Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 226 (2007) 29–58
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

High-order well-balanced finite volume WENO schemes for shallow water equation with moving water

Sebastian Noelle [a,*,1], Yulong Xing [b], Chi-Wang Shu [c,2]

[a] *Institute for Geometry and Applied Mathematics, RWTH Aachen, 52056 Aachen, Germany*
[b] *Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, United States*
[c] *Division of Applied Mathematics, Brown University, Providence, RI 02912, United States*

Received 16 January 2007; received in revised form 16 March 2007; accepted 26 March 2007
Available online 11 April 2007

Abstract

A characteristic feature of hyperbolic systems of balance laws is the existence of non-trivial equilibrium solutions, where the effects of convective fluxes and source terms cancel each other. Recently a number of so-called well-balanced schemes were developed which satisfy a discrete analogue of this balance and are therefore able to maintain an equilibrium state. In most cases, applications treated equilibria *at rest*, where the flow velocity vanishes. Here we present a new very high-order accurate, exactly well-balanced finite volume scheme for *moving flow* equilibria. Numerical experiments show excellent resolution of unperturbed as well as slightly perturbed equilibria.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Shallow water equation; Moving water equilibria; High-order upwind finite volume scheme; Well-balanced scheme

1. Introduction

A challenge in the numerical analysis of hyperbolic systems of balance laws is to maintain the fundamental equilibria, and to compute their perturbations accurately. Indeed, if a scheme cannot balance the effects of convective fluxes and source terms, it may introduce spurious oscillations near equilibria. In order to reduce these the grid must be refined more than necessary. On the other hand, well-balanced schemes promise to be efficient near equilibria. In many cases they are also very accurate away from equilibria.

Many recent papers (see [1–4,9,11,14,17,18,20,21,27,30–33] and the references therein) treat the lake at rest equilibrium for the shallow water equations:

* Corresponding author. Tel.: +49 241 80 93953; fax: +49 241 80 92317.
E-mail addresses: noelle@igpm.rwth-aachen.de (S. Noelle), xing@cims.nyu.edu (Y. Xing), shu@dam.brown.edu (C.-W. Shu).
[1] Research supported by the European network HYKE under EC contract HPRN-CT-2002-00282.
[2] Research supported by ARO Grant W911NF-04-1-0291 and NSF Grant DMS-0510345.

0021-9991/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.03.031

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

where h denotes the water height, u is the velocity of the fluid, b represents the bottom topography and g is the gravitational constant.

The lake at rest is given by

$$u = 0 \quad \text{and} \quad H := h + b = \text{constant}. \tag{1.2}$$

The somewhat unusual feature of this state is that it can be expressed by linear relations in the conservative variables $U = (h,hu)$: if $h > 0$, then (1.2) is equivalent to $hu = 0$ and $h = H - b$. This makes it straightforward to transform the conservative variables into equilibrium variables $V = (hu, gH)$ and vice versa.

Contrary to that, the general moving water steady-state solutions are given by

$$hu = \text{constant} \quad \text{and} \quad \frac{1}{2}u^2 + g(h+b) = \text{constant}. \tag{1.3}$$

It is significantly more difficult to obtain well-balanced schemes for such moving water steady states. In [8], Gosse developed a class of first-order accurate flux-vector-splitting schemes based on the theory of non-conservative products [7] which is well-balanced for general steady states, including moving water equilibria. The interface method developed by Jin [11] captures general equilibria with second-order accuracy. In [12], Jin and Wen designed such a well-balanced scheme, which relies on computing an integral exactly, where the integrand is only implicitly given by solutions to a cubic equation. Even though the point values of this integrand can be obtained at any given point, the integral itself cannot be obtained in closed form and must be approximated by a numerical quadrature. The exact well-balancedness of the scheme would then be replaced by the numerical quadrature error. In [13], the same authors designed another scheme which is computationally less expensive, but the scheme can only maintain the moving water steady state to second-order accuracy, not exactly. Wen [29] developed steady state preserving schemes by reconstructing in equilibrium variables. Russo [22] developed well-balanced central schemes on staggered grids which are second-order accurate and exactly well-balanced for subcritical (i.e. subsonic) moving equilibria. We make an attempt in this paper to design exactly well-balanced, high-order accurate schemes for moving water steady states.

High-order exactly well-balanced finite volume schemes for still water equilibria have already been developed in [18,20,32,33]. Here we treat the much more complex situation of moving water.

The equilibrium variables for the moving steady-state water are given by

$$V = (m, E), \tag{1.4}$$

where

$$m = hu \quad \text{and} \quad E = \frac{1}{2}u^2 + g(h+b). \tag{1.5}$$

The nonlinearity makes it non-trivial to invert the map $U \to V$. Moreover, there is no unique way to recover an equilibrium function $V(x)$ or even a single equilibrium state $\overline{V}$ from a set of conservative cell averages $\{\overline{U}_i\}$. Our solution to this problem, introduced in Sections 2 and 3.2, is one of the key ingredients in this paper. The crucial idea is to define implicitly a *reference equilibrium state* $\overline{V}_i = \overline{V}_i(\overline{U}_i)$ in each cell in such a way, that all $\overline{V}_i$ coincide with $\overline{V}$ once we are in equilibrium.

Having defined the reference states $\overline{V}_i$, we introduce an *equilibrium limiter* which guarantees that a possible equilibrium present in the cell averages $\{\overline{U}_i\}$ is maintained in the reconstruction $U(x)$. This is the second key building block of our well-balanced scheme, and together with the definition of the reference equilibrium states it lays the foundation of our well-balanced algorithm.

From here on our procedure is somewhat more standard and extends techniques from [1,18,32] and others. The main work which remains to be done is to define a well-balanced quadrature rule of the source term, and to study the singular boundary layer at the cell edges. We split the edges into two infinitesimal layers, a convective layer where the source term is not active, and a topographic layer, where the source term is present but the flow remains in equilibrium. In the interior of the cell, we derive a new well-balanced quadrature rule for

the moving water case, which must be limited carefully in order to satisfy the conditions of the Lax–Wendroff theorem while maintaining high-order accuracy for general solutions in smooth regions. All together this leads to a rather transparent formulation of our well-balanced finite volume scheme, where the role of conservative fluxes and source terms can be clearly distinguished.

The outline of the paper is as follows: In Section 2 we study the basic transform between conservative and equilibrium variables. The main part of the paper is contained in Section 3. In Section 3.1 we lay out the framework of the discretization and formulate sufficient conditions for high-order accuracy and convergence to weak solutions. In Section 3.2 we define the reference states $\overline{V}_i$ implicitly and introduce our new equilibrium limiter. In Section 3.3 we introduce the basic well-balanced quadrature rule for the source term in moving water and introduce the infinitesimal layers at the edges which separate the discontinuities in the conservative variables and the source term. In Section 3.4 we summarize our new high-order accurate well-balanced finite volume scheme. The proof of well-balanced property and convergence to weak solutions are presented in Section 3.5. In Section 4 we present one-dimensional numerical results: several challenging moving water equilibria are preserved up to machine accuracy for many timesteps, and small perturbations are sharply resolved. For smooth non-equilibrium flows we obtain the expected high-order convergence rates. In Section 5 we present a two-dimensional numerical example which is a small perturbation of a one-dimensional moving equilibrium. The two-dimensional scheme is a dimension by dimension generalization of our one-dimensional well-balanced scheme. The results are compared with those obtained from the traditional high-order WENO schemes and the advantage of using the well-balanced scheme is demonstrated. Finally, in Section 6 we draw some conclusions.

We would like to point out that much of our approach can be carried over directly to other classes of balance laws. All one needs to rederive is the pointwise mapping between conservative and equilibrium variables introduced in Section 2 and the estimate at the sonic point in Lemma 3.11.

2. Conservative and equilibrium variables

In this section we study the sets of conservative variables U and equilibrium variables V upon which our well-balanced scheme relies. As usual, the conservative variables are denoted by $U = (h, m) = (h, hu)$. Let

$$E := \frac{1}{2}u^2 + g(h + b) \tag{2.1}$$

be the total energy. For smooth solutions, the shallow water equations may be rewritten as

$$h_t + m_x = 0, \tag{2.2}$$

$$u_t + E_x = 0. \tag{2.3}$$

Thus the steady states (1.3) are given by $m \equiv constant$, $E \equiv constant$. This motivates the introduction of the *equilibrium variables*

$$V := (m, E). \tag{2.4}$$

In order to construct our well-balanced scheme, it is essential to transform the conservative variables U into the equilibrium variables V and vice versa. Due to the nonlinearity of the energy, it is not straightforward to establish such a transform.

2.1. Variable transformations

Given conservative variables U and a bottom function b, the energy E (and hence the equilibrium variables $V = V(U)$) can be easily computed by (2.4). The difficulty lies in finding the inverse transform $U = U(V)$. For this, we introduce the Froude number:

$$Fr := |u|/\sqrt{gh}, \tag{2.5}$$

which plays the same role as the Mach number in gas dynamics: A state is called sonic, sub- or supersonic if the Froude number equals, falls below or exceeds unity. We label the different flow regimes by the sign function

$$\sigma := \operatorname{sign}(Fr - 1), \tag{2.6}$$

so

$$\sigma = \begin{cases} 1 & \text{supersonic flow,} \\ 0 & \text{sonic flow,} \\ -1 & \text{subsonic flow.} \end{cases} \tag{2.7}$$

Suppose now that $V = (m,E)$ and b are given. Under which conditions can we recover the conservative variable h from this information, and thus establish the desired transform $U = U(V)$?

The following development will be well-familiar to readers with a background in hydraulic engineering, see e.g. the classical textbook of Chow [6]. We denote the part of the energy depending on h by

$$\varphi(h) := \frac{m^2}{2h^2} + gh. \tag{2.8}$$

The quantity φ/g is called "specific energy" in hydraulic engineering. Here m is considered to be a fixed parameter. Our task is to find a unique solution h such that

$$\varphi(h) = E - gb. \tag{2.9}$$

If $m = 0$, then one can solve (2.9) as long as $E - gb > 0$. If $m \neq 0$, then $\varphi(h)$ is positive and convex. Its unique minimum is (h_0, φ_0) with

$$gh_0 = (g|m|)^{2/3}, \quad \varphi_0 = \frac{3}{2}(g|m|)^{2/3}. \tag{2.10}$$

Note that h_0 is exactly the sonic point for the prescribed value of m. We also have a lower bound for the energy, given by

$$E_0 = \varphi_0 + gb = \frac{3}{2}(g|m|)^{2/3} + gb. \tag{2.11}$$

If $E < E_0$, there is no solution to (2.9). If $E = E_0$, there is the unique solution $h = h_0$. If $E > E_0$, there are two solutions, one supersonic and the other one subsonic.

It is instructive to normalize the variables via $\hat{h} := h/h_0$, $\hat{\varphi} := \varphi/\varphi_0$. Then

$$\hat{\varphi}(\hat{h}) = \frac{2}{3}\left(\frac{1}{2\hat{h}^2} + \hat{h}\right), \tag{2.12}$$

and the Froude number may be written as

$$Fr(\hat{h}) = \hat{h}^{-3/2}. \tag{2.13}$$

This shows that $\hat{h} = 1$, $\hat{h} > 1$ resp. $\hat{h} < 1$ correspond to sonic, sub- and supersonic states, see Fig. 1. If we introduce $\hat{E} := (E - gb)/\varphi_0$, then (2.9) becomes

$$\hat{\varphi}(\hat{h}) = \hat{E}. \tag{2.14}$$

We summarize our results in the following Definition and Lemma.

Definition 2.1. Let $m \in \mathbb{R}$ be given. A pair $(\widehat{E}, \sigma) \in \mathbb{R} \times \{-1, 0, 1\}$ (resp. a triple $(E, b, \sigma) \in \mathbb{R}^2 \times \{-1, 0, 1\}$) is an admissible state if either

$$\sigma = 0 \quad \text{and} \quad \widehat{E} = 1 \quad (\text{resp. } E = E_0) \tag{2.15}$$

or

$$|\sigma| = 1 \quad \text{and} \quad \widehat{E} > 1 \quad (\text{resp. } E > E_0). \tag{2.16}$$

Lemma 2.2. *Let m be given, and suppose that the pair $(\widehat{E}, \sigma)$ is admissible. Then there exists a unique solution*

$$\hat{h} = \hat{h}(\widehat{E}, \sigma) \tag{2.17}$$

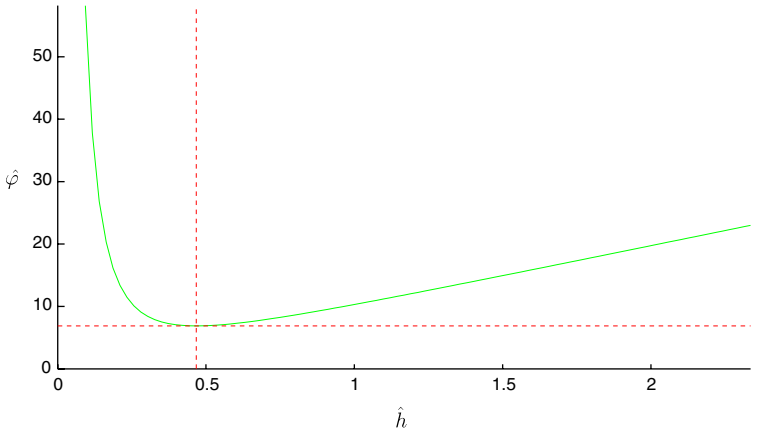

Fig. 1. The normalized function $\hat{\varphi}(\hat{h})$. Supersonic ($\hat{h} < 1$), sonic ($\hat{h} = 1$) and subsonic ($\hat{h} > 1$) regions.

such that

$$\begin{aligned} &\hat{h} < 1 \quad \text{for } \sigma = 1 \text{ (supersonic flow)} \\ &\hat{h} = 1 \quad \text{for } \sigma = 0 \text{ (sonic flow)} \\ &\hat{h} > 1 \quad \text{for } \sigma = -1 \text{ (subsonic flow).} \end{aligned} \tag{2.18}$$

We call $\widehat{h}(\hat{E}, \sigma)$ *the admissible solution of* (2.14).

