High-order Well-balanced Schemes

Sebastian Noelle, Yulong Xing, Chi-Wang Shu

Contents

Abstract

In this paper we review some recent work on high-order well-balanced schemes for hyperbolic systems of balance laws. A characteristic feature of such systems is the existence of non-trivial steady state solutions, where the effects of convective fluxes and source terms cancel each other. Well-balanced schemes satisfy a discrete analogue of this balance and are therefore able to maintain a steady state. We discuss two classes of schemes, one based on high-order accurate, non-oscillatory finite difference operators which are well-balanced for a general class of steady states, and the other one based on well-balanced quadratures, which can - in principle - be applied to all steady states. Hyperbolic systems of balance laws have a wide application, exemplified by shallow water equations (SWE) which have steady states *at rest*, where the flow velocity vanishes, and also the more challenging *moving flow* steady states. Numerical experiments show excellent resolution of unperturbed as well as slightly perturbed steady states.

Keywords: shallow water equations, fundamental steady states, high-order upwind finite volume schemes, well-balanced schemes

1. Introduction

In many applications we encounter hyperbolic balance laws, which in one dimension are in the form

$$U_t + f(U,x)_x = s(U,x) \tag{1.1}$$

where U is the solution vector, $f(U,x)$ is the flux and $s(U,x)$ is the source term. The source term may come from geometrical, reactive or other considerations. Examples of hyperbolic balance laws include the shallow water equation with a non-flat bottom topology, elastic wave equation [2], chemosensitive movement [16] and nozzle flow [14].

Comparing with the standard hyperbolic conservation laws, namely (1.1) with $s(U,x) = 0$, the numerical approximation to the balance laws (1.1) is usually not too much more difficult: we simply need to put the point values

(for finite difference schemes) or the cell averages (for finite volume schemes) of the source term $s(U,x)$ directly into the discretization of the spatial operator. There is, however, one noticeable exception. The balance law (1.1) often admits steady state solutions in which the source term $s(U,x)$ is exactly balanced by the flux gradient $f(U,x)_x$. Such steady state solutions are usually non-trivial (they are usually not polynomial functions of the spacial variable x) and they often carry important physical meaning (for example, the still water or steady moving water solution of the shallow water equation, to be studied in more detail later in this paper). The objective of well-balanced schemes is to preserve exactly some of these steady state solutions. The most important advantage of well-balanced schemes is hat they can accurately resolve small perturbations to such steady state solutions with relatively coarse meshes. In comparison, a non-well-balanced scheme will introduce truncation errors to the steady state solution, hence it cannot resolve small perturbations to such steady states unless the truncation error is already smaller than such perturbations, thus requiring a refined mesh. In Section 5 we will provide such examples. However, it is quite difficult to design well-balanced schemes which are high-order accurate and non-oscillatory in the presence of discontinuities in the solution.

In this paper we use the shallow water equation as a prototype to survey a few recently developed well-balanced high-order finite difference, finite volume and discontinuous Galerkin finite element methods. We attempt to explain the main ingredients in these algorithms which allow us to achieve the well-balanced property without losing other nice properties of the original scheme, such as high-order accuracy and non-oscillatory performance in the presence of solution discontinuities.

The paper is organized as follows. In Section 2 we first discuss a number of interesting steady states. Then we introduce the residual which need to be well-balanced near stationary states.

At this point the paper splits into two approaches: The first approach, see Section 3, applies to finite difference, finite volume and discontinuous Galerkin schemes. It treats steady states for which the source term can be decomposed into sums of products of the form (3.4). The challenge is to construct finite

difference operators which are high-order accurate and non-oscillatory for the conservative flux difference and the source term, and which coincide for both terms in the case of steady state solutions.

The second approach, designed for general steady states and finite volume schemes, is covered in Section 4. The key task is to find well-balanced quadratures for the integral of the residual, see equation (4.1). Subsection 4.2 presents a general framework to decompose this integral into suitable parts. Subsection 4.3 realizes this approach for moving water steady states for shallow water flows.

In Section 5 we present numerical results showing the accuracy and well-balanced properties of both classes of schemes for a number of challenging flows. Section 6 contains some concluding remarks.

It is perhaps surprising that the two approaches outlined in Sections 3 and 4 require such different techniques. Indeed, the reader might skip either section on a first reading, and then proceed to the numerical experiments in Section 5.

On the other hand, we hope that the presentation of both approaches in a single paper will provide a clear understanding that well-balancing requires a detailed study of the truncation error for each individual scheme (since the truncation error should disappear for discrete steady states). The broad set of ideas and techniques presented in this paper might be helpful to the reader developing his/her own version of high-order well-balancing in a new situation.

2. Preliminaries: steady states and the residual

In this section we introduce *equilibrium variables* which characterize smooth steady states, and discuss the *residual* which monitors the deviation of the system from stationary states. In particular, two forms of the residual are singled out which are the bases of the finite difference algorithms in Section 3 on one hand and the finite volume algorithm in Section 4 on the other hand.

We will generally refer to a time-independent solution of the hyperbolic balance law as a *steady state*. When we refer to pointwise or cell-wise local transformations, we may use the terms *equilibrium*-transvormation, -variable, -reconstruction, -limiting and so forth.

2.1. Steady states

Let us again consider the system of balance laws (1.1). For example, for the shallow water equations

$$U = (h, m)^T,\ f(U) = (m, m^2/h + gh^2/2)^T,\ s(U, x) = (0, -ghb_x(x))^T, \tag{2.1}$$

where h is the water height, m is the momentum (discharge in hydraulics), $b(x)$ is the prescribed bottom topography above a given reference height, and g is the gravitational acceleration.

Many such systems can be rewritten in the form

$$V_t + c(V, x)V_x = 0 \tag{2.2}$$

for some variable

$$V = V(U, x), \tag{2.3}$$

which we would like to call the *equilibrium variables*, since constant V implies a stationary state.

Note that constant V does not imply that U is constant, since V depends also on x through the variable function b. Therefore, one should expect non-trivial steady states.

For shallow water, the equilibrium variables are $V(U, x) = (m, E)$, where the equilibrium energy E is given by

$$E(U, b) = \frac{m^2}{2h^2} + g(h + b). \tag{2.4}$$

In the following we describe various classes $\mathcal{E}$ of stationary states.

Example 2.1 - (1) The class of all steady states, $\mathcal{E}_{tot}$.

(2) Smooth steady states $\mathcal{E}_{smooth}$.

(3) Conservation laws: Here $s(U, b) \equiv 0$ and $f(U) \equiv const.$ Stationary states include

- $\mathcal{E}_0$ = constant states.
- $\mathcal{E}_1$ = two constant states separated by a stationary shock or contact.

- $\mathcal{E}_2$ = gas dynamics with zero velocity, constant pressure, and any bounded measurable function for the density.

(4) Steady states for 1D scalar balance laws.

(5) 1D shallow water equations:

- The lake at rest $\mathcal{E}_{LaR}$, where $m \equiv 0$ and hence $E = g(h+b) \equiv const.$
- Smooth river flows $\mathcal{E}_{river}$, where m is nonzero.
- Waterfalls $\mathcal{E}_{waterfall}$ (discontinuous river flows)

(6) Separable source terms studied in [36].

(7) Geostrophic jets $\mathcal{E}_{jet}$ for 2D shallow water, where $(u,v) \equiv (u(y),0)$, $g(h+b)_y = fu$ and f is the coriolis force in the upper hemisphere.

(8) Multi-layer shallow water: Oceans at rest and moving oceans.

Remark 2.1 - (1) There are many more classes of steady states, especially in 2D.

(2) It is important to note that most well-balanced schemes are designed to preserve only a certain subclass of steady states exactly. Other steady states may be preserved approximately within a certain order of accuracy.

Section 3 treats steady states for which the source terms are separable in the sense of (3.4). This includes the lake at rest as a prototype. The main tool is the construction of a well-balanced class of finite difference operators. In Section 4 we outline a well-balanced finite volume approach. While the framework in Subsection 4.1 covers in principle all steady states, we carry out the specific steps for moving water flows in Subsection 4.3.

2.2. The residual

Let us again consider the system of balance laws (1.1). We are particularly interested in solutions close to steady states, where $U_t = 0$. Therefore, we introduce the residual

$$R := -f(U)_x + s(U,x). \tag{2.5}$$

Note that

$$U_t = R, \tag{2.6}$$

and the solution U deviates from steady state if and only if $R \neq 0$.

In Section 3 we will study a class of separable steady states satisfying (3.4). This assumption implies that

$$R = (-f(U) + t(U,x))_x \tag{2.7}$$

for stationary solutions, where $t(U,x)$ is determined by $s(U,x)$. Using this structure, we construct high-order accurate well-balanced finite difference operators.

In Section 4 we focus on finite volume schemes and hence consider cell averages $\overline{R}_i$ of the residual. Well-balanced quadratures are constructed for the regular and singular parts of these integrals.

3. Schemes based on well-balanced finite difference operators

In this section, we focus on a class of steady states for which the source term is separable in the sense of Assumption 3.2. We develop well-balanced high-order accurate finite difference operators for the residual. Based on these difference operators, we derive well-balanced finite difference, finite volume and discontinuous Galerkin schemes. The steady states under consideration include the lake at rest for the shallow water equations.

The one-dimensional hyperbolic system of conservation laws with source terms under consideration is given by (1.1). We start the discussion by presenting the well balanced finite difference scheme. The extension to finite volume and DG schemes is shown in the following subsections. Only one-dimensional balance law (1.1) is investigated in this section, although the generalization to the multi-dimensional case

$$U_t + f(U,x,y)_x + g(U,x,y)_y = s(U,x,y) \tag{3.1}$$

can be done in some situations. For example, we can easily generalize the proposed technique to the two-dimensional shallow water equations with lake at rest steady state.

3.1. Finite difference scheme

We first consider the case that (1.1) is a scalar balance law. The case of systems will be explored later. We are interested in preserving exactly certain steady state solutions U of (1.1):

$$f(U)_x = s(U,x). \tag{3.2}$$

We make two assumptions on the equation (1.1) and the steady state solution U of (3.2) that we are interested to preserve exactly.

Assumption 3.1. *The steady state solution U of (3.2) that we are interested to preserve satisfies*

$$V(U,x) = constant \tag{3.3}$$

for a known function $V(U,x)$.

Note that in [34, 35, 36] the equilibrium variables have been denoted by $a(U,x)$ instead of $V(U,x)$.

Assumption 3.2. *The source term $s(U,x)$ in (1.1) can be decomposed as*

$$s(U,x) = \sum_i s_i(V(U,x))\, t_i'(x) \tag{3.4}$$

for some functions s_i and t_i.

We will design a numerical scheme which can preserve exactly the steady state solutions U which satisfy Assumption 3.1, for a balance law (1.1) with a source term satisfying Assumption 3.2. We remark here that the shallow water system with a lake at rest steady state satisfies these assumptions, and will comment on this later in this subsection. The key idea to achieve a well-balanced scheme, is to decompose the source term as in Assumption 3.2 and to first design a linear scheme with an identical numerical approximation operator for the flux derivative and the derivatives in the decomposed source terms, when applied to the steady state solution that we would like to balance.

We define a linear finite difference operator D to be one satisfying $D(af_1 + bf_2) = aD(f_1) + bD(f_2)$ for constants a, b and arbitrary grid functions f_1 and

f_2. A scheme for (1.1) with a source term given by (3.4) is said to be a linear scheme if all the spatial derivatives are approximated by linear finite difference operators. Such a linear scheme would have a truncation error

$$D_0(f(U)) - \sum_i s_i(V(U,x))\, D_i(t_i(x)),$$

where D_i are linear finite difference operators used to approximate the spatial derivatives. We further restrict our attention to linear schemes which satisfy

$$D_0 = D_1 = \cdots = D \tag{3.5}$$

for the steady state solution. Notice that we only require that the finite difference operators become identical for the steady state solution that we are interested to preserve, for general solutions these finite difference operators can be different. For such linear schemes we have

Proposition 3.1. *For the balance law (1.1) with its source term given by (3.4), linear schemes with (3.5) for the steady state solutions satisfying (3.3) can preserve these steady state solutions exactly.*

The proof of this result is rather straightforward and can be found in [35].

