Dpartement Laboratoire National d'Hydraulique, Groupe Hydraulique Fluviale.* 1997.

[13] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput. Phys., 49 (1983), pp. 357–393.

[14] A. Harten, B. Engquist, S. Osher and S. R. Chakravarty. Uniformly high order accurate essentially non-oscillatory schemes. *J. Comput. Phys.*, 1987, 231-303.

[15] S. Jin. A steady state capturing method for hyperbolic systems with geometrical source terms. *Math. Model. Numer. Anal.,* 35, 2001, 631-646.

[16] S. Jin and X. Wen. An efficient method for computing hyperbolic systems with geometrical source terms having concentrations. *J. Comput. Math.,* 22, 2004, 230-249.

[17] K. H. Karlsen, S. Mishra and N.H. Risebro. A new class of well-balanced schemes for conservation laws with source terms, *Math. Comp.,* 78 (265), 2009, 55-78.

[18] A. Kurganov and D. Levy. Central-upwind schemes for the St. Venant system. *Math. Model. Num. Anal.,* 36, 2002, 397-425.

[19] A. Kurganov and E. Tadmor. New high resolution central schemes for non-linear conservation laws and convection-diffusion equations. *J. Comput. Phys*, 160(1), 241-282, 2000.

[20] P. G. LeFloch, J. M. Mercier and C. Rohde. Fully discrete entropy conservative schemes of arbitrary order. *SIAM J. Numer. Anal.,* 40 (5), 2002, 1968-1992.

[21] P. G. LeFloch and C. Rohde. High order schemes, entropy inequalities and non-classical shocks. *SIAM. J. Numer. Anal.,* 37, 2000, 2023-2060.

[22] R. J. LeVeque. Finite volume methods for hyperbolic problems. *Cambridge university press,* Cambridge, 2002.

[23] R. J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm *J. Comput. Phys.*, 146, 346 - 365, 1998.

[24] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[25] S. Noelle, N. Pankratz, G. Puppo and J. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *J. Comput. Phys.*, 213, 474-499, 2006.

[26] S. Noelle, Y. Xing, C. Shu. High order well-balanced finite volume WENO schemes for shallow water equation with moving water. *J. Comput. Phys.*, 226 (1), 29-58, 2007.

[27] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes *J. Comput. Phys.*, 226, 250-258, 1981

[28] P. L. Roe. Entropy conservative schemes for Euler equations. *Talk at HYP 2006, Lyon, France.* Unpublished, Lecture available from http://math.univ-lyon1.fr/ hyp2006.

[29] G. Russo. Central schemes for conservation laws with application to shallow water equations. *S. Rionero, G. Romano (Eds.) STAMM 2002*, Springer Verlag, Italia, 2005, 225-246.

[30] C.-W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys., 77 (1988), pp. 439–471.

[31] C.-W. Shu and S. Osher, Efficient implementation of essentially nonoscillatory shock-capturing schemes, II, J. Comput. Phys., 83 (1989), pp. 32–78.

[32] C.-W. Shu, Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws, Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, A. Quarteroni, ed., Lecture Notes in Math.,Springer, Berlin., 1697 (1998).

[33] E. Tadmor. Numerical viscosity and entropy conditions for conservative difference schemes. *Math. Comp.*, 43 (168), 369 -381, 1984.

[34] E. Tadmor. The numerical viscosity of entropy stable schemes for systems of conservation laws, I. *Math. Comp.,* 49, 91-103, 1987.

[35] E. Tadmor. Approximate solutions of nonlinear conservation laws. *Advanced Numerical approximations of Nonlinear Hyperbolic equations, A. Quarteroi ed.*, Lecture notes in Mathematics, Springer Verlag (1998), 1-149.

[36] E. Tadmor. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Act. Numerica.,* 451-512, 2004.

[37] E. Tadmor and W. Zhong. Entropy stable approximations of Navier-Stokes equations with no artificial numerical viscosity. *J. Hyperbolic. Differ, Equ.,* 3 (3), 2006, 529-559.

[38] E. Tadmor and W. Zhong. Energy preserving and stable approximations for the two-dimensional shallow water equations. *In Mathematics and computation: A comtemporary view,* Proc. of the third Abel symposium, Alesund, Norway. Springer, 2008, 67-94.