Written in non-scaled variables (h, m, E, b) we have shown.

Corollary 2.3. *Let* m *be given, and suppose that the triple* (E, b, σ) *is admissible. Then the unique admissible solution* $h = h(m, E, b, \sigma)$ *of* (2.9) *is given by*

$$h(m, E, b, \sigma) = \frac{(g|m|)^{2/3}}{g} \hat{h}(\widehat{E}, \sigma). \tag{2.19}$$

Given admissible values $(\widehat{E}, \sigma)$ it is straightforward to find the corresponding solution $\hat{h}$ by Newton's method: if $\sigma = 0$, then $\hat{h} = 1$. If $\sigma = 1$, make sure that the starting value $\hat{h}^0$ in Newton's method satisfies $\hat{h}^0 < 1$ and $\hat{\varphi}(\hat{h}^0) > \widehat{E}$. Then the sequence $\hat{h}^n$ generated by Newton's method is monotone and converges quadratically towards $\hat{h}(\widehat{E}, \sigma)$. Analogously, if $\sigma = -1$, assure that $\hat{h}^0 > 1$ and $\hat{\varphi}(\hat{h}^0) > \widehat{E}$ in order to obtain monotone, quadratic convergence.

Note that a similar variable transform has also been used in [12,13,22].

3. High-order well-balanced finite volume scheme

In this section, we design a high-order finite volume weighted essentially non-oscillatory (WENO) scheme for the shallow water Eq. (1.1), with the objective to maintain the general moving steady state (1.3). We will concentrate on the one-dimensional case. Two space dimensions are treated in Section 5 with a numerical example. The basic framework of the well-balanced scheme follows the one introduced by Audusse et al. [1], and later used in the recent papers [18,33]. However, the approximation of the flux and source terms requires more attention due to the complexity of the moving steady state.

3.1. Framework of the discretization

For simplicity we write the shallow water equations in the form

$$U_t + f(U)_x = s(U, b). \tag{3.1}$$

We discretize the computational domain with cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], i = 1, \ldots, N$. We denote the size of the ith cell by $\triangle x_i$ and the center of the cell by $x_i = \frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})$. The computational variables are $\overline{U}_i(t)$, which approximate the cell averages $\overline{U}(x_i, t) = \frac{1}{\triangle x_i}\int_{I_i} U(x,t)\mathrm{d}x$.

We solve an integrated version of (3.1) over the interval I_i. Our conservative finite volume scheme then takes the classical semidiscrete form

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{U}_i(t) = -\frac{1}{\Delta x_i}\left(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}\right) + \frac{1}{\triangle x_i}s_i =: \frac{1}{\Delta x_i}r_i. \tag{3.2}$$

where $\hat{f}_{i+\frac{1}{2}}$ is a consistent, Lipschitz continuous numerical flux for the homogeneous shallow water equations and s_i is a high-order approximation to the integral of the source term $\int_{I_i} s(h(x,t), b(x))\mathrm{d}x$. For later reference, we call the RHS of (3.2) the *residual* $r_i/\Delta x_i$. Thus a well-balanced scheme is one for which all residuals vanish at steady state.

As to the formal accuracy of the scheme, we have the following lemma:

Lemma 3.1. *The numerical scheme (3.2) is formally kth order accurate if the following holds in smooth regions*:

(i) $\hat{f}_{i+\frac{1}{2}} = f(U(x_{i+\frac{1}{2}}, t)) + \mathcal{O}((\triangle x_i)^k)$ *with a smooth error term* $\mathcal{O}((\triangle x_i)^k)$
(ii) $s_i = \int_{I_i} s(h,b)\mathrm{d}x + \mathcal{O}((\Delta x_i)^{k+1})$

The proof of this lemma is straightforward.

We choose a TVD Runge–Kutta discretization [26] in time. In order to complete the definition of the scheme, we need to introduce the spatial reconstruction, the source term discretization, and the numerical fluxes. These will be described in Sections 3.2 and 3.3. In Section 3.5 we will also prove that our scheme satisfies a Lax–Wendroff theorem, which assures that limits are weak solutions.

3.2. Equilibrium-limited reconstructions in the cell interior

Assume the initial values $\overline{U}_i$ and $\bar{b}_i$ are given. We apply the high-order accurate WENO reconstruction procedure [25,23] on $\bar{b}_i$ to obtain $b_i, b^{\pm}_{i+\frac{1}{2}}$, and the approximations of $b(x)$ at the relevant Gaussian points. If $b(x)$ is known at all points, this WENO reconstruction procedure is unnecessary. The WENO reconstruction procedure is based on a nonlinear, convex combination of lower order reconstructions from sub-stencils, with the combination coefficients depending on the local smoothness of the function in relevant cells. It can achieve uniformly high-order accuracy in smooth regions and can maintain a sharp, non-oscillatory discontinuity transition. We refer to [16,10,25,23] for more details.

At each time step t^n, we first apply the WENO reconstruction procedure to the variables $\overline{U}_i$ to obtain $U^{\pm}_{i+\frac{1}{2}}$, $\sigma^{\pm}_{i+\frac{1}{2}}$, and hence $V^{\pm}_{i+\frac{1}{2}}$. The reconstructed values U_i, σ_i and V_i at the center of the cell are also needed for the purpose of source term discretization.

Now we need to address one of the more subtle points of the well-balanced algorithm. Even if the initial data are in perfect equilibrium, say $V(x) \equiv \overline{V}$ for some constant equilibrium state $\overline{V}$, the WENO-reconstructed values $U_i, U^{\pm}_{i+\frac{1}{2}}$ and hence $V_i, V^{\pm}_{i+\frac{1}{2}}$ may not be in equilibrium any more. The problem comes from the total energy $E = \frac{1}{2}u^2 + g(h+b)$. First of all, the topography b may be a general function of x. Second, the velocity depends nonlinearly on height and momentum. For the lake at rest, the second problem disappears since $u = 0$. The first problem can be fixed by reconstructing not b, but $h + b$ and recovering $b_i, b^{\pm}_{i+\frac{1}{2}}$ as $(h+b)_i - h_i, (h+b)^{\pm}_{i+\frac{1}{2}} - h^{\pm}_{i+\frac{1}{2}}$, see [1,18].

For moving equilibria, this is much less straightforward. We recall our assumption that the topography $b(x)$ is smooth within each cell I_i. Let the conservative cell averages $\overline{U}_i = (\bar{h}_i, \bar{m}_i)$ be given. Now we assume that the cell is in equilibrium, with constant values (m,E). Certainly, $m = \bar{m}_i$. From (2.11), the minimal possible value of E at x is $E_0(x) = \frac{3}{2}(g|m|)^{2/3} + gb(x)$. Therefore,

$$E \geqslant \frac{3}{2}(g|\bar{m}_i|)^{2/3} + g\max_{x\in I_i} b(x) =: E_i^{\min}. \tag{3.3}$$

It is convenient to define two more average heights related to cell I_i, one the maximal supersonic average height h_i^+ and the other the minimal subsonic average height h_i^- via

$$h_i^+ := \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E_i^{\min}, b(x), 1)\mathrm{d}x \tag{3.4}$$

$$h_i^- := \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E_i^{\min}, b(x), -1)\mathrm{d}x. \tag{3.5}$$

Note that $h_i^+ < h_0 < h_i^-$, where h_0 is the critical, or sonic, point. Now we define a local reference energy $\overline{E}_i$ as follows:

Definition 3.2

(i) If $\bar{h}_i < h_i^+$, then define $E = \bar{E}_i$ to be the unique solution of

$$\bar{h}_i = \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E, b(x), 1)\mathrm{d}x. \tag{3.6}$$

(ii) If $\bar{h}_i > h_i^-$, then define $E = \overline{E}_i$ to be the unique solution of

$$\bar{h}_i = \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E, b(x), -1)\mathrm{d}x. \tag{3.7}$$

(iii) Otherwise, if $h_i^+ \leqslant \bar{h}_i \leqslant h_i-$, then set

$$\overline{E}_i := E_i^{\min}. \tag{3.8}$$

(iv) We call $\overline{V}_i := (\bar{m}_i, \overline{E}_i)$ the reference equilibrium values.

Remark 3.3

(1) Note that in the super- (resp. sub-)sonic case the function $h(\bar{m}_i, E, b(x), \pm 1)$ is strictly monotonely decreasing (increasing) with respect to E. This makes the solutions E of (3.6) resp. (3.7) well defined.
(2) Due to (3.4),(3.5), the reference energy $\overline{E}_i$ is continuous as a function of $\bar{h}_i$.
(3) In case (iii), the equilibrium solution becomes sonic for some point $x^* \in I_i$. The sonic point x^* is a maximum of b over cell I_i.

Lemma 3.4. *Suppose that $b(x)$ and $U(x)$ are piecewise smooth. Assume furthermore that for each cell I_i, there exists a constant equilibrium value $V_i^* = (m_i^*, E_i^*)$ such that*

$$V(U(x), b(x)) = V_i^* \quad \text{for } x \in I_i \tag{3.9}$$

(in other words, the data are locally in equilibrium). Let $\overline{U}_i$ be the local cell averages and let $\overline{V}_i$ be the reference equilibrium values given by Definition 3.2. Then

$$\overline{V}_i = V_i^* \quad \text{for all } i. \tag{3.10}$$

Proof. First suppose that the cell is entirely supersonic, $\sigma(x) \equiv 1$. Therefore, $h(\bar{m}_i, E, b(x), \sigma(x))$ is strictly decreasing as a function of E. Since $E_i^* \geqslant E_i^{\min}$,

$$h_i^+ = \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E_i^{\min}, b(x), 1)\mathrm{d}x \geqslant \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E_i^*, b(x), 1)\mathrm{d}x = \bar{h}_i. \tag{3.11}$$

We are therefore in case (i) of Definition 3.2, and $\overline{E}_i$ satisfies

$$\bar{h}_i = \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, \overline{E}_i, b(x), 1)\mathrm{d}x = \frac{1}{\Delta x_i}\int_{I_i} h(\bar{m}_i, E_i^*, b(x), 1)\mathrm{d}x. \tag{3.12}$$

By the monotonicity of $h(E)$, this implies that $\overline{E}_i = E_i^*$. Cells which are entirely subsonic can be treated analogously.

Now suppose that the cell is transsonic, i.e. σ changes sign within the cell at some point x^*. Then

$$E(x^*) = \frac{3}{2}(g|\bar{m}_i|)^{2/3} + gb(x^*) \leqslant E_i^{\min}. \tag{3.13}$$

On the other hand, $E = E_i^*$ is constant over the cell, and $E_i^* \geqslant E_i^{\min}$. This implies that $E_i^* = E_i^{\min}$.

It remains to show that $\overline{E}_i = E_i^{\min}$. Split the cell into $I_i^{\pm} := \{x \in I_i : \sigma(x) = \pm 1\}$ and $I_i^0 := \{x \in I_i : \sigma(x) = 0\}$. Then

$$\begin{aligned}\bar{h}_i &= \frac{1}{\triangle x_i}\left(\int_{I_i^+} h(\bar{m}_i, \overline{E}_i, b(x), 1)\mathrm{d}x + \int_{I_i^-} h(\bar{m}_i, \overline{E}_i, b(x), -1)\mathrm{d}x + \int_{I_i^0} h(\bar{m}_i, \overline{E}_i, b(x), 0)\mathrm{d}x\right)\\ &\geqslant \frac{1}{\triangle x_i}\left(\int_{I_i^+} h(\bar{m}_i, \overline{E}_i, b(x), 1)\mathrm{d}x + \int_{I_i^-} h(\bar{m}_i, \overline{E}_i, b(x), 1)\mathrm{d}x + \int_{I_i^0} h(\bar{m}_i, \overline{E}_i, b(x), 1)\mathrm{d}x\right) = h_i^+.\end{aligned} \tag{3.14}$$

Similarly, we obtain $\bar{h}_i \leqslant h_i^-$. Thus we are in case (iii) of Definition 3.2, and $\overline{E}_i = E_i^{\min} = E_i^*$. □

In actual implementation, we use a Gauss quadrature of sufficient accuracy to approximate the integral in (3.6)–(3.8). That is, the reference energy $\overline{E}_i$ is implicitly defined by the equation

$$\bar{h}_i = \frac{1}{\Delta x_i}\sum_{\alpha} \omega_\alpha h(\overline{hu}_i, \overline{E}_i, b_{i+\alpha}, \sigma(\overline{U}_i)). \tag{3.15}$$

A Newton iteration is then used to solve (3.15) with the initial guess of $\overline{E}_i$ being

$$\overline{E}_i^{(0)} := \frac{\overline{hu}_i^2}{2\bar{h}_i^2} + g(\bar{h}_i + \bar{b}_i).$$

The conclusion of Lemma 3.4 still holds for the reference value $\overline{E}_i$ defined in (3.15), if the given conservative cell average $\overline{U}_i$ is computed following the same quadrature. The relevance of Lemma 3.4 (and its discrete analogue) is that it provides an indicator that we have reached equilibrium, since in this case all the values $\overline{V}_i$ coincide.

Next we show how to use the local reference values $\overline{V}_i$ to modify the WENO-reconstructed values $V_{i+\frac{1}{2}}^{\pm}$ and V_i in such a way that they maintain any present global equilibrium state $\overline{V}$. For this we use the total variation bounded (TVB) [24] type limiter function:

$$\lim(w; \bar{w}_i, \bar{w}_{i\pm1}) := \bar{w}_i + \lambda(w - \bar{w}_i), \tag{3.16}$$

where

$$\lambda := \min\left(1, \frac{\sum_{j=i\pm1}|\bar{w}_j - \bar{w}_i|^2}{2|w - \bar{w}_i|^2}\right). \tag{3.17}$$

Of course, other limiters should be possible as well.