We now already have high-order well-balanced schemes for the balance laws under consideration. However, these schemes are linear, hence they will be oscillatory when the solution contains discontinuities. We would need to consider nonlinear schemes, namely schemes which are nonlinear even if the flux $f(U)$ and the source $s(U,x)$ in (1.1) are both linear functions of U, for example, high-order finite difference WENO schemes [3, 17, 21]. Next, we will use the fifth order finite difference WENO scheme as an example to demonstrate the basic ideas. We will not give the details of the base WENO schemes, and refer to [17, 30] for such details.

To present the basic ideas, we first consider the situation when the WENO scheme is used without a flux splitting (e.g. the WENO-Roe scheme as described in [17]). We notice that the WENO approximation to d_x where $d = f(U)$ can be eventually written out as

$$d_x|_{x=x_j} \approx \sum_{k=-r}^{r} a_k d_{k+j} \equiv D_d(d)_j \tag{3.6}$$

where $r = 3$ for the fifth order WENO approximation and the coefficients a_k depend nonlinearly on the smoothness indicators involving the grid function d. The key idea now is to use the finite difference operator D_d with $d = f(U)$ *fixed*, namely to use the same coefficients a_k obtained through the smoothness indicator of d, and apply it to approximate $t_i'(x)$ in the source terms (3.4). Thus

$$t_i'(x_j) \approx \sum_{k=-r}^{r} a_k \, t_i(x_{k+j}) = D_d \left(t_i(x)\right)_j .$$

Clearly, the finite difference operator D_d, obtained from the high-order WENO procedure and when $d = f(U)$ is fixed, is a high order accurate *linear* approximation to the first derivative for any grid function. Therefore the result of Proposition 3.1 is still valid and we conclude that the high-order finite difference WENO scheme as stated above, without the flux splitting, and with the special handling of the source terms described above, maintains exactly the steady state.

Now, we consider WENO schemes with a Lax-Friedrichs flux splitting, such as the WENO-LF and WENO-LLF schemes described in [17]. Here the flux $f(U)$ is written as a sum of $f^+(U)$ and $f^-(U)$, defined by

$$f^{\pm}(U) = \frac{1}{2}\left[f(U) \pm \alpha U\right] \tag{3.7}$$

where $\alpha = \max_U \left|\frac{\partial f(U)}{\partial U}\right|$ with the maximum being taken over either a local region (WENO-LLF) or a global region (WENO-LF), see [17, 30] for more details. We now make a modification to this flux splitting, by replacing $\pm\alpha U$ in (3.7) with $\pm\alpha \, \mathrm{sign}\left(\frac{\partial V(U,x)}{\partial U}\right) V(U,x)$. We would need to assume here that $\frac{\partial V(U,x)}{\partial U}$ does not change sign. The constant α should be suitably adjusted by the size of $\frac{\partial V(U,x)}{\partial U}$ in order to maintain enough artificial viscosity. The term $V(U,x)$ can also be replaced by $p(V(U,x))$ for any function p, whose choice should be such that $p(V(U,x))$ is as close to U as possible in order to emulate the original LF flux splitting with $\pm\alpha U$. This modification does not affect accuracy, which relies only on the fact $f(U) = f^+(U) + f^-(U)$. For the steady state solution satisfying (3.3), the artificial viscosity term $\pm\alpha \, \mathrm{sign}\left(\frac{\partial V(U,x)}{\partial U}\right) V(U,x)$ (or $\pm\alpha \, \mathrm{sign}\left(\frac{\partial p(V(U,x))}{\partial U}\right) p(V(U,x))$) in the Lax-Friedrichs flux splitting becomes a constant, and by the consistency of the WENO approximation, the

effect of these viscosity terms towards the approximation of $f(U)_x$ is zero. The flux splitting WENO approximation in this situation becomes simply $f^{\pm}(U) = \frac{1}{2}f(U)$, hence the steady state solution is preserved as before, if we simply split the derivatives in the source term as:

$$t_i'(x) = \frac{1}{2}t_i'(x) + \frac{1}{2}t_i'(x), \tag{3.8}$$

and apply the same flux splitting WENO procedure to approximate them with the nonlinear coefficients a_k coming from the WENO approximations to $f^{\pm}(U)$ respectively. This will guarantee (3.5). We thus obtain

Proposition 3.2. *The WENO-Roe, WENO-LF and WENO-LLF schemes as implemented above are exact for steady state solutions satisfying (3.3) and can maintain the original high-order accuracy.*

We now discuss the system case. The framework described for the scalar case can be applied to systems provided that we have certain knowledge about the steady state solutions to be preserved in the form of (3.3). Typically, for a system with m equations, V is a vector, and we would have m relationships in the form of (3.3):

$$V_1(U, x) = constant, \qquad \cdots \qquad V_m(U, x) = constant \tag{3.9}$$

for the steady state solutions that we would like to preserve exactly. We would then still aim for decomposing each component of the source term in the form of (3.4), where s_i could be arbitrary functions of $V_1(U, x), \cdots, V_m(U, x)$, and the functions s_i and t_i could be different for different components of the source vector. The remaining procedure is then the same as that for the scalar case and we again obtain well balanced high-order WENO schemes. We should also mention that local characteristic decomposition is typically used in high-order WENO schemes in order to obtain better non-oscillatory property for strong discontinuities. When computing the numerical flux at $x_{i+\frac{1}{2}}$, the local characteristic matrix R, consisting of the right eigenvectors of the Jacobian at $U_{i+\frac{1}{2}}$, is a constant matrix for fixed i. Hence this characteristic decomposition procedure does not alter the argument presented above for the scalar case. We refer to [34] for more details.

The shallow water equations (1.1)–(2.1) take the form

$$\begin{cases} h_t + (hu)_x = 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2} g h^2 \right)_x = -ghb_x, \end{cases} \tag{3.10}$$

The lake at rest solution satisfies (3.9) in the form

$$V_1 \equiv hu = 0, \qquad V_2 \equiv h + b = constant, . \tag{3.11}$$

The first component of the source term is 0. A decomposition of the second component of the source term in the form of (3.4) is

$$-ghb_x = -g\,(h+b)\,b_x + \frac{1}{2} g\ \left(b^2\right)_x \tag{3.12}$$

i.e. $s_1 = s_1(V_2) = -g\,(h+b)$, $s_2 = \frac{1}{2}g$, $t_1(x) = b(x)$, and $t_2(x) = b^2(x)$, which satisfies Assumption 3.2. Hence, the technique designed above can be used to obtain high-order well-balanced finite difference scheme for the shallow water equations with lake at rest solution (3.11). Two dimensional version of the shallow water equations can also be handled by the same technique [34, 36], and are not shown here. Some numerical results will be shown in Section 5 to demonstrate the good properties of these well-balanced high-order finite difference schemes.

3.2. Finite volume scheme

Following the idea of obtaining well-balanced schemes by decomposing the source terms, as shown in Section 3.1, we generalize finite volume WENO schemes to obtain high-order well-balanced schemes. The crucial difference between the finite volume and the finite difference WENO schemes is that the WENO reconstruction procedure for a finite volume scheme applies to the solution and not to the flux function values. As a consequence, finite volume schemes are more suitable for computations in complex geometry and for using adaptive meshes. The details of the finite volume WENO schemes can be found in [17, 27, 30]. However, because of a different computational framework, the maintenance of the well-balanced property requires different technical approaches.

The main idea in the previous subsection to design a well-balanced high-order finite difference WENO scheme is to decompose the source term into a sum of several terms, each of which is discretized independently using a finite difference formula consistent with that of approximating the flux derivative terms in the conservation law. We follow a similar idea here and decompose the integral of the source term into a sum of several terms, then compute each of them in a way consistent with that of computing the corresponding flux terms. We first consider the case that (1.1) is a scalar balance law. The case of systems will be explored later.

Similarly, we make some assumptions on the equation (1.1) and the steady state solution U of (3.2) that we are interested to preserve exactly:

Assumption 3.3. *The steady state solution U of (3.2) that we are interested to preserve satisfies*

$$V(U,x) \equiv \frac{U+p(x)}{q(x)} = constant \tag{3.13}$$

for some known functions $p(x)$ and $q(x)$.

Assumption 3.4. *The source term $s(U,x)$ in (1.1) can be decomposed as*

$$s(U,x) = \sum_j s_j(V(U,x))\, t_j'(x) \tag{3.14}$$

for some known functions s_j and t_j.

Note that Assumption 3.3 given here is more restrictive than that in Section 3.1, due to the additional difficulties related to the finite volume formulation.

We consider the semi-discrete formulation of the balance law

$$\frac{d}{dt}\bar{U}_i(t) = -\frac{1}{\triangle x_i}(f(U(x_{i+\frac{1}{2}}),t) - f(U(x_{i-\frac{1}{2}}),t)) + \frac{1}{\triangle x_i}\int_{I_i} s(U,x)dx. \tag{3.15}$$

The time discretization is usually performed by the classical high order Runge-Kutta method. Before stating our numerical scheme, we first present the procedure to reconstruct the pointwise values by the WENO reconstruction procedure, and then decompose the integral of the source term into several terms, with the objective of keeping the exact balance property without reducing the

high-order accuracy of the scheme. The scheme is then finally introduced with a minor change on the flux term, compared with the original WENO scheme.

The first step in building the algorithm is to reconstruct $U^{\pm}_{i+\frac{1}{2}}$ from the given cell averages $\bar{U}_i$ by the WENO reconstruction procedure, which are high order accurate approximations to the exact value $U(x_{i+\frac{1}{2}})$. It can be eventually written out as

$$U^{+}_{i+\frac{1}{2}} = \sum_{k=-r+1}^{r} w_k \bar{U}_{i+k} \equiv S^{+}_{\bar{U}}(\bar{U})_i, \quad U^{-}_{i+\frac{1}{2}} = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{U}_{i+k} \equiv S^{-}_{\bar{U}}(\bar{U})_i. \tag{3.16}$$

where $r = 3$ for the fifth order WENO approximation and the coefficients w_k and $\tilde{w}_k$ depend nonlinearly on the smoothness indicators involving the cell average $\bar{U}$. Here we obtain a linear operator $S^{\pm}_{\bar{U}}(v)$ (linear in v) which is obtained from a WENO reconstruction with fixed coefficients w_k calculated from the cell averages $\bar{U}$. A key idea here is to use the *linear* operators $S^{\pm}_{\bar{U}}(v)$ and apply them to reconstruct the functions $\bar{p}_i$ and $\bar{q}_i$. Thus

$$\begin{aligned} p^{+}_{i+\frac{1}{2}} &= S^{+}_{\bar{U}}(\bar{p})_i = \sum_{k=-r+1}^{r} w_k \bar{p}_{i+k}, \qquad p^{-}_{i+\frac{1}{2}} = S^{-}_{\bar{U}}(\bar{p})_i = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{p}_{i+k} \\ q^{+}_{i+\frac{1}{2}} &= S^{+}_{\bar{U}}(\bar{q})_i = \sum_{k=-r+1}^{r} w_k \bar{q}_{i+k}, \qquad q^{-}_{i+\frac{1}{2}} = S^{-}_{\bar{U}}(\bar{q})_i = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{q}_{i+k}. \end{aligned} \tag{3.17}$$

With the reconstructed values $p^{\pm}_{i+\frac{1}{2}}$ and $q^{\pm}_{i+\frac{1}{2}}$, we obtain the pointwise value of $V(U,x)$ by $V(U,x)^{\pm}_{i+\frac{1}{2}} = \frac{U^{\pm}_{i+\frac{1}{2}} + p^{\pm}_{i+\frac{1}{2}}}{q^{\pm}_{i+\frac{1}{2}}}$. Clearly, $p^{\pm}_{i+\frac{1}{2}}$ and $q^{\pm}_{i+\frac{1}{2}}$ are high-order accurate pointwise approximation to the function of $p(x)$ and $q(x)$ at the cell boundary $x_{i+\frac{1}{2}}$. Hence, $V(U,x)^{\pm}_{i+\frac{1}{2}}$ is a high-order approximation to $V(U(x_{i+\frac{1}{2}}), x_{i+\frac{1}{2}})$.