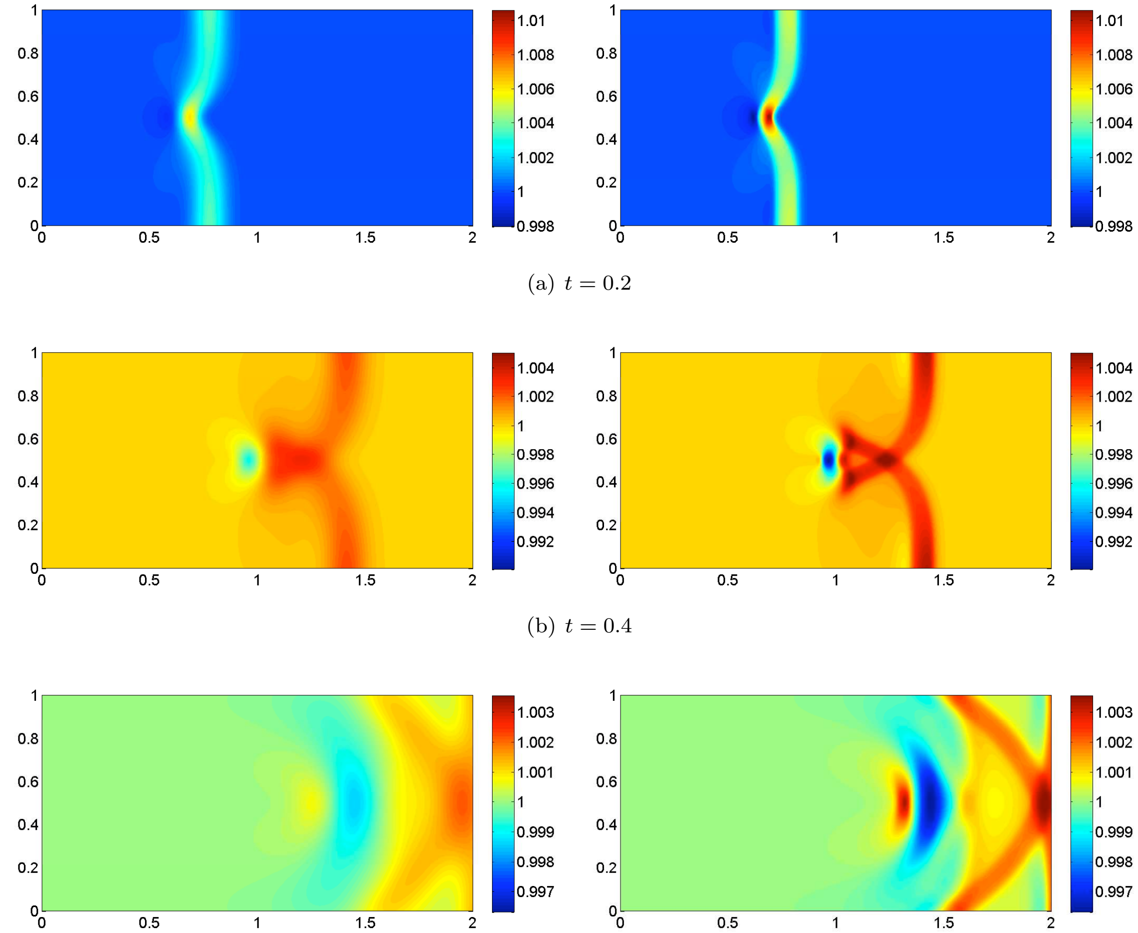
[39] M. E. Vazquez-Cendon. Improved treatment of source terms in upwind schmes for the shallow water equations in channels with irregular geometry. *J. Comput. Phys.,* 148, 1999, 497-526.

[40] G. B. Whitham. Linear and Nonlinear waves. *John Wiley and Sons.,* New York, 1999, 636 pp.

(Ulrik S.Fjordholm)
Seminar for Applied Mathematics (SAM)
Department of Mathematics, ETH Zürich,
HG J 48, Zürich -8092, Switzerland
E-mail address: ulriksf@gmail.com

(Siddhartha Mishra)
Seminar for Applied Mathematics (SAM)
Department of Mathematics, ETH Zürich,
HG G 57.2, Zürich -8092, Switzerland
E-mail address: smishra@sam.math.ethz.ch

(Eitan Tadmor)
Department of Mathematics
Center of Scientific Computation and Mathematical Modeling (CSCAMM)
Institute for Physical sciences and Technology (IPST)
University of Maryland
MD 20742-4015, USA
E-mail address: tadmor@cscamm.umd.edu



(a) $t = 0.2$

(b) $t = 0.4$

(c) $t = 0.6$

FIGURE 12. A simulation of the two-dimensional lake at rest with perturbation using the ES1 and ES2 scheme with 600×300 mesh points. Left column: ES1; right column: ES2.

Research Reports

No.	Authors/Title
10-26	*U.S. Fjordholm, S. Mishra and E. Tadmor* Well-balanced, energy stable schemes for the shallow water equations with varying topography
10-25	*U.S. Fjordholm and S. Mishra* Accurate numerical discretizations of non-conservative hyperbolic systems
10-24	*S. Mishra and Ch. Schwab* Sparse tensor multi-level Monte Carlo finite volume methods for hyperbolic conservation laws with random intitial data
10-23	*J. Li, J. Xie and J. Zou* An adaptive finite element method for distributed heat flux reconstruction
10-22	*D. Kressner* Bivariate matrix functions
10-21	*C. Jerez-Hanckes and J.-C. Nédélec* Variational forms for the inverses of integral logarithmic operators over an interval
10-20	*R. Andreev* Space-time wavelet FEM for parabolic equations
10-19	*V.H. Hoang and C. Schwab* Regularity and generalized polynomial chaos approximation of parametric and random 2nd order hyperbolic partial differential equations
10-18	*A. Barth, C. Schwab and N. Zollinger* Multi-Level Monte Carlo Finite Element method for elliptic PDE's with stochastic coefficients
10-17	*B. Kågström, L. Karlsson and D. Kressner* Computing codimensions and generic canonical forms for generalized matrix products
10-16	*D. Kressner and C. Tobler* Low-Rank tensor Krylov subspace methods for parametrized linear systems
10-15	*C.J. Gittelson* Representation of Gaussian fields in series with independent coefficients
10-14	*R. Hiptmair, J. Li and J. Zou* Convergence analysis of Finite Element Methods for $H(\mathrm{div};\Omega)$-elliptic interface problems