We apply the limiter separately to momentum m and energy E, and write the result symbolically as

$$\widetilde{V}_{i+\frac{1}{2}}^{\pm} = \lim\left(V_{i+\frac{1}{2}}^{\pm}; \overline{V}_i, \overline{V}_{i\pm1}\right). \tag{3.18}$$

Similarly, we compute the limited pointwise values $\tilde{V}_i$. Note that non-negative energies $E_{i+\frac{1}{2}}^{\pm}$ will remain non-negative. We have the following well-balanced property, which is important for the following steps:

Lemma 3.5. *At steady state, where $V(x) \equiv \overline{V}$, the limited values (3.18) satisfy*

$$\widetilde{V}_{i+\frac{1}{2}}^{\pm} = \widetilde{V}_i = \overline{V}_i = \overline{V} \quad \text{for all } i. \tag{3.19}$$

Therefore, we call (3.16)–(3.18) the equilibrium limiter.

Proof. If $V(x) \equiv \overline{V}$, then $\overline{V}_i = \overline{V}$ for all i due to (3.10). Therefore, the parameter λ in (3.16) and (3.17) vanishes, and

$$\lim(V^{\pm}_{i+\frac{1}{2}}; \overline{V}_i, \overline{V}_{i\pm1}) = \overline{V}_i = \overline{V}. \qquad \square \tag{3.20}$$

Remark 3.6. The limiter is inactive in smooth region if the solution is far from the steady state, even near smooth extrema, as can be verified by simple Taylor expansion. This guarantees that the limiter does not affect the high order accuracy of the scheme in smooth region for general solutions of (1.1).

The corresponding conservative variables are given by

$$\widetilde{U}^{\pm}_{i+\alpha} := U(\widetilde{V}^{\pm}_{i+\alpha}, b^{\pm}_{i+\alpha}, \sigma^{\pm}_{i+\alpha}) \quad \text{for } \alpha \in \left\{0, \frac{1}{2}\right\}. \tag{3.21}$$

As an immediate consequence of Lemma 3.5 we have

Corollary 3.7. *If $\overline{V}_{i\pm1} = \overline{V}_i$, then the equilibrium-limited values (3.21) satisfy*

$$V(\widetilde{U}_i, b_i) = V\left(\widetilde{U}^{\pm}_{i+\frac{1}{2}}, b^{\pm}_{i+\frac{1}{2}}\right) = \overline{V}_i. \tag{3.22}$$

3.3. Well-balanced quadrature rules for the residuum

In the previous section we have introduced the subtleties of the reconstruction in the interior of the cells, where the solution is smooth. In this section we will resolve the cell-boundary discontinuities in b and U by two layers, an equilibrium and a convective layer. In each of these layers as well as in the interior of the cell we will define the numerical residuum in a suitable way. This will result in a well-balanced residuum.

At the boundary, both the conservative variables $\widetilde{U}^{\pm}_{i+\frac{1}{2}}$ and the topography $b^{\pm}_{i+\frac{1}{2}}$ exhibit a jump discontinuity. As usual, the jump in the conservative variables is treated by an approximate Riemann solver. The jump in the topography will give rise to a δ-singularity in the source term, which has to be taken into account.

To derive our scheme, we separate the boundary into two layers, see Fig. 2. Take, for example the right boundary of cell i. To illustrate our approach, we introduce points $x_A < x_B < x_C := x_{i+\frac{1}{2}}$ which are separated by an infinitesimal distance. Together with these we introduce the values

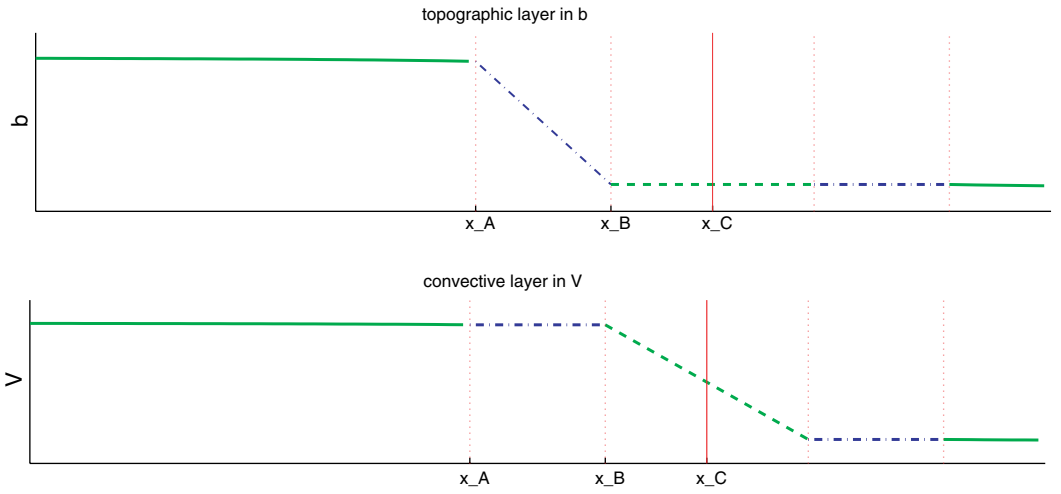

Fig. 2. The boundary layer model. Boundary between cells i and $i+1$. Top: discontinuous topography b (dash-dot). Bottom: shock-discontinuity in V (dashed).

$$(U_A, b_A) := \left(\widetilde{U}^-_{i+\frac{1}{2}}, b^-_{i+\frac{1}{2}}\right), \tag{3.23}$$

$$(U_B, b_B) := \left(\widetilde{U}^-_{i+\frac{1}{2}}, \hat{b}_{i+\frac{1}{2}}\right), \tag{3.24}$$

$$(U_C, b_C) := \left(\widetilde{U}_{i+\frac{1}{2}}, \hat{b}_{i+\frac{1}{2}}\right). \tag{3.25}$$

The values at point x_A are adjacent to the interior of the cell. The value $b^-_{i+\frac{1}{2}}$ is the WENO-reconstructed bottom topography, and $\widetilde{U}^-_{i+\frac{1}{2}}$ is the WENO-reconstructed and equilibrium limited conservative variable (3.21). At the point x_B the topography from the right of cell i and the left of cell $i+1$ is merged,

$$\hat{b}_{i+\frac{1}{2}} = \min\left(b^-_{i+\frac{1}{2}}, b^+_{i+\frac{1}{2}}\right). \tag{3.26}$$

The equilibrium variable remains constant, $V_B = V_A$, and the conservative variable changes accordingly to the new value

$$\widetilde{U}^-_{i+\frac{1}{2}} := U\left(\widetilde{V}^-_{i+\frac{1}{2}}, \hat{b}_{i+\frac{1}{2}}, \sigma^-_{i+\frac{1}{2}}\right). \tag{3.27}$$

Between the points x_B and x_C the topography remains unchanged. The point x_C marks the interface between cells i and $i+1$. The interface value $\widetilde{U}_{i+\frac{1}{2}}$ symbolizes the solution of the approximate Riemann problem,

$$f\left(\widetilde{U}_{i+\frac{1}{2}}\right) = \hat{f}_{i+\frac{1}{2}} = F\left(\widetilde{U}^-_{i+\frac{1}{2}}, \widetilde{U}^+_{i+\frac{1}{2}}\right). \tag{3.28}$$

We can therefore distinguish two boundary layers within each cell. We call $[x_B, x_C]$ the convective and $[x_A, x_B]$ the topographic, or equilibrium layer. When we consider the complete residual over cell i, we also introduce the points $x_{C'} = x_{i-\frac{1}{2}} < x_{B'} < x_{A'}$ as well as values $U_{C'}$, $U_{B'}$, $U_{A'}$.

We would like to remark that Castro, Gallardo, Parés and co-workers [4,20] have developed high-order well-balanced schemes based on the theory of non-conservative products [7]. It would be interesting to understand if the paths by which they connect values across an interface can be related to our subcell construction.

3.3.1. The residuum in the convective layer

In the convective layer, the topography is constant. Therefore, the source term vanishes. Adding contributions from the left and right boundaries of cell i we obtain

$$s_i^{\text{conv}} = 0 \tag{3.29}$$

$$r_i^{\text{conv}} = -f(U_C) + f(U_B) - f(U_{B'}) + f(U_{C'}) \tag{3.30}$$

$$= -f\left(\widetilde{U}_{i+\frac{1}{2}}\right) + f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) - f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right) + f\left(\widetilde{U}_{i-\frac{1}{2}}\right). \tag{3.31}$$

3.3.2. The residuum in the equilibrium layer

In the equilibrium layer the bottom b changes while the equilibrium variables $V = (m, E)$ remain constant. Let

$$R(x) := r(x)/\Delta x \quad \text{and} \quad S(x) := s(x)/\Delta x.$$

Since for the exact residuum

$$R(x) = -f(U)_x + S(x) = -f(U)_x - gh(x)b(x)_x = -um_x - hE_x = 0$$

the discrete residuum should vanish in the equilibrium layer. Therefore, we define

$$r_i^{\text{equi}} = 0, \tag{3.32}$$

and accordingly

$$s_i^{\text{equi}} = f(U_B) - f(U_A) + f(U_{A'}) - f(U_{B'}) = f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) - f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) + f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right) - f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right). \tag{3.33}$$

Thus we can express the source term as the convective flux difference and vice versa.

3.3.3. The interior residual

Let us denote the interior of the cell by $[x_L, x_R]$, with boundary values U_L, U_R, b_L, b_R. We assume that all jumps discontinuities of the topography are located at the cell boundaries, and that the reconstructions of the topography (denoted again by $b(x)$) are globally and uniformly Lipschitz continuous in the interior of the cells:

Assumption 3.8

(i) The exact bottom $b(x)$ is piecewise smooth with at most finitely many discontinuities.

(ii) The reconstructions $b_\Delta(x)$ are uniformly and globally Lipschitz continuous in the interior of the cells: There is a constant $L_b > 0$ such that for all $\Delta x > 0$, for all cells I_i and for all interior points $x,y \in I_i$,

$$|b_\Delta(x) - b_\Delta(y)| \leqslant L_b|x - y|. \tag{3.34}$$

(iii) There is a constant $C_B > 0$ such that for all Δx and for all edges $x_{i-\frac{1}{2}}$ where $b(x)$ is smooth,

$$|b^+_\Delta(x_{i-\frac{1}{2}}) - b^-_\Delta(x_{i-\frac{1}{2}})| \leqslant C_b\Delta x^2. \tag{3.35}$$

Given numbers a_L and a_R, let

$$Da := a_R - a_L, \quad \bar{a} := (a_L + a_R)/2 \tag{3.36}$$

be the difference and mean operators. For later use, we recall the product rule of differencing

$$D(ab) = \bar{a}Db + Da\bar{b}. \tag{3.37}$$

We would like to define a residuum r^R_L which is a high-order accurate discretization of the exact cell residuum

$$-\int_{I^R_L}(f(U)_x + ghb_x)\mathrm{d}x = -\int_{I^R_L}(um_x + hE_x)\mathrm{d}x \tag{3.38}$$

and vanishes for smooth equilibria, where m and E are constants. A standard discretization, which is well-balanced for the lake at rest ($u = 0 = D(h + b)$) is

$$-Df - g\bar{h}Db.$$

Now we will refine this discretization in such a way that it also balances moving equilibria. For this we augment the standard source term quadrature $-g\bar{h}Db$ by a term $\hat{s}^{\mathrm{int}}_i$,

$$r^{\mathrm{int}}_i = -Df - g\bar{h}Db + \hat{s}^{\mathrm{int}}_i. \tag{3.39}$$

In order to understand what well-balancing of moving equilibria requires we assume that $Dm = DE = 0$ and expand the flux difference in that case: using the product rule (3.37) we obtain

$$\begin{aligned} Df &= D(mu + gh^2/2) = \bar{m}Du + \bar{u}Dm + g\bar{h}Dh = \bar{m}Du + \bar{u}Dm + \bar{h}D(E - gb - u^2/2) \\ &= -g\bar{h}Db + (\bar{m} - \bar{h}\bar{u})Du = -g\bar{h}Db + \frac{1}{4}Dh(Du)^2. \end{aligned} \tag{3.40}$$

Thus the residuum r^{int}_i in (3.39) vanishes if and only if

$$\hat{s}^{\mathrm{int}}_i = Df + g\bar{h}Db = \frac{1}{4}Dh(Du)^2 \quad \text{for } Dm = DE = 0. \tag{3.41}$$

In addition to this, the correction should be small enough to admit convergence to weak solutions, namely

$$\hat{s}^{\mathrm{int}}_i = \mathrm{o}(\Delta x), \tag{3.42}$$

see the proof of Theorem 3.17. We also require that it is antisymmetric in the sense that

$$\hat{s}_i^{\text{int}}(Db, Dh, Du) = -\hat{s}_i^{\text{int}}(-Db, -Dh, -Du). \tag{3.43}$$

To summarize, we define the quadrature for the source term in the interior of the ith cell by

$$s_i^{\text{int}} = -g\bar{h}Db + \hat{s}_i^{\text{int}} \tag{3.44}$$

and require that $\hat{s}_i^{\text{int}}$ satisfies conditions (3.41)–(3.43). By construction, we have the following well-balancing result:

Lemma 3.9. *For balanced states, i.e. if $Dm = DE = 0$, then $r_i^{int} = 0$.*

3.3.4. Constructing the correction to the interior source term $\hat{s}_i^{\text{int}}$

The construction of the correction to the interior source term is quite subtle due to the notorious degeneracy at the sonic point. Fortunately, the final form of $\hat{s}_i^{\text{int}}$ is rather simple. We begin by the following identities which focus on the sonic point in equilibrium.

Lemma 3.10. *Let $Dm = DE = 0$, and let h_0 be the water height at the sonic point defined in (2.10). If $h_L = h_0$, then*

$$|Dh| = C_1|Db|^{1/2} \quad \text{with } C_1 := C_1(h_L, h_R) := \left(\frac{2h_R^2}{h_L + 2h_R}\right)^{1/2}. \tag{3.45}$$

Moreover,

$$\frac{1}{4}|Dh||Du|^2 = C_2|Db|^{3/2} \quad \text{with } C_2 := C_2(h_L, h_R) := \frac{gh_Lh_R}{2^{1/2}(h_L + 2h_R)^{3/2}}. \tag{3.46}$$

Analogous identities hold if $h_R = h_0$.