Now assume that U is the steady state solution satisfying (3.3), namely

$$V(U,x) = c \qquad \Leftrightarrow \qquad U + p(x) = c\, q(x)$$

for some constant c. If the cell averages $\bar{U}_i$, $\bar{p}_i$ and $\bar{q}_i$ are computed in the same fashion (e.g. all computed exactly, or all computed with the same numerical quadrature) from U, $p(x)$ and $q(x)$, then we clearly also have

$$\bar{U}_i + \bar{p}_i = c\, \bar{q}_i$$

for the same constant c. Since the reconstructed values $U^{\pm}_{i+\frac{1}{2}}$, $p^{\pm}_{i+\frac{1}{2}}$ and $q^{\pm}_{i+\frac{1}{2}}$ are computed from the cell averages $\bar{U}_j$, $\bar{p}_j$ and $\bar{q}_j$ with the *same* linear operators $S^{\pm}_{\bar{u}}(v)$, we clearly have

$$U^{\pm}_{i+\frac{1}{2}} + p^{\pm}_{i+\frac{1}{2}} = c\, q^{\pm}_{i+\frac{1}{2}}$$

for the same constant c, that is,

$$V(U,x)^{\pm}_{i+\frac{1}{2}} = c \tag{3.18}$$

for the same constant c. This is an important fact to design the well-balanced schemes.

Clearly, for a steady state solution U satisfying Assumptions 3.3 and 3.4,

$$\begin{aligned}\frac{d}{dx}\left(f(U) - \sum_j s_j(V(U,x))\, t_j(x)\right) &= f(U)_x - \sum_j s_j(V(U,x))\, t_j'(x)\\ &= f(U)_x - s(U,x) = 0.\end{aligned}$$

Therefore, $f(U) - \sum_j s_j(V(U,x))\, t_j(x)$ is a constant. We would need to choose suitably $(t_j)^{\pm}_{i+\frac{1}{2}}$, which should be high-order approximations to $t_j(x_{i+\frac{1}{2}})$ such that

$$f(U^{\pm}_{i+\frac{1}{2}}) - \sum_j s_j(V(U,x)^{\pm}_{i+\frac{1}{2}})\, (t_j)^{\pm}_{i+\frac{1}{2}} = constant \tag{3.19}$$

for a steady state solution U satisfying Assumptions 3.3 and 3.4. We will specify the choices of $(t_j)^{\pm}_{i+\frac{1}{2}}$ for the shallow water equations at the end of this subsection.

Finally, we need to decompose the integral of the source term in the follow-

ing way in order to obtain a well-balanced scheme

$$
\begin{aligned}
&\int_{I_i} s(U,x)dx = \sum_j \int_{I_i} s_j(V(U,x))t_j'(x)dx \\
= &\sum_j \left(\frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)\int_{I_i} t_j'(x)dx\right. \\
&\left. + \int_{I_i}\left(s_j(V(U,x)) - \frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)\right)t_j'(x)dx\right) \\
= &\sum_j \left(\frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)(t_j(x_{i+\frac{1}{2}}) - t_j(x_{i-\frac{1}{2}}))\right. \\
&\left. + \int_{I_i}\left(s_j(V(U,x)) - \frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)\right)t_j'(x)dx\right). \qquad (3.20)
\end{aligned}
$$

The purpose of this decomposition is to ensure that the integral of the source term equals the first term at the right hand side of (3.20) when $V(U,x) = const$, as the last term disappears in this case.

Now we are ready to describe the final form of the algorithm

$$
\frac{d}{dt}\bar{U}_i(t) = -\frac{1}{\triangle x_i}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}) + \frac{1}{\triangle x_i}\hat{s}_i, \qquad (3.21)
$$

with

(3.22)

$$
\hat{s}_i = \sum_j \left(\frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)\left((\hat{t}_j)_{i+\frac{1}{2}} - (\hat{t}_j)_{i-\frac{1}{2}}\right) + s_{i,j}\right)
$$

where $(\hat{t}_j)_{i+\frac{1}{2}}$ is a high-order approximation to $t_j(x_{i+\frac{1}{2}})$, whose definition will be described below, and $s_{i,j}$ is any high-order approximation to the integral

$$
\int_{I_i}\left(s_j(V(U,x)) - \frac{1}{2}\left(s_j(V(U,x)^+_{i-\frac{1}{2}}) + s_j(V(U,x)^-_{i+\frac{1}{2}})\right)\right)t_j'(x)\,dx. \qquad (3.23)
$$

The numerical flux $\hat{f}_{i+\frac{1}{2}}$ is defined by a monotone flux such as the Lax-Friedrichs flux

$$
F(U^-_{i+\frac{1}{2}}, U^+_{i+\frac{1}{2}}) = \frac{1}{2}\left[f(U^-_{i+\frac{1}{2}}) + f(U^+_{i+\frac{1}{2}}) - \alpha(U^+_{i+\frac{1}{2}} - U^-_{i+\frac{1}{2}})\right]. \qquad (3.24)
$$

We need to make a modification to this flux, by replacing $\alpha(U^+_{i+\frac{1}{2}} - U^-_{i+\frac{1}{2}})$ in (3.24) with $\alpha\,\mathrm{sign}(q(x))(V(U,x)^+_{i+\frac{1}{2}} - V(U,x)^-_{i+\frac{1}{2}})$. The numerical flux now becomes

(3.25)
$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2}\left[f(U^-_{i+\frac{1}{2}}) + f(U^+_{i+\frac{1}{2}}) - \alpha\,\mathrm{sign}(q(x))(V(U,x)^+_{i+\frac{1}{2}} - V(U,x)^-_{i+\frac{1}{2}})\right].$$

We would need to assume here that $q(x)$ in (3.3) does not change sign. The constant α should be suitably adjusted by the size of $\frac{1}{q(x)}$ in order to maintain enough artificial viscosity. This modification does not affect accuracy. For the steady state solution (3.13),

$$\alpha\,\mathrm{sign}(q(x))(V(U,x)^+_{i+\frac{1}{2}} - V(U,x)^-_{i+\frac{1}{2}}) = 0$$

because of (3.18). Hence, the effect of these viscosity terms becomes zero and the numerical flux turns out to be in a simple form

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2}\left[f(U^-_{i+\frac{1}{2}}) + f(U^+_{i+\frac{1}{2}})\right]. \tag{3.26}$$

Following this, we treat the approximation $(\hat{t}_j)_{i+\frac{1}{2}}$ in (3.22) in a similar way:

$$(\hat{t}_j)_{i+\frac{1}{2}} = \frac{1}{2}\left[(t_j)^-_{i+\frac{1}{2}} + (t_j)^+_{i+\frac{1}{2}}\right] \tag{3.27}$$

where, as mentioned before, $(t_j)^\pm_{i+\frac{1}{2}}$ are high order approximations to $t_j(x_{i+\frac{1}{2}})$ satisfying (3.19). Note that we implement (3.27) for the general case, not only for the steady solution. There is no viscosity term in the source term, compared with the numerical flux (3.25).

For the remaining source term $s_{i,j}$, we simply use a suitable high-order Gauss quadrature to evaluate the integral. The approximation of the values at those Gauss points are obtained by the WENO reconstruction procedure. It is easy to observe that high order accuracy is guaranteed for our scheme, and even if discontinuities exist in the solution, the non-oscillatory property is maintained.

Proposition 3.3. *The WENO-LF schemes as implemented above with (3.21), (3.22), (3.25) and (3.27) are exact for steady state solutions satisfying (3.13) and can maintain the original high-order accuracy for general solutions.*

The proof of this result is rather straightforward and can be found in [36]. The extension to the system case follows the same idea as that for the well-balanced finite difference schemes.

For the shallow water equations (3.10) with a lake at rest steady state solution (3.11), we take the same decomposition of the second component of the source term as in (3.12). We apply the WENO reconstruction to the function $(b(x),0)^T$, with coefficients computed from $(h,hu)^T$, to obtain $b^{\pm}_{i+\frac{1}{2}}$, and define

$$(t_1)^{\pm}_{i+\frac{1}{2}} = b^{\pm}_{i+\frac{1}{2}}, \qquad (t_2)^{\pm}_{i+\frac{1}{2}} = \left(b^{\pm}_{i+\frac{1}{2}}\right)^2.$$

Under these definitions and if the steady state $h+b=c$, $u=0$ for some constant c is reached, we have

$$\begin{aligned}
& f(U^-_{i+\frac{1}{2}}) - \sum_j s_j \left(V(U,x)^-_{i+\frac{1}{2}}\right)(t_j)^-_{i+\frac{1}{2}} \\
= \;& \frac{1}{2}g\left(h^-_{i+\frac{1}{2}}\right)^2 - \frac{1}{2}g\left(b^-_{i+\frac{1}{2}}\right)^2 + g\frac{1}{2}\left(h^-_{i+\frac{1}{2}} + b^-_{i+\frac{1}{2}} + h^+_{i-\frac{1}{2}} + b^+_{i-\frac{1}{2}}\right) b^-_{i+\frac{1}{2}} \\
= \;& \frac{1}{2}g\left(h^-_{i+\frac{1}{2}} + b^-_{i+\frac{1}{2}}\right)\left(h^-_{i+\frac{1}{2}} - b^-_{i+\frac{1}{2}}\right) + g\,c\,b^-_{i+\frac{1}{2}} \\
= \;& \frac{1}{2}g\,c\left(h^-_{i+\frac{1}{2}} - b^-_{i+\frac{1}{2}} + 2b^-_{i+\frac{1}{2}}\right) = \frac{1}{2}g\,c^2,
\end{aligned}$$

which is a constant. A similar manipulation leads to

$$f(U^+_{i+\frac{1}{2}}) - \sum_j s_j \left(V(U,x)^+_{i+\frac{1}{2}}\right)(t_j)^+_{i+\frac{1}{2}} = \frac{1}{2}g\,c^2.$$

Hence the high-order finite volume WENO schemes can be designed following the above idea for the shallow water equations.

3.3. Extension to discontinuous Galerkin scheme

We have successfully designed high-order well-balanced finite difference and finite volume WENO well-balanced scheme for a class of hyperbolic balance laws. In this subsection, we consider the generalization of these ideas to the Runge-Kutta discontinuous Galerkin (RKDG) methods. Well-balanced high-order RKDG schemes will be designed for a class of conservation laws satisfying Assumptions 3.3 and 3.4. The basic idea is the same as that for the finite volume

schemes, such as the technique of decomposing the source term and replacing the viscosity term in the numerical fluxes, because the RKDG methods can be considered as a generalization of finite volume schemes, even though they do not require a reconstruction and evolve the complete polynomial in each cell forward in time. The RKDG methods are therefore easier to use for multi-dimensional problems in complex geometry, than the finite volume schemes, as the complicated reconstruction procedure can be avoided. We refer to [8, 9, 10, 11, 12] for more details of RKDG methods.