Proof. Suppose wlog that $h_L = h_0$. Using the identity $m^2 = gh_L^3$ and the definition of the energy E we obtain that

$$\begin{aligned} Db &= \frac{1}{g}\left(DE - \frac{1}{2}m^2 D\left(\frac{1}{h^2}\right)\right) - Dh \\ &= \frac{1}{2g}gh_L^3\frac{h_L^2 - h_R^2}{h_L^2h_R^2} - Dh \qquad (3.47) \\ &= -\frac{h_L + 2h_R}{2h_R^2}(Dh)^2, \qquad (3.48) \end{aligned}$$

so

$$(Dh)^2 = -\frac{2h_R^2}{h_L + 2h_R}Db, \tag{3.49}$$

which proves (3.45). Since $Dm = DE = 0$, we have

$$Du = -\frac{m}{h_Lh_R}Dh \tag{3.50}$$

and therefore

$$s\frac{1}{4}Dh(Du)^2 = \frac{m^2}{4h_L^2h_R^2}Dh^3 = \frac{m^2}{4h_L^2h_R^2}\left(\frac{2h_R^2}{h_L + 2h_R}\right)^{3/2}|Db|^{3/2} = \frac{gh_Lh_R}{2^{1/2}(h_L + 2h_R)^{3/2}}|Db|^{3/2}, \tag{3.51}$$

which shows (3.46). □

Based on this lemma, we can now estimate the term $\frac{1}{4}Dh(Du)^2$ in all equilibrium situations.

Lemma 3.11. *Let $[x_L, x_R]$ be the interior of the cell, and suppose that $Dm = DE = 0$.*

(i) Suppose that there is no sonic point in $[x_L, x_R]$. Then

$$\frac{1}{4}|Dh||Du|^2 \leqslant C_3|Db|^{3/2} \tag{3.52}$$

where C_3 is defined in (3.60) below.

(ii) Suppose that there is a sonic point $x_0 \in [x_L, x_R]$. Let $b_0 := b(x_0)$. Then

$$\frac{1}{4}|Dh||Du|^2 \leqslant C_4(|b_R - b_0| + |b_L - b_0|)^{3/2}. \tag{3.53}$$

Here C_4 is defined in (3.62).

Proof

(i) Suppose wlog that we are in the supersonic case, and

$$h_L < h_R < 1. \tag{3.54}$$

Then a direct computation yields

$$gDh = -gh_0 \int_{\widehat{E}_L}^{\widehat{E}_R} \frac{1}{\hat{\varphi}'(\hat{h}(\widehat{E}))}\,\mathrm{d}\widehat{E} = \frac{3}{2}gh_0 \int_{\widehat{E}_L}^{\widehat{E}_R} \frac{\hat{h}(\widehat{E})^3}{\hat{h}(\widehat{E})^3 - 1}\,\mathrm{d}\widehat{E} = \frac{3}{2}gh_0 \int_{\widehat{E}_R}^{\widehat{E}_L} \frac{\hat{h}(\widehat{E})^3}{1 - \hat{h}(\widehat{E})^3}\,\mathrm{d}\widehat{E}. \tag{3.55}$$

Now we shift the normalized energy towards the sonic point $\widehat{E} = 1$. Let

$$\widetilde{E} := 1 + \widehat{E} - \widehat{E}_R. \tag{3.56}$$

Since the corresponding $\hat{h}(\widetilde{E})$ is still to the left of the sonic point, with $\hat{h}(\widetilde{E}_L) < \hat{h}(\widetilde{E}_R) = 1$, and since the integrand on the RHS of (3.55) is monotonically increasing in that region, we obtain

$$0 \leqslant \frac{\hat{h}(\widehat{E})^3}{1 - \hat{h}(\widehat{E})^3} \leqslant \frac{\hat{h}(\widetilde{E})^3}{1 - \hat{h}(\widetilde{E})^3}. \tag{3.57}$$

Therefore,

$$gDh \leqslant \frac{3}{2}gh_0 \int_{\widehat{E}_R}^{\widehat{E}_L} \frac{\hat{h}(\widetilde{E})^3}{1 - \hat{h}(\widetilde{E})^3}\,\mathrm{d}\widehat{E} = \frac{3}{2}gh_0 \int_1^{\widetilde{E}_L} \frac{\hat{h}(\widetilde{E})^3}{1 - \hat{h}(\widetilde{E})^3}\,\mathrm{d}\widetilde{E} = g(\tilde{h}_R - \tilde{h}_L) = gD\tilde{h}. \tag{3.58}$$

Here, $\tilde{h}_{L/R}$ are the shifted heights. Denoting the shifted bottom by $\tilde{b}_{L/R}$, we observe that $D\tilde{b} = Db$. Now we can apply Lemma 3.10 and conclude that

$$|Dh| \leqslant |D\tilde{h}| = \widetilde{C}_1|Db|^{1/2} \quad \text{with } \widetilde{C}_1 := \left(\frac{2\tilde{h}_L^2}{\tilde{h}_R + 2\tilde{h}_L}\right)^{1/2} \tag{3.59}$$

and

$$\frac{1}{4}|Dh|(Du)^2 \leqslant C_3|Db|^{3/2} \quad \text{with } C_3 := \frac{m^2}{4h_L^2h_R^2}\widetilde{C}_1^3. \tag{3.60}$$

This proves (3.52).

(ii) Suppose now that $h_L < h_0 < h_R$. Then

$$\begin{aligned}\frac{1}{4}Dh(Du)^2 &= \frac{m^2}{4h_L^2h_R^2}(Dh)^3 = \frac{m^2}{4h_L^2h_R^2}(|h_L - h_0| + |h_R - h_0|)^3 \\ &= \frac{m^2}{4h_L^2h_R^2}(C_1(h_L, h_0)|b_L - b_0|^{1/2} + C_1(h_0, h_R)|b_R - b_0|^{1/2})^3 \leqslant C_4(|b_L - b_0| + |b_R - b_0|)^{3/2}\end{aligned} \tag{3.61}$$

with

$$C_4 := (\max(C_1(h_L, h_0), C_1(h_0, h_R)))^3 \frac{2^{3/2} m^2}{4 h_L^2 h_R^2}. \tag{3.62}$$

This proves (3.53). □

We are now ready to define $\hat{s}_i^{\text{int}}$. If we simply would set $\hat{s}_i^{\text{int}} = \frac{1}{4} Dh(Du)^2$, this would satisfy (3.41 and 3.43), but it would violate (3.42) in case a shock happens to cross the cell during that timestep. But we can limit this expression as follows: Let

$$\hat{s}_i^{\text{int}} := \beta q(\alpha/\beta), \tag{3.63}$$

where

$$\alpha := \frac{1}{4} Dh(Du)^2, \tag{3.64}$$

$$\beta := \begin{cases} C_4(|b_R - b_0| + |b_L - b_0|)^{3/2} & \text{if} \quad \exists x_0 \in [x_l, x_R], \\ C_3 |Db|^{3/2} & \text{otherwise.} \end{cases} \tag{3.65}$$

Here q is an odd, monotonically increasing, function in $C^{1,1}(\mathbb{R})$ satisfying

$$q(y) = y \quad \text{for } |y| \leqslant 1 \tag{3.66}$$

$$|q(y)| \leqslant 2 \quad \text{for all } y \in \mathbb{R} \tag{3.67}$$

(this degree of smoothness suffices at least for fourth-order accuracy). We choose q to be the piecewise quadratic function

$$q(z) := \begin{cases} z & \text{for} \quad 0 \leqslant z \leqslant 1 \\ -\frac{1}{4}(1 - 6z + z^2) & \text{for} \quad 1 \leqslant z \leqslant 3 \\ 2 & \text{for} \quad z \geqslant 3 \\ -q(-z) & \text{for} \quad z < 0. \end{cases}$$

Clearly, the symmetry condition (3.43) is satisfied. Now we check the smallness condition (3.42). Let us assume, for example, that the cell contains a sonic point. By (3.34) the bottom is Lipschitz continuous in the interior of each cell. Therefore,

$$|\hat{s}_i^{\text{int}}| \leqslant 2\beta = 2C_4(|b_R - b_0| + |b_L - b_0|)^{3/2} \leqslant 2C_4 L_b^{3/2} \Delta x^{3/2}. \tag{3.68}$$

Treating the case without sonic point analogously, we obtain that

$$|\hat{s}_i^{\text{int}}| \leqslant C \Delta x^{3/2}. \tag{3.69}$$

Therefore, the smallness condition (3.42) is satisfied.

Next we will verify (3.41) at equilibria. By definition, $\alpha/\beta \leqslant 1$ if $Dm = DE = 0$. Therefore, $q(\alpha/\beta) = \alpha/\beta$ and

$$\hat{s}_i^{\text{int}} = \beta(\alpha/\beta) = \alpha = \frac{1}{4} Dh(Du)^2, \tag{3.70}$$

so condition (3.41) is satisfied.

It is straightforward to check that for smooth data, (3.39) is second-order accurate as a quadrature for the source term, since we add a third-order difference to the term $-Df - g\bar{h}Db$. It is also symmetric. Thus it can be raised to any order of accuracy by extrapolation, as we will see in the next section. It reduces to the standard well-balanced quadrature for the lake at rest when $u_L = u_R = 0$.

3.3.5. High-order accuracy via extrapolation

The interior residual (3.39) is so far only second-order accurate. But we can directly adapt the extrapolation technique used in the paper of Noelle et al. [18], and obtain a high-order discretization.

We first subdivide each cell into N subcells and apply the quadrature (3.44) to all subcells.

Then we can have the following quadratures S_N:

$$S_N = \sum_{j=1}^{N} s_i^{\text{int}}(U_{j-1}^+, U_j^-, b_{j-1}^+, b_j^-), \tag{3.71}$$

where the subscript j means the value at the point $x_{i-\frac{1}{2}} + j\Delta x/N$. In the case of steady state, we have the following fact:

$$S_N = \sum_{j=1}^{N} s_i^{\text{int}}(U_{j-1}^+, U_j^-, b_{j-1}^+, b_j^-) = \sum_{j=1}^{N}(f(U_j^-) - f(U_{j-1}^+)) = f(U_N^-) - f(U_0^+) = f\left(U_{i+\frac{1}{2}}^-\right) - f\left(U_{i-\frac{1}{2}}^+\right).$$

This shows that S_N is also a second-order well-balanced approximation to the source term. Hence any linear combination of S_i is also a well-balanced approximation. Due to (3.43) the quadrature S_1 in (3.44) is second-order accurate and symmetric, therefore, there exists an asymptotic expansion:

$$S_N = S + c_1\left(\frac{\Delta x}{N}\right)^2 + c_2\left(\frac{\Delta x}{N}\right)^4 + \cdots, \tag{3.72}$$

where S represents the source term. Then the idea of extrapolation can provide an approximation to S with any order of accuracy by the combination of S_N. A well-balanced fourth-order approximation is given by

$$\frac{4S_2 - S_1}{3}. \tag{3.73}$$

Compared with the second-order discretization S_1, the fourth-order well-balanced scheme here needs one additional reconstructed point value at the cell center per cell, which is necessary for the computation of S_2. With this high-order discretization of the source term, the numerical scheme is complete, and we will show later that this scheme is in fact well-balanced.

3.4. Summary of the one-dimensional scheme

The fourth-order well-balanced scheme is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{U}_i := \frac{1}{\Delta x_i}\left(-F\left(\widetilde{U}_{i+\frac{1}{2}}^-, \widetilde{U}_{i+\frac{1}{2}}^+\right) + F\left(\widetilde{U}_{i-\frac{1}{2}}^-, \widetilde{U}_{i-\frac{1}{2}}^+\right) + s_i\right). \tag{3.74}$$

Here the function $F(\cdot,\cdot)$ is a conservative, Lipschitz continuous numerical flux consistent with the shallow water flux, i.e. $F(U,U) = f(U)$ for all U. The left and right values $\widetilde{U}_{i+\frac{1}{2}}^{\pm}$ at the cell interface are defined in (3.27).

From (3.29), (3.33) and (3.44) the total source term s_i is given by

$$s_i := \frac{4S_2 - S_1}{3} + f\left(\widetilde{U}_{i-\frac{1}{2}}^+\right) - f\left(\widetilde{U}_{i-\frac{1}{2}}^+\right) + f\left(\widetilde{U}_{i+\frac{1}{2}}^-\right) - f\left(\widetilde{U}_{i+\frac{1}{2}}^-\right), \tag{3.75}$$

where $\widetilde{U}_{i-\frac{1}{2}}^{\pm}$ is defined in (3.21). The extrapolated interior source term $(4S_2 - S_1)/3$ is defined by

$$S_1 := s_i^{\text{int}}\left(\widetilde{U}_{i-\frac{1}{2}}^+, \widetilde{U}_{i+\frac{1}{2}}^-, b_{i-\frac{1}{2}}^+, b_{i+\frac{1}{2}}^-\right) \tag{3.76}$$

$$S_2 := \left(s_i^{\text{int}}\left(\widetilde{U}_{i-\frac{1}{2}}^+, \widetilde{U}_i, b_{i-\frac{1}{2}}^+, b_i\right) + s_i^{\text{int}}\left(\widetilde{U}_i, \widetilde{U}_{i+\frac{1}{2}}^-, b_i, b_{i+\frac{1}{2}}^-\right)\right) \tag{3.77}$$

and the well-balanced quadrature of the source term s_i^{int} is given by (3.44)

$$s_i^{\text{int}}(U_L, U_R, b_L, b_R) := -g\bar{h}Db + \hat{s}_i^{\text{int}}, \tag{3.78}$$

where $\hat{s}_i^{\text{int}}$ is given by (3.63)–(3.65), and satisfies conditions (3.41)–(3.43). The scheme is completed by a TVD Runge–Kutta discretization [26] in time.

Algorithm 3.12. An implementation of this algorithm consists of the following steps:

1. Compute the initial cell average of U and bottom b based on the initial data. Apply the WENO reconstruction to $\bar{b}_i$ to obtain point values of b (may be ignored if bottom b is prescribed as a function of x).

2. At each time step, apply the usual WENO reconstruction procedure to the cell averages $\overline{U}_i$, and obtain $U^{\pm}_{i+\frac{1}{2}}$, hence $V^{\pm}_{i+\frac{1}{2}}$. Compute U_i and V_i to obtain fourth-order accuracy.
3. Compute the reference value $\overline{V}_i$ as the implicit solution of Eqs. (3.6)–(3.8).
4. Apply the equilibrium limiter (3.16) to the cell averages $\overline{V}_i, \overline{V}_{i\pm1}$, and to the point-values $V^{\pm}_{i+\frac{1}{2}}, V_i$, to get the limited values $\widetilde{V}^{\pm}_{i+\frac{1}{2}}$ and $\widetilde{V}_i$.
5. Compute the numerical fluxes on the RHS of (3.74).
6. Compute the high-order discretization of the source terms (3.75)–(3.78).
7. Apply a TVD Runge–Kutta scheme [26] to (3.74) to advance $\overline{U}_i(t)$ in time.

3.5. Well-balanced property and convergence to weak solutions

We begin this section by proving that our scheme is well-balanced for equilibria made of piecewise smooth regions separated by stationary shocks. Then we show that in the more general, non-stationary case limits of the scheme are weak solutions.

Collecting the results of the previous section it is straightforward to prove the following:

Theorem 3.13. *The WENO scheme (3.74)–(3.78) maintains smooth moving steady-state solutions (1.3) exactly and is high-order accurate. The same holds for the fully discrete scheme.*

Proof. Suppose that the initial data are a moving steady state, $V(x) \equiv \overline{V}$. Then Lemma 3.4 implies that all reference values $\overline{V}_i$ coincide with $\overline{V}$. Corollary 3.7 implies that $V(\widetilde{U}_i, b_i) = V(\widetilde{U}^{\pm}_{i+\frac{1}{2}}, b^{\pm}_{i+\frac{1}{2}}) = \overline{V}_i$. Now Lemma 3.9 implies that the interior residual vanishes, $r_i^{\text{int}} = 0$. Since we know from (3.32) that there is no residual in the topographic layer, $r_i^{\text{equi}} = 0$, it remains to show that the residual in the convective layer, r_i^{conv} vanishes as well. For this we study not only the values $\widetilde{U}^{+}_{i-\frac{1}{2}}$ and $\widetilde{U}_{i-\frac{1}{2}}$, but also the corresponding value $\widetilde{U}^{-}_{i-\frac{1}{2}}$ from the neighboring cell I_{i-1}. Since $\overline{V}_{i-1} = \overline{V}_i$, it follows that $\widetilde{U}^{-}_{i-\frac{1}{2}} = \widetilde{U}^{+}_{i-\frac{1}{2}}$ and hence $\hat{f}_{i-\frac{1}{2}} = f(\widetilde{U}^{+}_{i-\frac{1}{2}})$. Therefore,

$$r_i^{\text{conv}} = -\hat{f}_{i+\frac{1}{2}} + f\left(\widetilde{U}^{-}_{i+\frac{1}{2}}\right) + \hat{f}_{i-\frac{1}{2}} - f\left(\widetilde{U}^{+}_{i-\frac{1}{2}}\right) = 0 \tag{3.79}$$

and

$$r_i = r_i^{\text{int}} + r_i^{\text{equi}} + r_i^{\text{conv}} = 0, \tag{3.80}$$

so both the semidiscrete and the fully discrete schemes will preserve moving steady states.

We can easily check the two conditions of Lemma 3.1 are satisfied for our scheme. This proves the high-order accuracy. □

We can extend the well-balancedness result of the previous theorem to the case of piecewise smooth equilibrium solutions, where the smooth equilibria $Dm = DE = 0$ are separated by stationary shocks. Note that each smooth region will have its own constant value E_{loc}, and the Rankine–Hugoniot condition determines the jump in energy across the shock.

Theorem 3.14. *The WENO scheme (3.74)–(3.78) maintains piecewise smooth moving steady-state solution (1.3) exactly, if the stationary shocks separating the smooth regions are all located at cell boundaries, and computed by Roe's numerical flux function. The limiter procedure (3.16) in this case is replaced by a one-sided limiter for the two cells next to the shock. The same holds for the fully discrete scheme.*

Proof. The proof is completely analogous to that of the previous theorem. We only have to note that for a stationary shock located at $x_{i-\frac{1}{2}}$, Roe's solver gives

$$\hat{f}_{i-\frac{1}{2}} = f\left(\widetilde{U}^{+}_{i-\frac{1}{2}}\right).$$

This yields

$$r_i^{\mathrm{conv}} = -\hat{f}_{i+\frac{1}{2}} + f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) + \hat{f}_{i-\frac{1}{2}} - f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right) = 0. \tag{3.81}$$

The rest of the argument remains unchanged. □

Next we verify a Lax–Wendroff Theorem, that limits of our scheme are weak solutions. Let us first define the class of weak solutions which we have in mind. Let $\Omega = \mathbb{R} \times [0,T]$ be the domain and let $\varphi \in C^1(\Omega)$ be a test function. The difficulty is to give meaning to the source term integral

$$\int\int \varphi g h b_x \,\mathrm{d}x\,\mathrm{d}t \tag{3.82}$$

over the set where both b and h are discontinuous. The term hb_x has been called *non-conservative product* and has been extensively studied in the literature, see e.g. [19] and the references therein. We divide Ω into a regular set Ω_{reg} where the measure $hb_x \mathrm{d}x\,\mathrm{d}t$ is regular with respect to Lebesgue measure $\mathrm{d}x\,\mathrm{d}t$ (i.e. the topography b is Lipschitz continuous) and a singular set $\Omega_{\mathrm{sing}} = \Omega \backslash \Omega_{\mathrm{reg}}$. We assume that the singular set is a curve parametrized by t,

$$\Omega_{\mathrm{sing}} = \{(y(t),t) | 0 \leqslant t \leqslant T\} \tag{3.83}$$

(of course Ω_{sing} might also be a union of finitely many such curves). Then we blow up Ω_{sing} and shrink the set Ω_{reg} correspondingly using a parameter $\delta > 0$,

$$\Omega^\delta_{\mathrm{sing}} := \bigcup_{0\leqslant t\leqslant T} [y(t)-\delta, y(t)+\delta] \times \{t\} \tag{3.84}$$

$$\Omega^\delta_{\mathrm{reg}} := \Omega \setminus \Omega^\delta_{\mathrm{sing}}. \tag{3.85}$$

Clearly, we can define the integral over the regular set as

$$\int\int_{\Omega_{\mathrm{reg}}} \varphi g h b_x \,\mathrm{d}x\,\mathrm{d}t := \lim_{\delta\to 0} \int\int_{\Omega^\delta_{\mathrm{reg}}} \varphi g h b_x \,\mathrm{d}x\,\mathrm{d}t. \tag{3.86}$$

The treatment of the non-conservative product over the singular set is more involved:

$$\int\int_{\Omega_{sing}} \varphi g h b_x \,\mathrm{d}x\,\mathrm{d}t := \lim_{\delta\to 0} \int\int_{\Omega^\delta_{sing}} \varphi g h(b^\delta(x,t), m, E, \sigma) b^\delta_x(x,t) \mathrm{d}x\,\mathrm{d}t, \tag{3.87}$$

where for each t, $b^\delta(\cdot,t)$ is the continuous piecewise linear function on Γ_δ interpolating the three values

$$b^\delta(y(t)-\delta,t) = b_L(t), \quad b^\delta(y(t),t) = \min(b_L(t), b_R(t)), \quad b^\delta(y(t)+\delta,t) = b_R(t), \tag{3.88}$$

$h(b^\delta, m, E, \sigma)$ is the function defined in (2.19), and the equilibrium values m resp. E are the one-sided limits of m resp. E at that side where $b = \min(b_L, b_R)$. Wlog assume $b_L < b_R$, so $m = m_L$ and $E = E_L$.

Now we introduce the primitive of the function gh in (2.19) via

$$H(b,m,E,\sigma) := \int_{b_0(m,E)}^{b} gh(\hat{b}, m, E, \sigma)\mathrm{d}\hat{b}, \tag{3.89}$$

where we may choose $b_0(m,E) := \frac{1}{g}(E - \max_h(\frac{m^2}{2h^2} + gh))$. This allows us to rewrite the integral over the singular set $\Omega^\delta_{\mathrm{sing}}$ as

$$\int\int_{\Omega^\delta_{sing}} \varphi g h(b^\delta(x,t), m, E, \sigma) b^\delta_x(x,t) \mathrm{d}x\,\mathrm{d}t \tag{3.90}$$

$$= \int\int_{\Omega^\delta_{\mathrm{sing}}} \varphi \frac{\mathrm{d}}{\mathrm{d}x} H(b^\delta(x,t), m, E, \sigma) \mathrm{d}x\,\mathrm{d}t. \tag{3.91}$$

Taking the limit $\delta \to 0$, we obtain

$$\int\int_{\Omega_{sing}} \varphi g h b_x \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \varphi(y) \left(\lim_{\delta\to 0} \int_{y-\delta}^{y+\delta} \frac{\mathrm{d}}{\mathrm{d}x} H(b^\delta(x,t), m, E, \sigma) \mathrm{d}x \right) \mathrm{d}t = \int_0^T \varphi(y) DH \,\mathrm{d}t, \tag{3.92}$$

where as usual $DH := H(b_R, m, E, \sigma) - H(b_L, m, E, \sigma)$. If we introduce the average

$$\overline{\overline{gh}} := \frac{1}{b_R - b_L} \int_{b_L}^{b_R} gh(b, m, E, \sigma) \mathrm{d}b,$$

then we have the identity

$$DH = \overline{\overline{gh}} Db. \tag{3.93}$$

Now we are able to formulate the definition of a weak solution:

Definition 3.15. A function $U \in L^\infty(\Omega)$ is a weak solution of (1.1) if for all test functions $\varphi \in C^1(\Omega)$

$$\int\int_\Omega (\varphi_t U + \varphi_x f(U)) \mathrm{d}x \, \mathrm{d}t = \int_{\partial\Omega} (f(U), U) \cdot n\varphi \, \mathrm{d}S + \int\int_{\Omega_{\mathrm{reg}}} \varphi g h b_x \, \mathrm{d}x \, \mathrm{d}t + \int_{\Omega_{\mathrm{sing}}} \varphi \overline{\overline{gh}} Db \, \mathrm{d}t. \tag{3.94}$$

Remark 3.16

(i) For a systematic introduction to weak solutions using the theory of non-conservative products we refer to [7,19] and the references therein. The reader should note that definition (3.92) of non-conservative products is not the only one possible. Other definitions would lead to different classes of weak solutions, and in order to approximate them one would have to adapt the quadrature rules for the source.

(ii) LeRoux and collaborators [15,5] have constructed a solution operator to the Riemann-problem with variable bottom.

(iii) Our definition of a weak solution is motivated by considering the particular steady solution of a waterfall over steep or discontinuous terrain, see Figs. 3 and 4. The water flows in supercritically from the left with $(b, m, E, \sigma) = (b_L, m_L, E_L, 1)$ and $h = h(b_L, m_L, E_L, 1)$. As the water flows down the slide, b decreases. As can be seen from the supercritical region in Fig. 1, the height h decreases correspondingly to the value $h_C = h(b_R, m_L, E_L, 1)$, and the flow accelerates to $u = m/h_C$. Across the stationary shock momentum remains constant ($m_R = m_L$), the height jumps to the value $h_R = \frac{h_C}{2}(-1 + \sqrt{1 + k^2})$ with $k^2 = \frac{8m^2}{gh_C^3}$. It is interesting to observe that the equilibrium energy decreases by a cubic term,

$$E_R = E_L - \frac{g(Dh)^3}{4h_C h_R}.$$

As the slide becomes infinitely steep, the waterfall converges to a weak solution in the sense of Definition 3.15.

We are now ready to prove the following Lax–Wendroff theorem:

Theorem 3.17. *Suppose that according to Assumption 3.8, the bottom is piecewise smooth, contains at most finitely many jump discontinuities, and the reconstructions $b_\Delta(x)$ are uniformly and globally Lipschitz continuous*

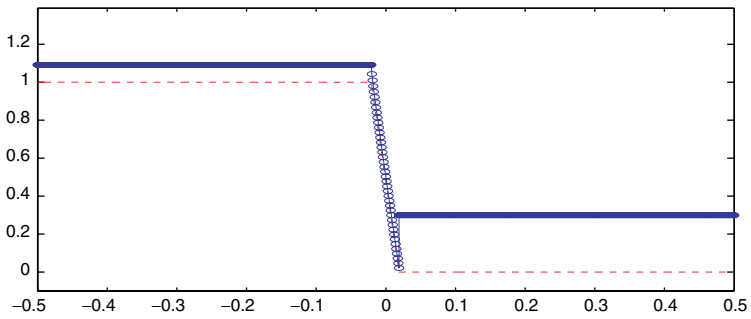

Fig. 3. Waterfall: a (finitely or infinitely) steep slide followed by a stationary shock. Dashed line: bottom topography. Circles: Water surface.

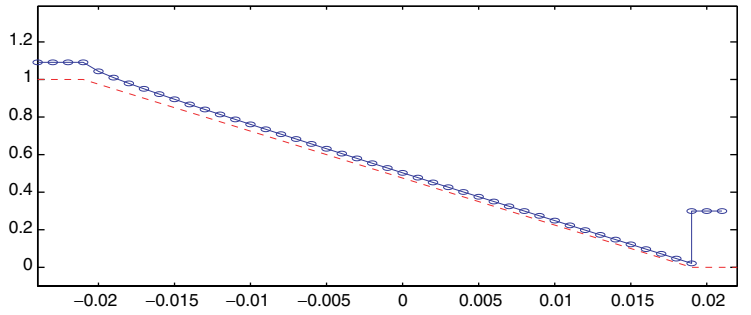

Fig. 4. Detail of waterfall. $x \in [-0.024, 0.022]$.

in the interior of the cells. Suppose furthermore that the approximate solutions U_Δ defined by Algorithm 3.12 converge uniformly almost everywhere to a function $U \in L^\infty(\Omega)$. Then U is a weak solution of (1.1).

Remark 3.18. Note that we admit discontinuous solution and jumps in the topography which satisfy (3.34).