The semi-discrete DG schemes for (1.1) take the form

$$\begin{aligned} &\textstyle\int_{I_j} \partial_t U_h(x,t) v_h(x) dx - \int_{I_j} f(U_h(x,t)) \partial_x v_h(x) dx + \hat{f}_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) \\ &\qquad -\hat{f}_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) = \textstyle\int_{I_j} s(U_h(x,t),t) v_h(x) dx \end{aligned} \tag{3.28}$$

$$\int_{I_j} U_h(x,0) v_h(x) dx = \int_{I_j} U_0(x) v_h(x) dx. \tag{3.29}$$

First, we define a high-order approximation $V_h(U_h, x) = \frac{U_h + p_h}{q_h}$ to $V(U_h, x)$, where p_h and q_h are L^2 projections of p and q into V_h, see (3.29) for such a projection. Now assume that U is the steady state solution satisfying (3.3), namely

$$U(x) + p(x) = c\, q(x)$$

for some constant c, and U_h is the L^2 projection of this steady state solution. Clearly, since the L^2 projection is a linear operator,

$$U_h(x) + p_h(x) = c\, q_h(x)$$

for the same constant c at every point x. This implies

$$V_h(U_h, x) = \frac{U_h(x) + p_h(x)}{q_h(x)} = c.$$

For such steady state solution U satisfying Assumptions 3.3 and 3.4, we have

$$\frac{d}{dx}\left(f(U) - \sum_j s_j(V(U,x))\, t_j(x) \right) = 0.$$

We would need to suitably choose a function $(t_j)_h$, which should be a high-order approximation to t_j and should satisfy the condition

$$f(U_h(x)) - \sum_j s_j(V_h(U_h(x),x))(t_j)_h(x) = constant \tag{3.30}$$

for all x. The construction of $(t_j)_h$ will be shown for the shallow water equations in the end of this subsection.

Similar to the decomposition of the source term in the well balanced finite volume schemes (3.20), we decompose the integral of the source term on the right hand side of (3.28) as:

$$\begin{aligned}
&\int_{I_i} s(U_h,x)v_h dx \\
=\ &\sum_j \left(\frac{1}{2}\left(s_j(V(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V(U_h,x)^-_{i+\frac{1}{2}})\right)\int_{I_i} t_j'(x)v_h dx \right. \\
&\left. + \int_{I_i}\left(s_j(V(U_h,x)) - \frac{1}{2}\left(s_j(V(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V(U_h,x)^-_{i+\frac{1}{2}})\right)\right) t_j'(x)v_h dx\right) \\
=\ &\sum_j \left(\frac{1}{2}\left(s_j(V(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V(U_h,x)^-_{i+\frac{1}{2}})\right)\right. \\
&\quad \cdot\left(t_j(x_{i+\frac{1}{2}})v_h(x^-_{i+\frac{1}{2}}) - t_j(x_{i-\frac{1}{2}})v_h(x^+_{i-\frac{1}{2}}) - \int_{I_i} t_j(x)v_h'(x)dx\right) \\
&\left. + \int_{I_i}\left(s_j(V(U_h,x)) - \frac{1}{2}\left(s_j(V(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V(U_h,x)^-_{i+\frac{1}{2}})\right)\right) t_j'(x)v_h dx\right).
\end{aligned}$$

We then replace this source term with a high-order approximation of it given by

$$\begin{aligned}
&\sum_j \left(\frac{1}{2}\left(s_j(V_h(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V_h(U_h,x)^-_{i+\frac{1}{2}})\right)\right. \\
&\quad \cdot\left((\hat{t}_j)_{h,i+\frac{1}{2}}v_h(x^-_{i+\frac{1}{2}}) - (\hat{t}_j)_{h,i-\frac{1}{2}}v_h(x^+_{i-\frac{1}{2}}) - \int_{I_i} (t_j)_h(x)v_h'(x)dx\right) \\
&\left. + \int_{I_i}\left(s_j(V_h(U_h,x)) - \frac{1}{2}\left(s_j(V_h(U_h,x)^+_{i-\frac{1}{2}}) + s_j(V_h(U_h,x)^-_{i+\frac{1}{2}})\right)\right) t_j'(x)v_h dx\right)
\end{aligned}$$

where $(\hat{t}_j)_{h,i+\frac{1}{2}}$ is a high-order approximation to $t_j(x_{i+\frac{1}{2}})$, whose definition

follows (3.25) and (3.27) from Section 3.2

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2}\left[f((U_h)^-_{i+\frac{1}{2}}) + f((U_h)^+_{i+\frac{1}{2}}) - \alpha\,\mathrm{sign}(q(x))(V_h(U_h,x)^+_{i+\frac{1}{2}} - V_h(U_h,x)^-_{i+\frac{1}{2}})\right],$$

$$(\hat{t}_j)_{h,i+\frac{1}{2}} = \frac{1}{2}\left[(t_j)_h(x^-_{i+\frac{1}{2}}) + (t_j)_h(x^+_{i+\frac{1}{2}})\right].$$

Usually, we perform the limiter on the function U_h after each Runge-Kutta stage. Now, our purpose is to maintain the steady state solution U which satisfies $V(U,x) = constant$. The above limiter procedure could destroy the preservation of such steady state, since if the limiter is enacted, the resulting modified solution U_h may no longer satisfy $V_h(U_h,x) = constant$. We therefore propose to first check whether any limiting is needed based on the function $V_h(U_h,x)$ in each Runge-Kutta stage, where the cell averages of $V_h(U_h,x)$ (needed to implement the TVB limiter) are computed by a suitable Gauss quadrature. If a certain cell is flagged by this procedure needing limiting, then the actual limiter is implemented on U_h, not on $V_h(U_h,x)$. When the limiting procedure is implemented this way, if the steady state U satisfying $V(U,x) = constant$ is reached, no cell will be flagged as requiring limiting since $V_h(U_h,x)$ is equal to the same constant, hence U_h will not be limited and therefore the steady state is preserved.

This finishes the description of the RKDG schemes. We can clearly observe that the accuracy is maintained (see Table 7 in Section 5). We also state below the proposition claiming the exact preservation of the steady state solution (3.3). The proof is straightforward and is therefore omitted.

Proposition 3.4. *The RKDG schemes as stated above are exact for steady state solutions satisfying (3.13) and can maintain the original high-order accuracy for general solutions.*

The extension of the well-balanced high-order RKDG schemes to the system case follows the same idea as that for the well-balanced finite volume schemes.

For the shallow water equations (3.10) with a lake at rest steady state solution (3.11), we can easily verify that the definitions of $(t_i)_h$,

$$(t_1)_h(x) = b_h(x), \qquad (t_2)_h(x) = (b_h(x))^2$$

where $b_h(x)$ is the L^2 projection of $b(x)$ to the finite element space V_h, lead to

$$f(U_h) - \sum_j s_j(V_h(U_h, x))(t_j)_h = \frac{1}{2} g\, c^2$$

when the steady state $h + b = c$, $u = 0$ is reached, satisfying our requirement.

A new approach to obtain well balanced methods, by calculating the source term exactly at the steady state, was introduced for RKDG methods to save computational cost in [37]. The traditional RKDG methods are shown to be capable of maintaining certain steady states exactly, if a small modification on either the initial condition or the flux is provided. We refer the interesting reader to [37]

4. Schemes based on well-balanced quadrature

In this section, we follow [24] and develop an alternative approach to well-balancing. It is based on well-balanced quadrature rules for cell averages of the residual defined in (4.1) and leads to a high-order accurate finite volume scheme which is well-balanced for moving water steady states.

The section is organized as follows: Based on the residual of the balance law introduced in Subsection 2.2, we define (in Subsection 4.1) a general class of semidiscrete finite volume schemes and give a definition for such schemes to be well-balanced for a steady state $\overline{V}$. For each building block of these schemes - the piecewise smooth reconstruction of the data, the quadrature of the regular part of the residual in the interior of the cells, and the reconstruction of the singular part of the residual at the cell interfaces - we define a notion of well-balancing (Subsection 4.2). Theorem 4.1 states that these conditions guarantee that the overall scheme is well-balanced. In Subsection 4.3 we realize this general program using equilibrium reconstructions. In particular, we treat the 1D shallow water equations as a prototype and consider the lake at rest, river flows and waterfalls aligned with the grid. Except for discontinuities which are not aligned with the grid, we can therefore balance general 1D steady state solutions for the shallow water equations. Many of the techniques presented here can be adapted to other classes of balance laws. In Subsections $4.4 - 4.5$ we discuss some interesting aspects of related schemes.

4.1. Framework of the finite volume discretization

Here we consider a general balance law in the form (2.6). Let

$$\overline{U}(x_i,t) \approx \frac{1}{\triangle x_i}\int_{I_i} U(x,t)\,dx$$
$$\overline{R}_i(t) \approx \frac{1}{\triangle x_i}\int_{I_i} R(x,t)\,dx \tag{4.1}$$

be approximate cell averages of the solution and the residual. Then we consider semidiscrete schemes of the form

$$\frac{d}{dt}\overline{U}_i(t) = \overline{R}_i \quad \text{for } i = 1,\dots,N. \tag{4.2}$$

Definition 4.1. *The scheme* (4.2) *is* well-balanced for a steady state $\overline{V}$ *if*

$$\overline{R}_i = 0 \quad \text{for } i = 1,\dots,N \tag{4.3}$$

whenever the original data are in steady state, i.e.

$$V(U,x) \equiv \overline{V} = \textit{constant}. \tag{4.4}$$

Remark 4.1 - (1) Such schemes have also been called *exactly well-balanced* in the literature, in order to distinguish them from *approximately well-balanced* schemes, for which

$$\overline{R}_i = \mathcal{O}(\triangle x^p) \tag{4.5}$$

for steady state data (4.4), where p should be higher than the order of consistency of the overall scheme.

(2) Since the solution U and also the topography b may be discontinuous, we consider R to be a bounded Borel measure over Ω, i.e. $R \in \mathcal{M}(\Omega)$. In general, R has both regular and singular parts with respect to Lebesgues' measure, and therefore it is not straightforward to give meaning to the integral in (4.1), or to define a consistent quadrature for this integral. However, these difficulties cannot and should not be avoided, and we believe that discussing them directly in terms of measures makes the presentation of several recent well-balanced finite volume schemes most transparent. This point of view is closely related to the

work on non-conservative products of measures in [13, 26]

(3) We begin with schemes which are semidiscrete in time. Later, we will use Runge-Kutta time discretizations as in [15, 29, 31] to derive fully discrete schemes.

As is well known from conservation laws, the difficulty in discretizing (4.2) arises from discontinuities in the solution. If the flux function f is nonlinear, the solution U will develop shocks in finite time. For stationary shocks $f(U)$ is continuous due to the Rankine-Hugoniot condition, so $U_t = 0$ at the shock. If the shock is unsteady, then $f(U)_x$ and hence U_t become a Dirac measure, and $U(x,t)$ jumps as the shock passes by.

For balance laws, it is desirable to treat also discontinuities in the data b, which may be given either by the problem itself or by the discretization.

In general, discontinuities in the flux $f(U)$ or the data b will lead to singular parts in the measure R. The term $f(U)_x$ can be treated classically via the theory of weak solutions of conservation laws. Singularities in the source term are less well understood.

4.2. Regular and singular parts of the residual

In order to evaluate the integral on the RHS of (4.1), we split R into its regular and singular parts with respect to Lebesgue measure dx,

$$R = R_{reg} + R_{sing}. \tag{4.6}$$

Analogously, we split the cell averages of the residual via

$$\overline{R}_i = \overline{R}^i_{reg} + \overline{R}^i_{sing}. \tag{4.7}$$

We assume that the singular parts of the residual are concentrated at the cell interfaces, and decompose $\overline{R}^i_{sing}$ into

$$\overline{R}^i_{sing} = \overline{R}^{i-1/2+}_{sing} + \overline{R}^{i+1/2-}_{sing}, \tag{4.8}$$

so

$$\overline{R}_i = \overline{R}^i_{reg} + \overline{R}^{i-1/2+}_{sing} + \overline{R}^{i+1/2-}_{sing} \tag{4.9}$$

and

$$\frac{d}{dt}\overline{U}_i(t) = \overline{R}^i_{reg} + \overline{R}^{i-1/2+}_{sing} + \overline{R}^{i+1/2-}_{sing}. \tag{4.10}$$

For the rest of Section 4, we give an overview how to treat the regular and the singular components of the residual on the RHS of (4.10). In Theorem 4.1, we give general sufficient conditions which guarantee that the scheme (4.10) is well-balanced for a steady state $\overline{V}$. In Subsection 4.3, we discuss a number of schemes and steady states for which these conditions are satisfied.