Proof. Since the proof of the Lax–Wendroff Theorem is classical, we focus on the terms which are due to our new well-balanced quadrature. For simplicity we restrict ourselves to the semidiscrete scheme and the case that the bottom topography is independent of time. Let φ be a test function and $\varphi_i = \varphi(x_i)$. From (3.74) we have to study the term

$$\sum_i \Delta x_i \varphi_i \left(\frac{\mathrm{d}}{\mathrm{d}t} \overline{U}_i - \frac{1}{\Delta x_i} \left(-F(\widetilde{U}^-_{i+\frac{1}{2}}, \widetilde{U}^+_{i+\frac{1}{2}}) + F(\widetilde{U}^-_{i-\frac{1}{2}}, \widetilde{U}^+_{i-\frac{1}{2}}) + s_i \right) \right). \tag{3.95}$$

We are particularly interested in the source term. From (3.75) we obtain

$$\sum_i \Delta x_i \varphi_i \frac{s_i}{\Delta x_i} = \sum_i \varphi_i \left(\frac{4S_2 - S_1}{3} + f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right) - f\left(\widetilde{U}^+_{i-\frac{1}{2}}\right) + f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) - f\left(\widetilde{U}^-_{i+\frac{1}{2}}\right) \right). \tag{3.96}$$

The flux differences on the RHS of (3.96) result from s_i^{equi} as defined in (3.33) and are differences across the topographic, or equilibrium layer where the bottom may jump. From (3.40) we have

$$Df = -g\bar{h} Db + \frac{1}{4} Dh (Du)^2.$$

Since we are in the equilibrium layer, where $Dm = DE = D\sigma = 0$, the height is given by $h = h(b, m, E, \sigma)$, see (2.19). This implies the identities

$$\begin{aligned} b &= \frac{1}{g}\left(E - \frac{m^2}{2h^2}\right) - h \\ \frac{\mathrm{d}b}{\mathrm{d}h} &= \frac{m^2}{gh^3} \\ Db &= \frac{m^2 \bar{h}}{g h_L^2 h_R^2} Dh \\ DH &= \frac{m^2}{h_L h_R} Dh. \end{aligned} \tag{3.97}$$

From these we obtain that

$$Df = -DH. \tag{3.98}$$

Thus the source term in the equilibrium layer converges to the last term on the RHS of (3.94) if the bottom is discontinuous. If the bottom is smooth, according to (3.35) the WENO reconstruction $b_{i\pm\frac{1}{2}}$ will contain at most a jump of $\mathcal{O}(\Delta x^2)$. Therefore,

$$Df = -\frac{DH}{Db} Db = \mathcal{O}(\Delta x^2)$$

as well and the corresponding term in (3.96) will vanish in the limit.

Next we study the terms S_1 and S_2. We begin with S_1 as defined by (3.76) and (3.44):

$$S_1 = s_i^{\text{int}}\left(\widetilde{U}^+_{i-\frac{1}{2}}, \widetilde{U}^-_{i+\frac{1}{2}}, b^+_{i-\frac{1}{2}}, b^-_{i+\frac{1}{2}}\right) = -g\bar{h}Db + \hat{s}_i^{\text{int}}.$$

Note that we are now in the regular set Ω_{reg}, where the topography is smooth. Therefore, the term

$$-\sum_i \Delta x_i \varphi_i g\bar{h}\frac{Db}{\Delta x_i}$$

will converge to the corresponding source term on the LHS of (3.94). By (3.34) and (3.42) the remaining term is

$$-\sum_i \varphi_i \hat{s}_i^{\text{int}} = o(\Delta x)\sum_i \varphi_i = o(1) \quad \text{as} \quad \Delta x \to 0$$

and hence vanishes in the limit. The term S_2 can be treated by the same argument. This concludes the proof. □

4. One-dimensional numerical results

In this section we present numerical results of our fourth-order finite volume WENO scheme satisfying the well-balanced property for the one-dimensional shallow water equation (1.1). In all the examples, time discretization is by the classical third-order TVD Runge–Kutta method [26], and the CFL number is taken as 0.6, except for the accuracy tests where smaller time step is taken to ensure that spatial errors dominate. The gravitation constant g is taken as 9.812 m/s^2.

To measure the extra cost of our well-balancing, we have compared runtimes for two of the numerical tests below. The CPU times of the new scheme are 72% and 84%, respectively, more than those of a traditional non-well-balanced WENO scheme with trivial treatment of the source term.

4.1. Well-balanced test

The purpose of the first test problems is to verify the well-balanced property of our algorithm towards the moving steady-state solution. These steady-state problems are classical test cases for transcritical and subcritical flows, and they are widely used to test numerical schemes for shallow water equations. For example, they have been used as a test case in [27]. Here, our purpose is to maintain these steady-state solutions exactly.

The bottom function is given by

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \leqslant x \leqslant 12, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

for a channel of length 25 m. Three steady states, subcritical or transcritical flow with or without a steady shock will be investigated.

(a) *Transcritical flow without a shock:*

The initial condition is given by

$$E = \frac{1.53^2}{2\times 0.66^2} + 9.812\times 0.66, \quad m = 1.53, \tag{4.2}$$

together with the boundary condition

- *upstream*: the discharge $hu = 1.534$ m^2/s is imposed;
- *downstream*: the water height $h = 0.66$ m is imposed when the flow is subcritical.

This steady state should be exactly preserved. We compute the solution until $t = 20$ using $N = 200$ uniform mesh points. The computed surface level $h + b$ and the bottom b are plotted in Fig. 5. In order to demonstrate that the steady state is indeed maintained up to round-off error, we use single precision and double precision to perform the computation, and show the L^1 and L^∞ errors for the water height h and the discharge hu (note: neither h nor hu in this case is a constant or polynomial function!) in Table 1 for different precisions. We can clearly see that the L^1 and L^∞ errors are at the level of round-off errors for different precisions, verifying the well-balanced property (see Table 1):

(b) *Transcritical flow with a shock. The initial condition is given by*

$$E = \begin{cases} \frac{3}{2}(9.812 \times 0.18)^{\frac{2}{3}}) + 9.812 \times 0.2 & \text{if } x \leqslant 11.665504281554291, \\ \frac{0.18^2}{2\times 0.33^2} + 9.812 \times 0.33 & \text{otherwise,} \end{cases} \qquad m = 0.18, \tag{4.3}$$

together with the boundary condition

- *upstream*: the discharge $hu = 0.18\ \mathrm{m}^2/\mathrm{s}$ is imposed;
- *downstream*: the water height $h = 0.33\ \mathrm{m}$ is imposed.

This steady state should be exactly preserved. As we mentioned in Section 3.5, we only discuss the case when the shock is exactly located at the cell boundary. Hence we shift the computational domain to put the shock at the cell boundary. As we mentioned in Theorem 3.14, for this case when stationary shock exists, we need to use the Roe's flux to compute the approximate Riemann problem (3.28), and replace the limiter procedure (3.16) by a one-sided limiter for the two cells next to the shock, i.e., the following formula is used instead of (3.16) and (3.17):

$$\lim(w; \bar{w}_i, \bar{w}_{i\pm1}) := \bar{w}_i + \lambda(w - \bar{w}_i), \tag{4.4}$$

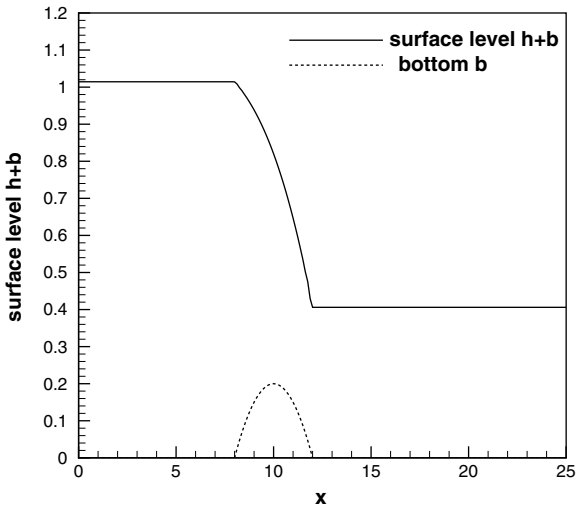

Fig. 5. The surface level $h + b$ and the bottom b for the transcritical flow without a shock.

Table 1
$L1$ and L^∞ errors for different precisions for the transcritical flow without a shock

Precision	L^1 error		L^∞ error	
	h	hu	h	hu
Single	2.19E − 08	4.74E − 09	1.61E − 06	1.19E − 07
Double	1.15E − 16	3.21E − 16	5.55E − 16	1.33E − 15

where

$$\lambda := \min\left(1, \frac{|\bar{w}_{i-1} - \bar{w}_i|^2}{|w - \bar{w}_i|^2}\right), \quad \text{or} \quad \min\left(1, \frac{|\bar{w}_{i+1} - \bar{w}_i|^2}{|w - \bar{w}_i|^2}\right), \tag{4.5}$$

depending on whether the cell is on the left or right side of the shock. Also, we mentioned that the left and right approximated values of bottom at the shock must be exact, so that the Roe's flux can capture this shock exactly. Here we compute the solution until $t = 20$ using $N = 400$ uniform mesh points. The computed surface level $h + b$ and the bottom b are plotted in Fig. 6. In order to demonstrate that the steady state is indeed maintained up to round-off error, we use single precision and double precision to perform the computation, and show the L^1 and L^∞ errors for the water height h and the discharge hu in Table 2 for different precisions. We can clearly see that the L^1 and L^∞ errors are at the level of round-off errors for different precisions, verifying the well-balanced property.

(c) *Subcritical flow*. The initial condition is given by

$$E = 22.06605, \quad m = 4.42, \tag{4.6}$$

together with the boundary condition

- *upstream*: the discharge $hu = 4.42\ \mathrm{m}^2/\mathrm{s}$ is imposed;
- *downstream*: the water height $h = 2\ \mathrm{m}$ is imposed. when the flow is subcritical.

This steady state should be exactly preserved. We compute the solution until $t = 20$ using $N = 200$ uniform mesh points. The computed surface level $h + b$ and the bottom b are plotted in Fig. 7. In order to demonstrate that the steady state is indeed maintained up to round-off error, we use single precision and double precision to perform the computation, and show the L^1 and L^∞ errors for the water height h and the discharge hu in Table 3

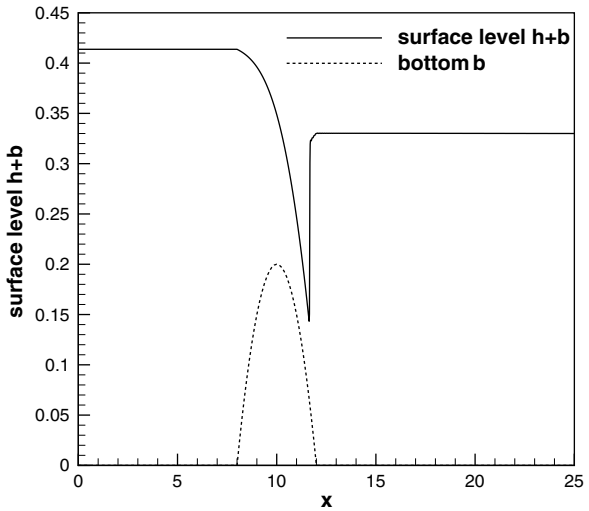

Fig. 6. The surface level $h + b$ and the bottom b for the transcritical flow with a shock.

Table 2
L^1 and L^∞ errors for different precisions for the transcritical flow with a shock

Precision	L^1 error		L^∞ error	
	h	hu	h	hu
Single	2.78E − 09	2.74E − 09	3.87E − 07	2.53E − 07
Double	1.06E − 15	1.23E − 15	8.37E − 14	8.32E − 14

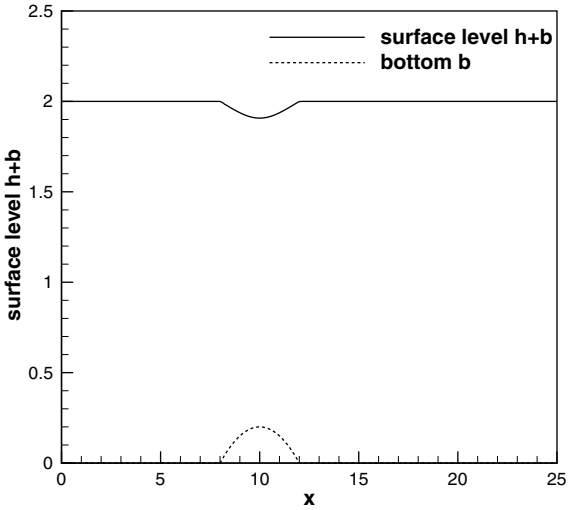

Fig. 7. The surface level $h + b$ and the bottom b for the subcritical flow.

Table 3
L^1 and L^∞ errors for different precisions for the subcritical flow

Precision	L^1 error		L^∞ error	
	h	hu	h	hu
Single	4.62E − 07	3.23E − 07	6.81E − 06	7.23E − 06
Double	1.44E − 17	8.84E − 17	6.66E − 16	1.77E − 15

for different precisions. We can clearly see that the L^1 and L^∞ errors are at the level of round-off errors for different precisions, verifying the well-balanced property.

4.2. Testing the orders of accuracy

In this example we will test the high-order accuracy of our schemes for a smooth solution. Following the examples presented in [30], we have the bottom function and initial conditions

$$b(x) = \sin^2(\pi x), \quad h(x,0) = 5 + e^{\cos(2\pi x)}, \quad (hu)(x,0) = \sin(\cos(2\pi x)), \quad x \in [0,1]$$

with periodic boundary conditions. Since the exact solution is not known explicitly for this case, we use the fifth-order finite volume non-well-balanced WENO scheme with $N = 12{,}800$ cells to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. We compute up to $t = 0.1$ when the solution is still smooth (shocks develop later in time for this problem). Table 4 contains the

Table 4
L^1 errors and numerical orders of accuracy for the example in Section 4.2

No. of cells	CFL	h		hu	
		L^1 error	Order	L^1 error	Order
25	0.6	1.48E − 02		9.78E − 02	
50	0.6	2.41E − 03	2.68	1.97E − 02	2.31
100	0.4	2.97E − 04	3.02	2.58E − 03	2.93
200	0.3	2.44E − 05	3.61	2.13E − 04	3.60
400	0.2	1.03E − 06	4.56	8.97E − 06	4.57
800	0.1	3.49E − 08	4.89	2.95E − 07	4.93

L^1 errors for the cell averages and numerical orders of accuracy for the finite volume schemes, respectively. Notice that the CFL number we have used decreases with the mesh size and is recorded in Table 4. We can easily observe the fifth-order accuracy for the WENO schemes. Note that the fifth-order WENO reconstruction has been used in space, but the source term is approximated by a fourth-order accurate extrapolation. Hence the approximation of the source term in the algorithm contributes less to the overall error. This phenomena has been investigated in [18].