4.2.1. The regular part of the residual

Suppose that

$$\tilde{U}(x) \approx U(x) \tag{4.11}$$

$$\tilde{b}(x) \approx b(x) \tag{4.12}$$

are piecewise smooth reconstructions of the cell averages $\bar{U}_i, \bar{b}_i$ over the cells I_i. Let $x_i^{(1)} \dots x_i^{(p)}$ be quadrature points within cell I_i, to be used in the quadrature (4.15) below. In Subsection 4.3 we will develop reconstructions with the following property:

Definition 4.2. *Suppose that the original data (U, b) are in steady state, i.e. (4.4) holds for some steady state $\overline{V}$. Suppose furthermore that $(\bar{U}_i, \bar{b}_i)$ are the cell averages of the data (U, b). Then the reconstruction $(\tilde{U}, \tilde{b})$ of (U, b) is* well-balanced for the steady state $\overline{V}$ and the quadrature points $x_i^{(1)} \dots x_i^{(p)} \in I_i$ *if*

$$\tilde{V}(x_i^{(j)}) := V(\tilde{U}(x_i^{(j)}), \tilde{b}(x_i^{(j)})) \equiv \overline{V} \quad \text{for} \quad j = 1 \dots p. \tag{4.13}$$

In analogy to (2.5), let

$$\tilde{R} := R(\tilde{U}, \tilde{b}) = -f(\tilde{U})_x + s(\tilde{U}, \tilde{b}) \tag{4.14}$$

be the approximate residual. Let

$$K(\tilde{R}; I_i) := \sum_{j=1}^{p} \omega_j \tilde{R}(x_i^{(j)}) \;\approx\; \frac{1}{\triangle x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{R}(x)dx \tag{4.15}$$

be a quadrature of the approximate residual $\tilde{R}$ over the interior of cell I_i and let

$$\overline{R}^i_{reg} := K(\tilde{R}; I_i). \tag{4.16}$$

We will study quadratures with the following property:

Definition 4.3. *The quadrature* (4.16) *is* well-balanced for the steady state $\overline{V}$ *if*

$$\overline{R}^i_{reg} = 0 \tag{4.17}$$

for all $(\tilde{U}, \tilde{b})$ *which satisfy* (4.13).

4.2.2. The singular part of the residual

We now turn to the singular part of the residual. Let us focus upon an interface $x_{i+1/2}$. For infinitesimal small ε, we introduce a boundary layer $(x_{i+1/2} - \varepsilon, x_{i+1/2} + \varepsilon)$. Within this layer, we construct bounded continuous functions $\hat{U}_\varepsilon(y)$ and $\hat{b}_\varepsilon(y)$, where $y = x - x_{i+1/2}$. The boundary values are

$$\hat{U}_\varepsilon(\pm\varepsilon) = \tilde{U}(x_{i+1/2}\pm) = \tilde{U}^\pm_{i+1/2} \tag{4.18}$$

$$\hat{b}_\varepsilon(\pm\varepsilon) = \tilde{b}(x_{i+1/2}\pm) = \tilde{b}^\pm_{i+1/2}, \tag{4.19}$$

where $\tilde{U}$ and $\tilde{b}$ are the piecewise smooth reconstructions from (4.11), (4.12). Now we define the singular parts of the residual on the RHS of (4.10) via

$$\overline{R}^{i+1/2-}_{sing} := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{-\varepsilon}^{0} \hat{R}_\varepsilon(y) dy \tag{4.20}$$

$$\overline{R}^{i+1/2+}_{sing} := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{0}^{\varepsilon} \hat{R}_\varepsilon(y) dy, \tag{4.21}$$

where

$$\hat{R}_\varepsilon(y) := R(\hat{U}_\varepsilon(y), \hat{b}_\varepsilon(y)). \tag{4.22}$$

Definition 4.4. *The approximation* (4.20)–(4.21) *of the singular parts of the residual is* well-balanced for the steady state $\overline{V}$ *if*

$$\tilde{V}_{i+1/2-} = \tilde{V}_{i+1/2+} = \overline{V} \tag{4.23}$$

implies

$$\overline{R}_{sing}^{i+1/2-} = \overline{R}_{sing}^{i+1/2+} = 0. \tag{4.24}$$

4.2.3. The general well-balancing theorem

So far we have introduced a notion of well-balancing for each building-block of the semi-discrete finite volume scheme. Combining them we can immediately established the following theorem:

Theorem 4.1. *Consider the scheme*

$$\frac{d}{dt}\overline{U}_i(t) = \overline{R}_{reg}^{i} + \overline{R}_{sing}^{i-1/2+} + \overline{R}_{sing}^{i+1/2-}, \tag{4.25}$$

where $\overline{R}_{reg}^{i}$ *is given by* (4.16), $\overline{R}_{sing}^{i+1/2-}$ *by* (4.20) *and* $\overline{R}_{sing}^{i-1/2+}$ *by* (4.21) *with* i *replaced by* $(i-1)$. *Suppose that for a constant steady state* $\overline{V}$, *the reconstruction* $(\tilde{U}, \tilde{b})$ *in* (4.11), (4.12), *the quadrature* (4.16) *and the approximate singular residua are well-balanced according to Definitions 4.2, 4.3, and 4.4. Then the scheme* (4.25) *is well-balanced for the steady state* $\overline{V}$ *in the sense of Definition 4.1, i.e.*

$$\frac{d}{dt}\overline{U}_i(t) \equiv 0. \tag{4.26}$$

This finishes our general discussion of well-balanced schemes. In the next subsection, we will construct several schemes which fall into the framework outlined in Theorem 4.1, among them the recent second order scheme of Audusse et al. [1] and the high-order schemes of Castro, Pares et al. and the authors [5, 23, 24].

4.3. Realization via equilibrium reconstructions

In this section we will show that some recent schemes fall into the framework outlined in the previous section. In Subsections 4.3.1–4.3.3, we will therefore

verify the well-balancing properties for the reconstruction, the quadrature, and the singular layer as introduced in Definitions 4.2, 4.3 and 4.4. Once this has been done, Theorem 4.1 implies that the overall schemes are well-balanced according to Definition 4.1.

4.3.1. Smooth reconstruction in the cell interior

Hydrostatic reconstructions in the cell interior

We begin with a so-called *hydrostatic reconstruction* that preserves the discharge m and the waterlevel $\eta = h + b$. Here we follow Audusse et al. [1]. They begin the reconstruction process by reconstructing the discharge m, the water level η and the bottom b. Then in [1, (2.8)], they define the reconstructed height as

$$\tilde{h}(x) := \tilde{\eta}(x) - \tilde{b}(x). \tag{4.27}$$

Therefore, if the discharge and the water level are constant to begin with, they will remain constant during the reconstruction. In particular, the lake at rest ($m = 0$, $\eta = const$) is preserved throughout the reconstruction. Let us mention in passing that h as defined in (4.27) is later on truncated by Audusse et al. in order to guarantee positivity of the water height. This is done in such a way that the well-balancing relation (4.27) is preserved (see [1, (2.9), (2.13)]). Therefore, the hydrostatic reconstruction is well-balanced for the lake at rest according to Definition 4.2.

Equilibrium reconstruction in the cell interior

In [24, Sect.3.2] the authors devised a reconstruction which preserves all one-dimensional steady states for the shallow water equation. While we refer to that paper for the details, we would like to give the key idea in a nutshell. Given cell averages $(\overline{U}_i)$ and a bottom function $b(x)$, we choose local reference values $\overline{V}_i$ of the equilibrium variables. These are defined implicitly by the requirement that

$$\frac{1}{\triangle x_i} \int_{I_i} U(\overline{V}_i, x) dx = \overline{U}_i. \tag{4.28}$$

Let us pause for a moment and discuss this relation carefully: $U(V, x)$ is the inverse of $V(U, x)$, i.e. $U(V(U, x), x) = U$. Relation (4.28) chooses $\overline{V}_i$ as the

unique (see the paragraphs preceding [24, Def.3.2]) local equilibrium such that the corresponding conserved variables $U(\overline{V}_i, b(x))$ have the same cell average $\overline{U}_i$ as the numerical data. It is proven in [24, Def.3.2] that, if the data $U(x)$ and $b(x)$ are in local equilibrium ($V(U(x), x) \equiv \overline{V}$ for all cells I_i), then the reference equilibrium states $\overline{V}_i$ computed via (4.28) coincide with the true local steady state $\overline{V}$.

The reconstruction is completed by limiting the reconstruction $\tilde{V}(x)$ with respect to the reference values $\overline{V}_i$ (see [24, (3.18)]). The argument that our reconstruction is well-balanced for *all* steady states is now straightforward: if the data are globally in equilibrium (i.e. (4.4) holds for a global steady state $\overline{V}$), then $\overline{V}_i = \overline{V}$ for all cells, and the equilibrium-limiter [24, (3.18)] enforces that $\tilde{V}(x) \equiv \overline{V}$.

A well-balanced reconstruction due to Castro, Pares et al.
In [5] Castro, Gallardo, Lopez and Pares start by computing $\overline{V}_i$ as in (4.28). This gives the low order accurate equilibrium reconstruction

$$\tilde{U}^*(x) := U(\overline{V}_i, b(x))$$

which is only based upon the values within the i^{th} cell. Let us keep i and hence $\tilde{U}^*$ fixed. To find the high-order correction Castro et al. compute a reconstruction polynomial

$$Q_i(x) = p(x|(I_j, \overline{U}_j - \bar{\tilde{U}}_j^*), j = i-k, \ldots, i+k)$$

which interpolates the differences of the cell averages $\overline{U}_j$ and the cell averages of the low order reconstruction $\tilde{U}^*$. Note that $\overline{U}_j$ only coincides with $\bar{\tilde{U}}_j^*$ if $j = i$. Finally Castro et al. reconstruct U by

$$\tilde{U}_i(x) := \tilde{U}_i^*(x) + Q_i(x).$$

It is proven in [5] that this reconstruction is high-order accurate, and well-balanced if $\overline{V}_i = \overline{V}$ for any i.

4.3.2. Well-balanced quadrature in the cell interior

In this section we start from the smooth, well-balanced reconstructions $\tilde{U}$, $\tilde{b}$ which we constructed in the previous section and derive well-balanced interior

quadratures $K(\tilde{R}, I_i)$ (cf. (4.15)). In all of this section, we restrict ourselves to the shallow water equations.

For conciseness, we use the following notation: suppose that

$$a, b : \Omega \to \mathbb{R}$$

are real-valued functions defined on our spatial domain. For a fixed cell I_i, let

$$\tilde{a}, \tilde{b} : I_i \to \mathbb{R}$$

be smooth reconstructions over the interior of the cell. Then we denote the difference and mean operators by

$$D\tilde{a} := \tilde{a}^-_{i+1/2} - \tilde{a}^+_{i-1/2}, \qquad \bar{a} := (\tilde{a}^-_{i+1/2} + \tilde{a}^+_{i-1/2})/2 \tag{4.29}$$

For later use, we observe the discrete product rule of differencing

$$D(\tilde{a}\tilde{b}) = \bar{a}\, D\tilde{b} + D\tilde{a}\, \bar{b}. \tag{4.30}$$

Quadrature for the lake at rest

We begin with a widely used quadrature which is well-balanced for the lake at rest ($m \equiv 0$, $h + b \equiv \bar{\eta}$). For any smooth $\tilde{U}$, $\tilde{b}$

$$\begin{aligned} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{R}(x)dx &= \int_{x_{i-1/2}}^{x_{i+1/2}} (-f_2(\tilde{U})_x - g\tilde{h}\tilde{b}_x)(x)dx \\ &= -Df_2(\tilde{U}) - g \int_{x_{i-1/2}}^{x_{i+1/2}} (\tilde{h}\tilde{b}_x)(x)dx. \end{aligned} \tag{4.31}$$

Therefore we need to define a quadrature for the integral of the source term. We will use the two nodes $x_i^{(1)} = x_{i-1/2}, x_i^{(2)} = x_{i+1/2}$ and approximate both $\tilde{b}$ and $\tilde{h}$ by linear functions. This gives

$$\int_{x_{i-1/2}}^{x_{i+1/2}} (\tilde{h}\tilde{b}_x)(x)dx = \frac{\tilde{h}_{i-1/2} + \tilde{h}_{i+1/2}}{2} \, (\tilde{b}_{i+1/2} - \tilde{b}_{i-1/2}) = \bar{h}\, D\tilde{b} \tag{4.32}$$

Inserting (4.31) and (4.32) into (4.15) we obtain the quadrature

$$K(\tilde{R}, I_i) := \frac{1}{\triangle x_i} \left(-Df_2(\tilde{U}) - g\bar{h}D\tilde{b} \right). \tag{4.33}$$

A simple calculation shows that this is well-balanced for the lake at rest: If

$$\tilde{m}_{i-1/2} = \tilde{m}_{i+1/2} = 0, \quad \tilde{b}_{i-1/2} + \tilde{h}_{i-1/2} = \tilde{b}_{i+1/2} + \tilde{h}_{i+1/2} = \bar{\eta}, \tag{4.34}$$

then $f_2(\tilde{U}_{i\pm 1/2}) = \frac{g}{2}(\tilde{h}_{i\pm 1/2})^2$, so

$$-Df_2(\tilde{U}) = -\frac{g}{2} D\left(\tilde{h}^2\right) = -g\ \bar{h}D\tilde{h} = -g\ \bar{h}D(\bar{\eta} - \tilde{b}) = g\ \bar{h}D\tilde{b}. \tag{4.35}$$

Plugging (4.35) into (4.33) we immediately obtain that

$$K(\tilde{R}, I_i) = 0 \tag{4.36}$$

for the lake at rest.

Quadrature for moving water steady states

In [24] we refined the quadrature (4.33) to include moving steady states. The key observation is that

$$\begin{aligned} Df_2(\tilde{U}) &= D(\tilde{m}\tilde{u} + g\tilde{h}^2/2) \\ &= \bar{m}D\tilde{u} + \bar{u}D\tilde{m} + g\bar{h}D\tilde{h} \\ &= \bar{m}D\tilde{u} + \bar{u}D\tilde{m} + \bar{h}D(\tilde{E} - g\tilde{b} - \tilde{u}^2/2) \\ &= \bar{u}D\tilde{m} + \bar{h}D\tilde{E} - g\bar{h}D\tilde{b} + (\bar{m} - \bar{h}\bar{u})D\tilde{u}. \end{aligned} \tag{4.37}$$

Noting that $\bar{m} - \bar{h}\bar{u} = D\tilde{h}D\tilde{u}/4$, we obtain that

$$Df_2(\tilde{U}) = \bar{u}D\tilde{m} + \bar{h}D\tilde{E} - g\bar{h}D\tilde{b} + \frac{1}{4}D\tilde{h}(D\tilde{u})^2. \tag{4.38}$$

Therefore, for a non-stationary steady state, where $D\tilde{m} = D\tilde{E} = 0$,

$$-Df_2(\tilde{U}) - g\bar{h}D\tilde{b} + \frac{1}{4}D\tilde{h}(D\tilde{u})^2 = 0. \tag{4.39}$$

As a consequence, the quadrature

$$K(\tilde{R}, I_i) := \frac{1}{\triangle x_i} \left(-Df_2(\tilde{U}) - g\bar{h}D\tilde{b} + \frac{1}{4}D\tilde{h}(D\tilde{u})^2 \right) \tag{4.40}$$

is well-balanced for general steady states in 1D. For smooth flows, the cubic correction term $\frac{1}{4}D\tilde{h}(D\tilde{u})^2$ is so small that it does not affect the order of the quadrature rule. In [24] we showed how to limit this term when the jumps $D\tilde{h}$ and $D\tilde{u}$ are no longer of the order of the gridsize.

4.3.3. Singular layers at the cell boundaries

In Subsection 4.2.2 we introduced a general framework of well-balanced singular layers. It would be convenient if the equilibrium values are constant throughout the singular layer. But this can be done only if $\tilde{V}_{i+1/2-} = \tilde{V}_{i+1/2+} = \overline{V}$. In this case we set

$$\hat{U}_\varepsilon(y) = U(\overline{V}, \hat{b}_\varepsilon(y)), \tag{4.41}$$

where $\hat{b}_\varepsilon(\cdot)$ is any smooth reconstruction of the bottom topography. From (4.22) we immediately obtain that

$$\hat{R}_\varepsilon(y) = R(\hat{U}_\varepsilon(y), \hat{b}_\varepsilon(y)) = 0. \tag{4.42}$$

In the general case, there is no straightforward construction of the residual in the singular layer. However, we will mimic the construction (4.41)–(4.42) as much as possible in suitable parts of the interval $[-\varepsilon, \varepsilon]$.

For this we turn to the schemes proposed in [1, 6, 24], which are all related as follows: The continuous piecewise linear topography is defined by the four values $y = -\varepsilon, -\varepsilon/2, \varepsilon/2, \varepsilon$, corresponding to the points $x_{i+1/2} - \varepsilon, x_{i+1/2} - \varepsilon/2, x_{i+1/2} + \varepsilon/2, x_{i+1/2} + \varepsilon$. At these points our piecewise linear $\hat{b}_\varepsilon$ takes the values

$$\hat{b}_\varepsilon(y) := \begin{cases} \tilde{b}^{\pm}_{i+1/2} & \text{for} \quad y = \pm\varepsilon \\ \hat{b}_{i+1/2} & \text{for} \quad y = \pm\varepsilon/2, \end{cases} \tag{4.43}$$

where the intermediate value near the interface $\hat{b}_{i+1/2}$ still needs to be determined. Certainly, it should be a suitable convex combinations of the values $\tilde{b}^{\pm}_{i+1/2}$ at the endpoints. This intermediate value of the topography was defined slightly differently in each of [1, 24, 6], and we will discuss this in the following subsection.

Adjacent to the interiors of the left and right neighboring cells, i.e. in the intervals $[-\varepsilon, -\varepsilon/2]$ and $[\varepsilon/2, \varepsilon]$, we keep the equilibrium values constant, and define $\hat{U}_\varepsilon$ and $\hat{R}_\varepsilon$ via (4.41)–(4.42). In [24] we have called the set $[-\varepsilon, -\varepsilon/2] \cup [\varepsilon/2, \varepsilon]$, for which the residual vanishes, the *equilibrium layer*. By construction,

the values $\hat{U}_\varepsilon(\pm\varepsilon/2)$ are

$$\hat{U}_\varepsilon(-\frac{\varepsilon}{2}) = U(\tilde{V}(x_{i+1/2-}), \hat{b}_{i+1/2}) =: \hat{U}_{i+1/2-} \tag{4.44}$$

$$\hat{U}_\varepsilon(\frac{\varepsilon}{2}) = U(\tilde{V}(x_{i+1/2+}), \hat{b}_{i+1/2}) =: \hat{U}_{i+1/2+} \,. \tag{4.45}$$

Note that the function $U(V, b)$ used in (4.44)–(4.45) depends strongly on the particular steady state under consideration. For example it differs for the lake at rest considered in [1] and the moving water treated in [24].

In the remaining interval $[-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}]$, which we called *convective layer* in [24], the topography is constant, $\hat{b}_\varepsilon(y) \equiv \hat{b}_{i+1/2}$. Therefore, as for the exact solution, where

$$R(x) = -\partial_x f(U(x)),$$

the approximate residual should reduce to a conservative flux difference. For this, we define an approximate flux $\hat{f}_\varepsilon(y)$ as follows. In the center $y = 0$, the flux will be an approximate Riemann solver $\hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+})$, and at the endpoints $y = \pm\varepsilon/2$, it takes the values $f(\hat{U}_{i+1/2\pm})$. In between, $\hat{f}_\varepsilon$ may be any continuous function, e.g. piecewise linear. Then we set

$$\hat{R}_\varepsilon(y) := -\partial_y \hat{f}_\varepsilon(y). \tag{4.46}$$

From here, we can easily evaluate the singular parts of the residual in (4.20)–(4.21) and obtain

$$\overline{R}_{sing}^{i+1/2-} = -\hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+}) + f(\hat{U}_{i+1/2-}) \tag{4.47}$$

$$\overline{R}_{sing}^{i+1/2+} = -f(\hat{U}_{i+1/2+}) + \hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+}). \tag{4.48}$$

Lemma 4.1. *The approximation* (4.47)–(4.48) *of the singular parts of the residual is* well-balanced *in the sense of Definition 4.4.*

Proof – Since the residual vanishes in the equilibrium layer, it is sufficient to show that

$$\hat{U}_{i+1/2-} = \hat{U}_{i+1/2+}, \tag{4.49}$$

since then

$$f(\hat{U}_{i+1/2-}) = \hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+}) = f(\hat{U}_{i+1/2+}) \tag{4.50}$$

and, from (4.47)–(4.48),

$$\overline{R}_{sing}^{i+1/2-} = \overline{R}_{sing}^{i+1/2+} = 0. \tag{4.51}$$

So suppose that we are in local equilibrium in the sense of Definition 4.4, i.e. $\tilde{V}_{i+1/2-} = \tilde{V}_{i+1/2+} = \overline{V}$ for some steady state $\overline{V}$. Then

$$\hat{U}_{i+1/2-} = \hat{U}_{i+1/2+} = U(\overline{V}, \hat{b}_{i+1/2}), \tag{4.52}$$

which is (4.49). □

4.4. On the choice of the intermediate bottom $\hat{b}_{i+1/2}$

Now we focus upon an interface $x_{i+1/2}$ and the two values $\tilde{b}^{\pm}_{i+1/2}$, which represent the jump of the bottom at the cell interface. We require that the intermediate value $\hat{b}_{i+1/2}$ satisfies

$$\min\{\tilde{b}^{-}_{i+1/2}, \tilde{b}^{+}_{i+1/2}\} \le \hat{b}_{i+1/2} \le \max\{\tilde{b}^{-}_{i+1/2}, \tilde{b}^{+}_{i+1/2}\}. \tag{4.53}$$

In [1, (2.9)], Audusse et al. choose

$$\hat{b}_{i+1/2} := \max\{\tilde{b}^{-}_{i+1/2}, \tilde{b}^{+}_{i+1/2}\}. \tag{4.54}$$

Together with an appropriate CFL restriction and a suitable flux function, (4.54) guarantees positivity of the waterheight. It has been used by various authors, including the present authors, and we consider it to be the standard choice.

However, there is an important case which suggests that the standard choice should sometimes be replaced. While the choice of Audusse et al. is particularly useful at the shore of a lake with subcritical velocity, we consider the rather different situation of a waterfall in steep, fast mountainous rivers (see [24] for details). In Figure 1 we show this stationary moving water flow. The water flows in supercritically from the left until it hits a steep (or even discontinuous) descent. Flowing down, the water accelerates. Due to conservation of mass the water height decreases until the flow becomes critical. It is then stopped by a stationary shock, or bore. Behind the shock, the water moves on slowly with subcritical velocity.