4.3. A small perturbation of a moving steady-state water

The following test case is chosen to demonstrate the capability of the proposed scheme for computations on the perturbation of a steady-state solution, which cannot be captured well by a non well-balanced scheme.

In the Section 4.1, we presented three steady-state solutions and showed that our numerical schemes did maintain them exactly. In this test case, we impose to them a small perturbation 0.01 on the height in the interval $[5.75, 6.25]$.

Theoretically, this disturbance should split into two waves, propagating to the left and right respectively. Many numerical methods have difficulty with the calculations involving such small perturbations of the water surface. The solution obtained on a 200 cell uniform grid with simple transmissive boundary conditions, compared with the results using 2000 uniform cells, is shown in Fig. 8 for the transcritical flow without a shock, in Fig. 9 for the transcritical flow with a shock and in Fig. 10 for the subcritical flow. The stopping time T is set as 1.5 for the first and third flow, 3 for the second flow. At this time, the downstream-traveling water pulse has already passed the bump. We can clearly see that there are no spurious numerical oscillations and the resolution for the propagated small perturbation is very good.

4.4. The dam-break-problem over a rectangular bump

In this traditional test case we simulate the dam breaking problem over a rectangular bump, which produces a rapidly varying flow over a discontinuous bottom topography. This example was used in [28,30,18].

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leqslant 1500/8, \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

for $x \in [0, 1500]$. The initial conditions are

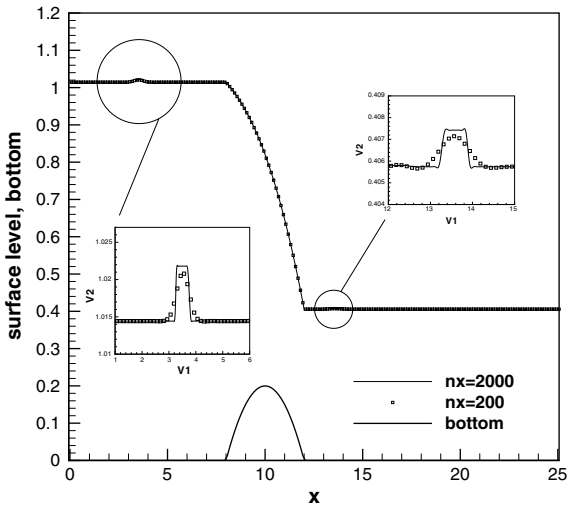

Fig. 8. Small perturbation of the transcritical flow without a shock.

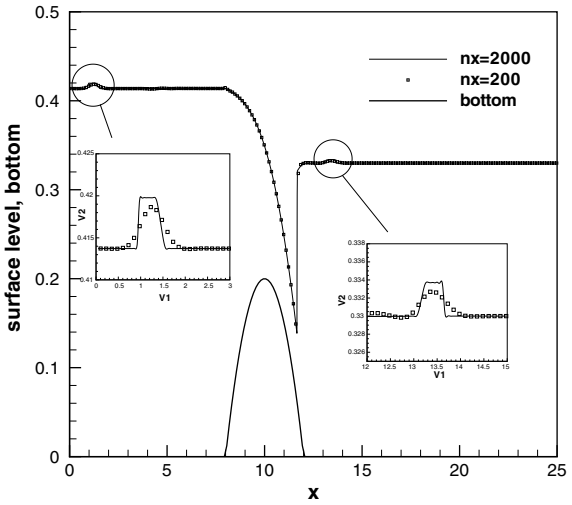

Fig. 9. Small perturbation of the transcritical flow with a shock.

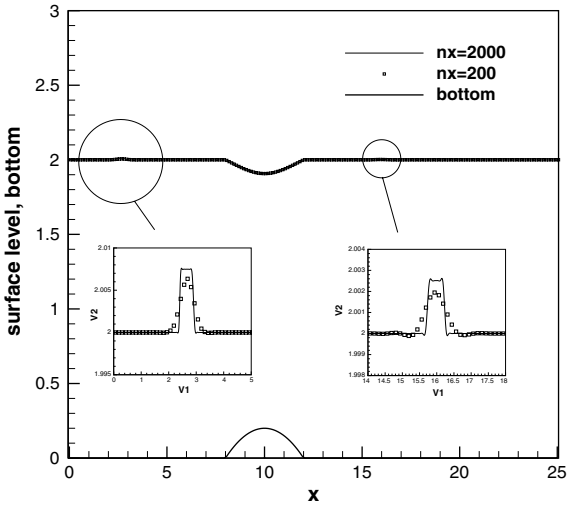

Fig. 10. Small perturbation of the subcritical flow.

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 20 - b(x) & \text{if } x \leqslant 750, \\ 15 - b(x) & \text{otherwise.} \end{cases} \tag{4.8}$$

We use open boundary conditions on both sides. In the beginning, we observe the standard rarefaction and shock waves which form the solution of the Riemann problem of the homogeneous shallow water equations. The numerical results with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Fig. 11 at ending time $t = 15$ s. At time $T \approx 17$, the waves reach the discontinuous edges of the bottom. After that, a part of the wave is transmitted, another part reflected, and a remaining part becomes a standing wave. Later on, this wave system keeps interacting. When the time T reaches 60, six waves appear in our solution. The numerical results with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Fig. 12 at the ending time $t = 60$ s.

In this example, the water height $h(x)$ is discontinuous at the points $x = 562.5$ and $x = 937.5$. Our scheme works well for this example, giving well resolved, non-oscillatory solutions using 400 cells which agree with the converged results using 4000 cells.

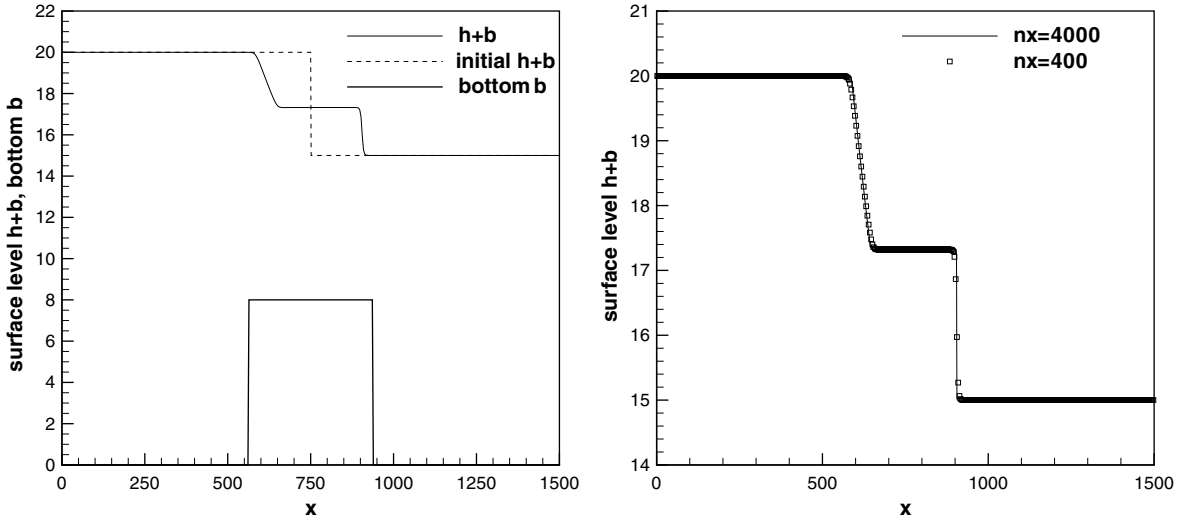

Fig. 11. The surface level $h + b$ for the dam breaking problem at time $t = 15$ s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

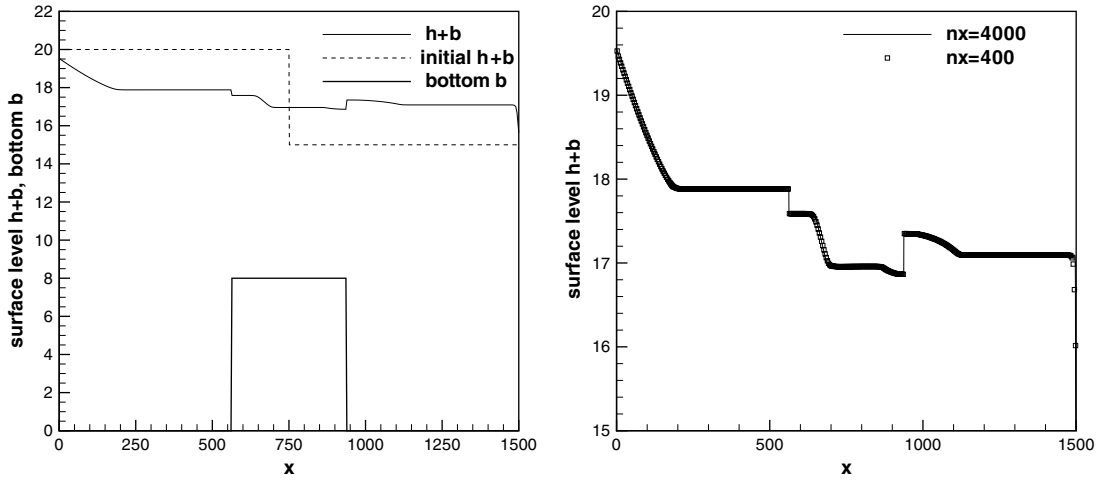

Fig. 12. The surface level $h + b$ for the dam breaking problem at time $t = 60$ s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

5. A two-dimensional example

The shallow water system in two space dimensions takes the form:

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0, \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x + (huv)_y = -ghb_x, \\ (hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2}gh^2\right)_y = -ghb_y, \end{cases} \tag{5.1}$$

where again h is the water height, (u, v) is the velocity of the fluid, b represents the bottom topography and g is the gravitational constant.

Our focus of this paper is on one-dimensional problems. In two spatial dimensions there is an abundance of steady states, and it is much more difficult to identify the interesting ones. It is in principle possible to extend

the techniques in this paper to obtain well-balanced schemes for some of the truly two-dimensional moving water steady states, but the procedure is significantly more complicated and we will not discuss such extensions in this paper. In this section we only consider our one-dimensional well-balanced scheme designed in previous sections, trivially generalized to two dimensions by using our well-balanced WENO algorithm in both directions. In the x direction, we first apply the usual WENO reconstruction procedure to obtain $U^{\pm}_{i+\frac{1}{2},j}$, where $U = (h, hu, hv)^T$. Then based on $h^{\pm}_{i+\frac{1}{2},j}$ and $hu^{\pm}_{i+\frac{1}{2},j}$, we repeat steps 2–4 of Algorithm 3.12 to obtain $\tilde{h}^{\pm}_{i+\frac{1}{2},j}$ and $\tilde{hu}^{\pm}_{i+\frac{1}{2},j}$. By keeping hv unchanged, we define $\widetilde{U}^{\pm}_{i+\frac{1}{2},j}$ as $(\tilde{h}^{\pm}_{i+\frac{1}{2},j}, \tilde{hu}^{\pm}_{i+\frac{1}{2},j}, hv^{\pm}_{i+\frac{1}{2},j})^T$ and then follow steps 5–6 of Algorithm 3.12. Notice that this procedure should be carried out for more than one quadrature points along the edge $x = x_{i+\frac{1}{2}}$ rather than just for the middle point $y = y_j$ as described above, in order to ensure higher than second-order accuracy. We are using $U^{\pm}_{i+\frac{1}{2},j}$ at the middle point $y = y_j$ above simply to demonstrate the ideas. The same procedure is applied to the y direction, where hu remains unchanged. We take a numerical

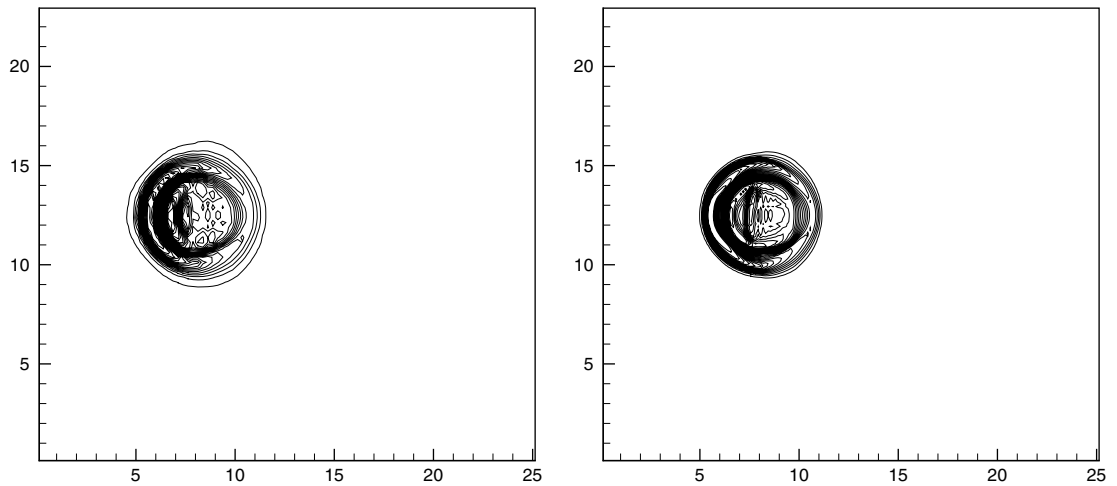

Fig. 13. The contours of the difference between the height h and the initial steady state (5.2) for the problem in Section 5 at time $t = 0.5$. 30 uniformly spaced contour lines from -0.009 to 0.012. Left: results with a 100×100 uniform mesh. Right: results with a 200×200 uniform mesh.