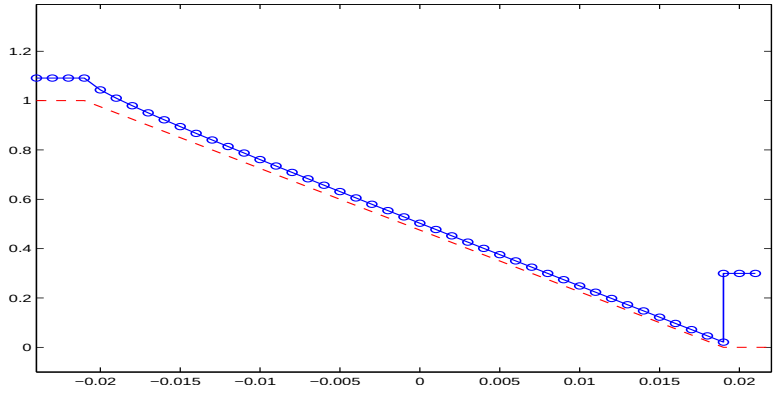

Figure 1 – Waterfall: a (finitely or infinitely) steep slide followed by a stationary shock. Dashed line: bottom topography. Circles: Water surface.

We obtain the same solution for the Riemann problem with discontinuous bottom. The interesting observation is that the hydrodynamic problem (the stationary shock) is resolved at the bottom of the topography, so we should replace (4.54) by

$$\hat{b}_{i+1/2} := \min\{\tilde{b}^-_{i+1/2}, \tilde{b}^+_{i+1/2}\}. \tag{4.55}$$

In [6], Castro, Pardo, and Parés formulate well-balanced schemes with a general choice of $\hat{b}_{i+1/2}$ satisfying only (4.53). The optimal choice of $\hat{b}_{i+1/2}$ remains an open problem, which seems to be closely related to the non-uniqueness of the Riemann problem, see [7, 18]. Meanwhile we recommend to use (4.54) since it is positivity preserving.

4.5. A note on the conservative character of the fluxes.

In discussions of recent well-balanced schemes, it was argued that the numerical fluxes in [1, (2.15)] were not conservative (see [20] for the definition of conservative numerical fluxes). We would like to clarify this point: In our notation (see (4.2)), the first order scheme of Audusse et al. reads

$$\Delta x_i \frac{d}{dt}\overline{U}_i(t) = \mathcal{F}_l(\overline{U}_i, \overline{U}_{i+1}, b_i, b_{i+1}) - \mathcal{F}_r(\overline{U}_{i-1}, \overline{U}_i, b_{i-1}, b_i) = \Delta x_i \, \overline{R}_i \tag{4.56}$$

with fluxes [1, (2.16)]

$$\mathcal{F}_r(\overline{U}_{i-1}, \overline{U}_i, b_{i-1}, b_i) = \hat{f}(\overline{U}_{i-1}, \overline{U}_i) + \begin{pmatrix} 0 \\ \frac{g}{2}h_i^2 - \frac{g}{2}h_{i-1/2+}^2 \end{pmatrix} \tag{4.57}$$

$$\mathcal{F}_l(\overline{U}_i, \overline{U}_{i+1}, b_i, b_{i+1}) = \hat{f}(\overline{U}_i, \overline{U}_{i+1}) + \begin{pmatrix} 0 \\ \frac{g}{2}h_i^2 - \frac{g}{2}h_{i+1/2-}^2 \end{pmatrix} \tag{4.58}$$

Indeed,

$$\mathcal{F}_r(\overline{U}_{i-1}, \overline{U}_i, b_{i-1}, b_i) \neq \mathcal{F}_l(\overline{U}_{i-1}, \overline{U}_i, b_{i-1}, b_i),$$

so these fluxes are not conservative in the sense of the Lax-Wendroff theorem.

Let us compare this to the scheme (4.10) derived in this section:

$$\frac{d}{dt}\overline{U}_i(t) = \overline{R}_{reg}^i + \overline{R}_{sing}^{i-1/2+} + \overline{R}_{sing}^{i+1/2-}. \tag{4.59}$$

From (4.16), (4.47) and (4.48)

$$\triangle x_i \, \overline{R}_{reg}^i = -f(\tilde{U}_{i+1/2-}) + f(\tilde{U}_{i-1/2+}) + \triangle x_i \, \bar{S}_{reg}^i \tag{4.60}$$

$$\triangle x_i \, \overline{R}_{sing}^{i+1/2-} = -\hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+}) + f(\hat{U}_{i+1/2-}) \tag{4.61}$$

$$\triangle x_i \, \overline{R}_{sing}^{i-1/2+} = -f(\hat{U}_{i-1/2+}) + \hat{f}(\hat{U}_{i-1/2-}, \hat{U}_{i-1/2+}), \tag{4.62}$$

where the regular part of the source term is given by

$$\triangle x_i \, \bar{S}_{reg}^i = -\begin{pmatrix} 0 \\ g\bar{h}_i D\tilde{b}_i \end{pmatrix}. \tag{4.63}$$

Setting

$$Df_i := \hat{f}(\hat{U}_{i+1/2-}, \hat{U}_{i+1/2+}) - \hat{f}(\hat{U}_{i-1/2-}, \hat{U}_{i-1/2+}),$$

the scheme (4.59) reads

$$\triangle x_i \frac{d}{dt}\overline{U}_i(t) = -Df_i + [f(\tilde{U}_{i-1/2+}) - f(\hat{U}_{i-1/2+})] + [f(\hat{U}_{i+1/2-}) - f(\tilde{U}_{i+1/2-})] + \triangle x_i \bar{S}^i_{reg}. \tag{4.64}$$

Following the arguments in [1, 24] or the present paper, one can check that the schemes defined by (4.64) and (4.56) coincide. However, the curious nonconservative flux differences appear also in the form (4.64). We will now show that the two flux differences in the square brackets are precisely the singular source terms in the left and right equilibrium layers, where the source term jumps.

From (4.22) and (4.42) we know that the residual vanishes in the equilibrium layer. More precisely, it consists of a non-zero flux difference and a non-zero source term which balance each other:

$$\triangle x_i \bar{R}^{i-1/2+}_{equil} = -f(\tilde{U}_{i-1/2+}) + f(\hat{U}_{i-1/2+}) + \triangle x_i \bar{S}^{i-1/2+}_{sing} = 0 \tag{4.65}$$

$$\triangle x_i \bar{R}^{i+1/2-}_{equil} = -f(\hat{U}_{i+1/2-}) + f(\tilde{U}_{i+1/2-}) + \triangle x_i \bar{S}^{i+1/2-}_{sing} = 0. \tag{4.66}$$

Therefore,

$$\triangle x_i \bar{S}^{i-1/2+}_{sing} = f(\tilde{U}_{i-1/2+}) - f(\hat{U}_{i-1/2+}) \tag{4.67}$$

$$\triangle x_i \bar{S}^{i+1/2-}_{sing} = f(\hat{U}_{i+1/2-}) - f(\tilde{U}_{i+1/2-}) \tag{4.68}$$

and the scheme (4.64) can be rewritten in the natural form

$$\frac{d}{dt}\overline{U}_i(t) = -Df_i + \triangle x_i \left(\bar{S}^{i-1/2+}_{sing} + \bar{S}^i_{reg} + \bar{S}^{i+1/2-}_{sing}\right) \tag{4.69}$$

which clearly distinguishes conservative flux differences, regular and singular source terms.

5. Numerical examples for the shallow water equations

We have successfully designed high-order well-balanced schemes by different approaches. In Section 3, high-order well-balanced finite difference, finite volume and RKDG schemes are designed for a class of hyperbolic balance laws, which include the shallow water equations with the lake at rest steady state. The key idea towards the well-balanced property is a special decomposition of the source term. Fifth order finite difference, finite volume WENO schemes and third order finite element RKDG scheme are implemented, and we denote them by FD5, FV5-D (D for well-balanced finite differencing) and RKDG3 respectively. In Subsection 4.3, high order finite volume schemes which are well-balanced for the steady river flow of the shallow water equations have been presented. The well-balanced property relies on a special equilibrium reconstruction and non-trivial quadrature of the source term, while the high-order accuracy comes from the high-order WENO reconstruction and extrapolation of the source term. Fourth order accuracy can be obtained. We denote these finite volume schemes as FV4-Q (Q for well-balanced quadrature). Note that we have two well-balanced finite volume schemes, FV5-D and FV4-Q, obtained through different approaches.

In this section we provide numerical results to demonstrate the good properties of these well-balanced schemes, when applied to the shallow water equations. The examples in Sections 5.1–5.3 show well-balancing steady states to machine accuracy, high order of accuracy for unsteady solutions, and small perturbations of steady states. The last two examples (discontinuous bottom, Section 5.4 and 2D pertubation, Section 5.5) go somewhat beyond the scope of the numerical analysis of Sections 3 and 4. They provide some preliminary insight for which applications the methods might still work, even though this may not yet be proven.

In all numerical tests, time discretization is by the classical third order TVD Runge-Kutta method [31]. For finite volume, finite difference WENO schemes, the CFL number is taken as 0.6, except for the accuracy tests where smaller time steps are taken to ensure that spatial errors dominate. For the third order RKDG scheme, the CFL number is 0.18. For the TVB limiter implemented in the RKDG scheme, the TVB constant M (see [10, 28] for its

definition) is taken as 0 in most numerical examples, unless otherwise stated. The gravitation constant g is taken as $9.812 m/s^2$ during the computation.

5.1. Well-balanced tests

The purpose of the first test problems is to verify the well balanced property of our algorithms. Note that FV4-Q is capable of capturing steady river flows, and FV5-D, FD5, RKDG3 are designed for capturing the lake at rest. Hence, two different test problems are proposed here. A fifth-order Gauss quadrature is employed to compute the initial value in the finite volume and DG approaches.

5.1.1. Lake at rest

This test is shown to verify that FV5-D, FD5 and RKDG3 indeed maintain the well-balanced property over a non-flat bottom. We choose two different functions for the bottom topography given by ($0 \le x \le 10$):

$$b(x) = 5\, e^{-\frac{2}{5}(x-5)^2}, \tag{5.1}$$

which is smooth, and

$$b(x) = \begin{cases} 4 & \text{if } 4 \le x \le 8, \\ 0 & \text{otherwise,} \end{cases} \tag{5.2}$$

which is discontinuous. The initial data is the stationary solution:

$$h + b = 10, \qquad hu = 0.$$

This steady state should be exactly preserved. We compute the solution until $t = 0.5$ using $N = 200$ uniform cells. In order to demonstrate that the well-balanced property is indeed maintained up to round-off error, we use single precision, double precision and quadruple precision to perform the computation, and show the L^1 and L^∞ errors for the water height h (note: h in this case is not a constant function!) and the discharge hu in Tables 1 and 2 for the two bottom functions (5.1) and (5.2) and different precisions. For the RKDG

Table 1 – L^1 and L^∞ errors for different precisions for the steady solution with a smooth bottom (5.1).

		L^1 error		L^∞ error	
	precision	h	hu	h	hu
FD5	single	3.13E-07	1.05E-05	9.54E-07	4.85E-05
	double	1.24E-15	2.34E-14	7.11E-15	8.65E-14
	quadruple	1.62E-33	2.11E-32	6.16E-33	8.74E-32
FV5-D	single	4.07E-06	3.75E-05	1.33E-05	1.33E-04
	double	2.50E-14	2.23E-13	7.64E-14	7.97E-13
	quadruple	3.49E-33	2.90E-32	1.39E-32	9.62E-32
RKDG3	single	6.44E-06	2.44E-05	2.57E-05	1.75E-04
	double	6.82E-15	2.90E-14	2.84E-14	2.14E-13
	quadruple	9.06E-31	3.92E-33	8.05E-29	1.12E-31

method, the errors are computed based on the numerical solutions at cell centers. 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.

We have also computed stationary solutions using initial conditions which are not the steady state solutions and letting time evolve into a steady state, obtaining similar results with the well-balanced property.