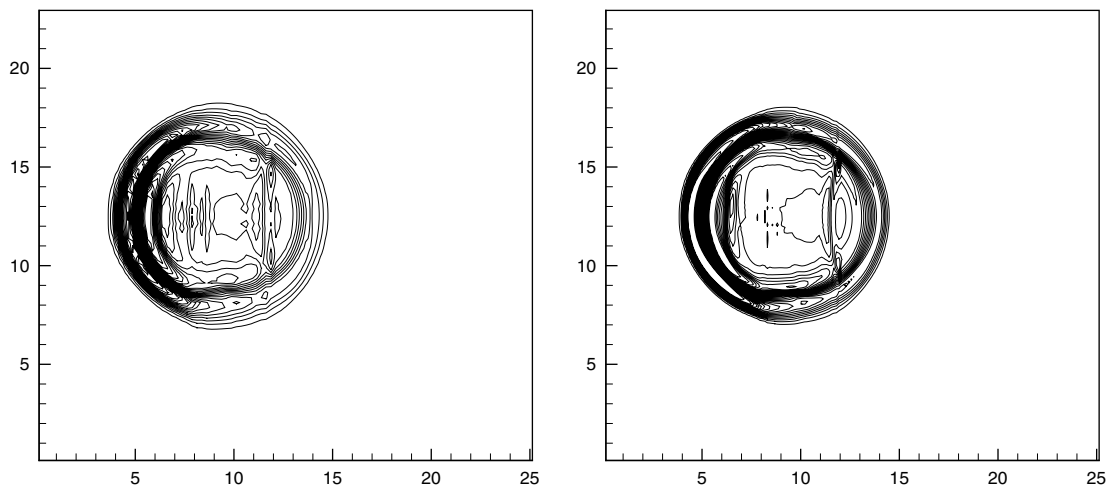

Fig. 14. The contours of the difference between the height h and the initial steady state (5.2) for the problem in Section 5 at time $t = 1$. 30 uniformly spaced contour lines from -0.005 to 0.008. Left: results with a 100×100 uniform mesh. Right: results with a 200×200 uniform mesh.

example which is a two-dimensional perturbation of a one-dimensional moving steady water, and compare our scheme with the regular, non-well-balanced WENO scheme.

Similar to the one-dimensional case, we use the classical third-order TVD Runge–Kutta time discretization with CFL number 0.6.

We solve the system in the rectangular domain $[0,25]\times[0,25]$. The bottom topography is given by

$$b(x,y)=\begin{cases}0.2-0.05(x-10)^2 & \text{if } 8\leqslant x\leqslant 12,\\ 0 & \text{otherwise.}\end{cases} \tag{5.2}$$

Notice that the bottom is a function of x only. A steady-state solution can be computed from:

$$\frac{1}{2}u^2+g(h+b)=22.06605,\quad hu(x,y,0)=4.42,\quad hv(x,y,0)=0. \tag{5.3}$$

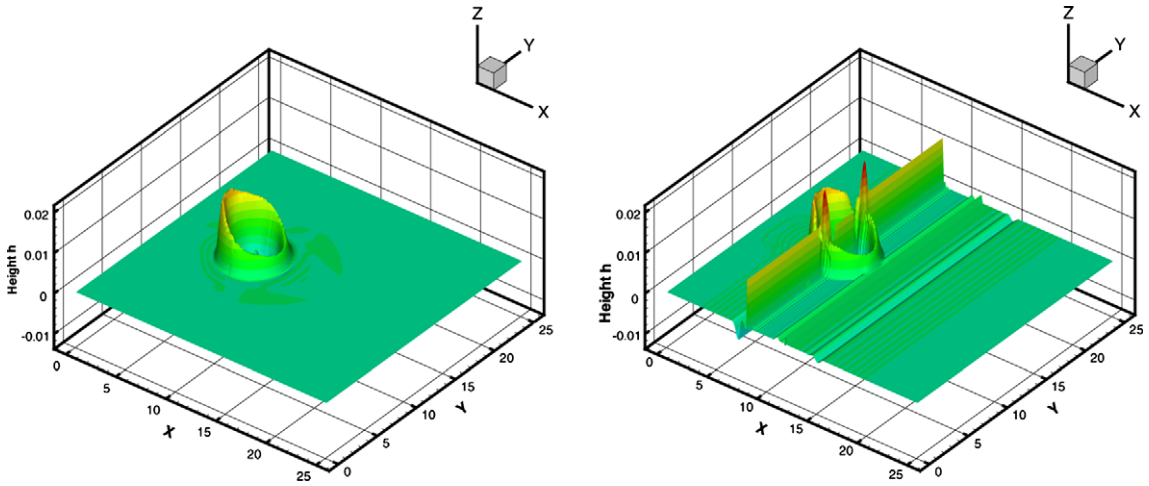

Fig. 15. The 3D figure of the difference between the height h and the initial steady state (5.2) for the problem in Section 5 at time $t=0.5$ with a 200×200 uniform mesh. Left: results based on well-balanced scheme. Right: results based on non-well-balanced scheme.

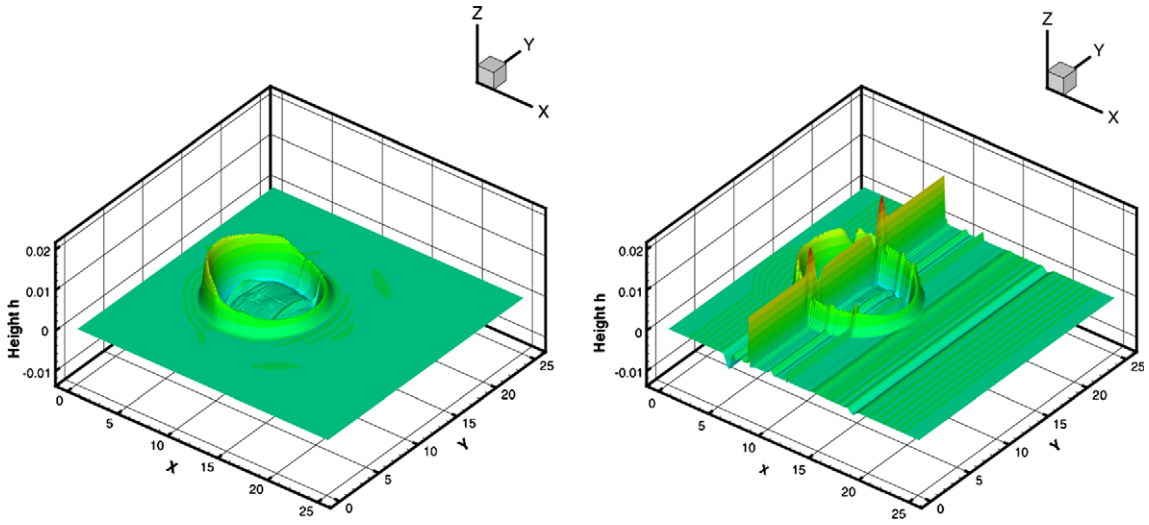

Fig. 16. The 3D figure of the difference between the height h and the initial steady state (5.2) for the problem in Section 5 at time $t=1$ with a 200×200 uniform mesh. Left: results based on well-balanced scheme. Right: results based on non-well-balanced scheme.

These data correspond precisely to the one-dimensional subcritical steady state of (4.6), and the cross section of the unperturbed solution can be seen in Fig. 7. Our initial condition is given by a two-dimensional small perturbation of that steady state, where h is perturbed upward by 0.05 in the box $6.5 \leqslant x \leqslant 7.5$, $12 \leqslant y \leqslant 13$. Figs. 13 and 14 display the disturbance as it interacts with the hump, on two different uniform meshes with 100×100 cells and 200×200 cells for comparison. The difference between the height h and the initial steady state (5.2) is presented at different times $t = 0.5$ and $t = 1$. We also run the same numerical test with the well-balanced fifth-order finite volume WENO scheme for the lake at rest. Note that this scheme is not well-balanced for moving equilibria. The comparison of the numerical results are presented in Figs. 15 and 16. The results indicate that our well-balanced scheme can resolve the complex small features of the flow very well, without spurious features which do appear in the results obtained with the regular non-well-balanced WENO scheme.

6. Concluding remarks

In this paper we have constructed well-balanced schemes of arbitrary order of accuracy for the moving steady-state solutions of the shallow water equations. The new schemes extend the techniques used in our previous work for still steady water [18,33]. Special reconstruction procedure and source term discretization are introduced such that the resulting WENO schemes balance the moving steady-state solution to machine accuracy. In this first implementation, the new code needs about 80% more CPU time than a traditional, non-well-balanced WENO scheme with trivial treatment of the source term. Numerical examples are given to demonstrate the well-balanced property, accuracy, good capturing of the small perturbation to the steady-state solutions, and non-oscillatory shock resolution of the proposed numerical method. Although the new schemes are designed for the shallow water equation, the idea can be generalized to many other balance laws.

Acknowledgements

This work was started at a workshop held at the ''American Institute of Mathematics'' (AIM) in Palo Alto. We thank AIM for providing a great environment for discussing and collaborating. Part of this work was done while the first author was in residence at CMA, ''Center of Mathematics for Application'' at Oslo University, and he thanks CMA and its members for their generous hospitality.

References

[1] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows, SIAM J. Sci. Comput. 25 (2004) 2050–2065.
[2] D.S. Bale, R.J. LeVeque, S. Mitran, J.A. Rossmanith, A wave-propagation method for conservation laws with spatially varying flux functions, SIAM J. Sci. Comput. 24 (2002) 955–978.
[3] N. Botta, R. Klein, S. Langenberg, S. Lützenkirchen, Well-balanced finite volume methods for nearly hydrostatic flows, J. Comp. Phys. 196 (2004) 539–565.
[4] M. Castro, J.M. Gallardo, C. Parés, High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products, Applications to shallow-water systems, Math. Comp. 75 (2006) 1103–1134.
[5] A. Chinnayya, A.Y. LeRoux, N. Seguin, A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon, Int. J. Fin. 1 (2004) (electronic).
[6] V.T. Chow, Open-channel Hydraulics, McGraw Hill, New York, 1959.
[7] G. Dal Maso, P. Lefloch, F. Murat, Definition and weak stability of nonconservative products, J. Math. Pures Appl. 74 (1995) 483–548.
[8] L. Gosse, A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms, Comput. Math. Appl. 39 (2000) 135–159.
[9] J.M. Greenberg, 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] G. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes, J. Comput. Phys. 126 (1996) 202–228.
[11] S. Jin, A steady-state capturing method for hyperbolic systems with geometrical source terms, Math. Model. Numer. Anal. (M^2AN) 35 (2001) 631–646.
[12] S. Jin, X. Wen, An efficient method for computing hyperbolic systems with geometrical source terms having concentrations, J. Comput. Math. 22 (2004) 230–249.

[13] S. Jin, X. Wen, Two interface type numerical methods for computing hyperbolic systems with geometrical source terms having concentrations, SIAM J. Sci. Comp. 26 (2005) 2079–2101.
[14] A. Kurganov, D. Levy, Central-upwind schemes for the Saint-Venant system, Math. Model. Numer. Anal. (M^2AN) 36 (2002) 397–425.
[15] A.Y. Leroux, Discrétisation des termes sources raides dans les problèmes hyperboliques, In Systèmes hyperboliques: noveau schemas et nouvelles applications. Ecoles CEA-EDF-INRIA ''problèmes non linéaires appliqués'', INRIA Rocquencourt (France), March 1998. <http://www-gm3.univ-mrs.fr/leroux/publications/ay.le_roux.html> (in French).
[16] X.-D. Liu, S. Osher, T. Chan, Weighted essentially nonoscillatory schemes, J. Comput. Phys. 115 (1994) 200–212.
[17] M. Lukáčová-Medvid'ová, S. Noelle, M. Kraft, Well-balanced finite volume evolution Galerkin methods for the shallow water equations, J. Comput. Phys. 221 (2007) 122–147.
[18] S. Noelle, N. Pankratz, G. Puppo, J.R. Natvig, Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows, J. Comput. Phys. 213 (2006) 474–499.
[19] C. Parés, ;Numerical methods for nonconservative hyperbolic systems. A theoretical framework, SIAM J. Numer. Anal. 44 (2006) 300–321.
[20] C. Parés, M. Castro, On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems, Math. Model. Numer. Anal. (M^2AN) 38 (2004) 821–852.
[21] G. Russo, Central schemes for balance laws, Hsyperbolic problems: theory, numerics, applications, vols. I, II (Magdeburg, 2000), 821–829, Internat. Ser. Numer. Math., 140, 141, Birkhäuser, Basel, 2001.
[22] G. Russo, Central schemes for conservation laws with application to shallow water equations, in: S. Rionero, G. Romano (Eds.), Trends and Applications of Mathematics to Mechanics: STAMM 2002, Springer Verlag, Italia SRL, 2005, pp. 225–246.
[23] J. Shi, C. Hu, C.-W. Shu, A technique of treating negative weights in WENO schemes, J. Comput. Phys. 175 (2002) 108–127.
[24] C.-W. Shu, TVB uniformly high-order schemes for conservation laws, Math. Comp. 49 (1987) 105–121.
[25] C.-W. Shu, Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws, in: B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor, A. Quarteroni (Eds.), Advanced numerical approximation of nonlinear hyperbolic equations, Lecture Notes in Mathematics, vol. 1697, Springer, 1998, pp. 325–432.
[26] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (1988) 439–471.
[27] M.E. Vazquez-Cendon, Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry, J. Comput. Phys. 148 (1999) 497–526.
[28] S. Vukovic, L. Sopta, ENO and WENO schemes with the exact conservation property for one-dimensional shallow water equations, J. Comput. Phys. 179 (2002) 593–621.
[29] X. Wen, A steady state capturing and preserving method for computing hyperbolic systems with geometrical source terms having concentrations, J. Comput. Phys. 219 (2006) 322–390.
[30] Y. Xing, 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.
[31] Y. Xing, C.-W. Shu, High order well-balanced finite difference WENO schemes for a class of hyperbolic systems with source terms, J. Sci. Comput. 27 (2006) 477–494.
[32] Y. Xing, 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.
[33] Y. Xing, 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, Comm. Comput. Phys. 1 (2006) 100–134.