5.1.2. Steady river flow

We pick different test problems for FV4-Q, to verify the well balanced property 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 [32]. Here, our purpose is to maintain these steady state solutions exactly.

The bottom function is given by:

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12, \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

Table 2 – L^1 and L^∞ errors for different precisions for the steady solution with a nonsmooth bottom (5.2).

		L^1 error		L^∞ error	
	precision	h	hu	h	hu
FD5	single	2.28E-07	3.61E-06	1.91E-06	2.37E-05
	double	9.05E-15	5.88E-14	3.55E-15	4.46E-14
	quadruple	1.30E-33	1.40E-32	4.62E-33	5.64E-32
FV5-D	single	6.50E-06	2.61E-05	1.91E-05	1.53E-04
	double	1.73E-14	5.88E-14	4.62E-14	2.43E-13
	quadruple	2.69E-32	9.30E-32	5.85E-32	3.04E-31
RKDG3	single	5.76E-07	3.54E-07	9.54E-07	1.18E-06
	double	1.41E-15	8.90E-16	3.55E-15	2.83E-15
	quadruple	2.69E-31	1.62E-35	8.06E-29	8.18E-34

for a channel of length $25m$. 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:

$$E = \frac{1.53^2}{2 \times 0.66^2} + 9.812 \times 0.66, \qquad m = 1.53, \tag{5.4}$$

together with the boundary condition

- upstream: The discharge hu=1.53 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 Figure 2. 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 Tables 3 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.

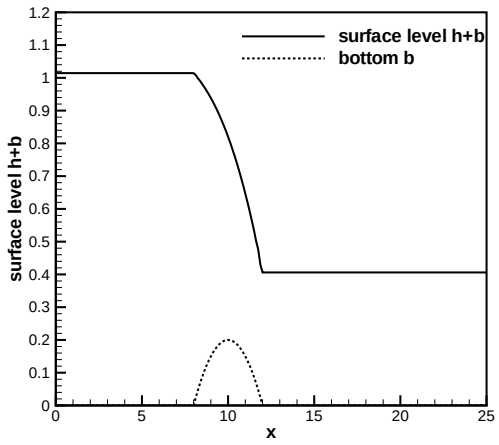

Figure 2 – The surface level $h+b$ and the bottom b for the transcritical flow without a shock.

Table 3 – $L1$ and L^∞ errors for different precisions for the transcritical flow without a shock.

		$L1$ error		L^∞ error	
	precision	h	hu	h	hu
FV4-Q	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

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

(5.5)
$$E = \begin{cases} \frac{3}{2}(9.812 \times 0.18)^{\frac{2}{3}}) + 9.812 \times 0.2 & \text{if } x \leq 11.665504281554291, \\ \frac{0.18^2}{2 \times 0.33^2} + 9.812 \times 0.33 & \text{otherwise,} \end{cases} \qquad m = 0.18,$$

together with the boundary condition

- upstream: The discharge hu=0.18 m^2/s is imposed.
- downstream: The water height h=0.33 m is imposed.

This steady state should be exactly preserved. As we mentioned in Subsection 4.3, 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. For this case when stationary shock exists, we need to use the Roe's flux to compute the approximate Riemann problem, and replace the limiter procedure by a one-sided limiter for the two cells next to the shock. Also, 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 Figure 3. 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 Tables 4 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.

Table 4 – L^1 and L^∞ errors for different precisions for the transcritical flow with a shock.

		$L1$ error		L^∞ error	
	precision	h	hu	h	hu
FV4-Q	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

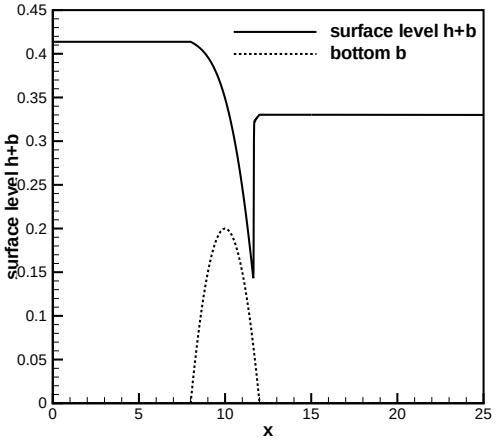

Figure 3 – The surface level $h+b$ and the bottom b for the transcritical flow with a shock.

c): Subcritical flow. The initial condition is given by:

$$E = 22.06605, \qquad m = 4.42, \tag{5.6}$$

together with the boundary condition

- upstream: The discharge hu=4.42 m^2/s is imposed.
- downstream: The water height h=2 m is imposed.

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 Figure 4. 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 Tables 5 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.

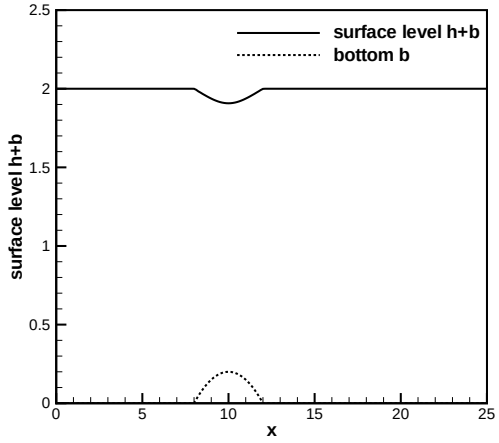

Figure 4 – The surface level $h+b$ and the bottom b for the subcritical flow.

5.2. Testing the orders of accuracy

In this example we will test the high-order accuracy of our schemes for a smooth solution. There are some known exact solutions to the shallow water equation with non-flat bottom in the literature, such as some stationary solutions, but they are not generic test cases for accuracy. We have therefore chosen to use the following bottom function and initial conditions

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

Table 5 – L^1 and L^∞ errors for different precisions for the subcritical flow.

	L^1 error		L^∞ error	
precision	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

with periodic boundary conditions, see [34]. Since the exact solution is not known explicitly for this case, we use the fifth order finite volume 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). Tables 6 and 7 contain the L^1 errors for the cell averages for FV4-Q, FV5-D and RKDG3, and for the point values for FD5, and numerical orders of accuracy. We can clearly see that the designed order of accuracy is achieved. For the RKDG scheme, the TVB constant M is taken as 32. Notice that the CFL number we have used for the finite volume scheme decreases with the mesh size and is recorded in Tables 6 and 7. For the RKDG method, the CFL number is fixed at 0.18. We note that fifth-order accuracy is observed for FV4-Q. 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 [23].

5.3. A small perturbation of a steady-state water

The following test cases are chosen to demonstrate the capability of the proposed schemes for computations on the perturbation of a steady state solution, which cannot be captured well by a non well-balanced scheme. For the same reason as in Section 5.1, two test cases are proposed for different algorithms.

5.3.1. Perturbation of a lake at rest

The following quasi-stationary test case was proposed by LeVeque [19]. It was chosen to demonstrate the capability of the proposed scheme for computations on a rapidly varying flow over a smooth bed, and the perturbation of a stationary state. We test it on FV5-D, FD5 and RKDG3 methods.

The bottom topography consists of one hump:

$$b(x) = \begin{cases} 0.25(\cos(10\pi(x-1.5))+1) & \text{if } 1.4 \le x \le 1.6, \\ 0 & \text{otherwise,} \end{cases} \tag{5.7}$$

Table 6 – L^1 errors and numerical orders of accuracy for the example in Section 5.2.

No. of cells	FV4-Q				
	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
	FV5-D				
	CFL	h		hu	
		L^1 error	order	L^1 error	order
25	0.6	1.48E-02		9.45E-02	
50	0.6	2.40E-03	2.63	1.98E-02	2.26
100	0.4	2.97E-04	3.01	2.58E-03	2.93
200	0.3	2.43E-05	3.61	2.13E-04	3.60
400	0.2	1.02E-06	4.57	8.96E-06	4.57
800	0.1	3.26E-08	4.97	2.85E-07	4.97

The initial conditions are given with

$$(5.8) \qquad (hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 1 - b(x) + \epsilon & \text{if } 1.1 \le x \le 1.2, \\ 1 - b(x) & \text{otherwise,} \end{cases}$$

where ϵ is a non-zero perturbation constant. Two cases have been run: $\epsilon = 0.2$ (big pulse) and $\epsilon = 0.001$ (small pulse). Theoretically, for small ϵ, this disturbance should split into two waves, propagating left and right at the characteristic speeds $\pm\sqrt{gh}$. Many numerical methods have difficulty with the calculations involving such small perturbations of the water surface. Both sets of initial conditions are shown in Figure 5. The solution at time t=0.2s for the big pulse $\epsilon = 0.2$, obtained on a 200 cell uniform grid with simple transmissive

Table 7 – L^1 errors and numerical orders of accuracy for the example in Section 5.2.

No. of	FD5				
	CFL	h		hu	
		L^1 error	order	L^1 error	order
25	0.6	1.70E-02		1.06E-01	
50	0.6	2.17E-03	2.97	1.95E-02	2.45
100	0.6	3.33E-04	2.71	2.83E-03	2.78
200	0.6	2.36E-05	3.82	2.04E-04	3.80
400	0.6	9.67E-07	4.61	8.38E-06	4.61
800	0.6	3.38E-08	4.84	2.94E-07	4.83
	RKDG3				
	CFL	h		hu	
		L^1 error	order	L^1 error	order
25	0.6	2.35E-03		2.12E-02	
50	0.6	1.15E-04	4.36	1.01E-03	4.39
100	0.4	1.24E-05	3.20	1.09E-04	3.21
200	0.3	1.02E-06	3.59	8.97E-06	3.60
400	0.2	1.11E-07	3.19	9.79E-07	3.19
800	0.1	1.30E-08	3.09	1.14E-07	3.08

boundary conditions, and compared with a 3000 cell solution, is shown in Figure 6 for the FD5, in Figure 7 for the FV5-D and in Figure 8 for the RKDG3. The results for the small pulse $\epsilon = 0.001$ are shown in Figures 9, 10 and 11. At this time, the downstream-traveling water pulse has already passed the bump. We can clearly see that there are no spurious numerical oscillations.

5.3.2. Perturbation of steady river flow

In subsection 5.1.2, 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], and

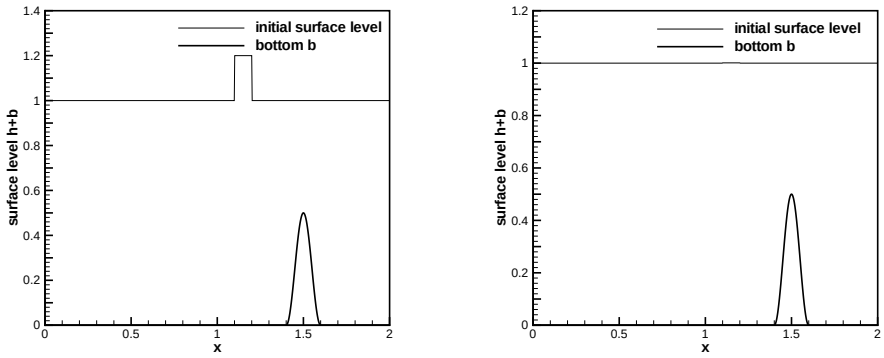

Figure 5 – The initial surface level $h + b$ and the bottom b for a small perturbation of a steady-state water. Left: a big pulse ϵ=0.2; right: a small pulse ϵ=0.001.

check whether the FV4-Q method captures it well. We remark that FV5-D, FD5 and RKDG3 are not well balanced for these steady states and they all fail to capture this perturbation on coarse meshes.

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 Figure 12 for the transcritical flow without a shock, in Figure 13 for the transcritical flow with a shock and in Figure 14 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.

5.4. The dam breaking problem over a rectangular bump

In this example we use the SW model to simulate the dam breaking problem over a rectangular bump, which involves a rapidly varying flow over a discontinuous bottom topography. This example was used in [33].

It is not yet settled whether SW models give meaningful predictions for flows

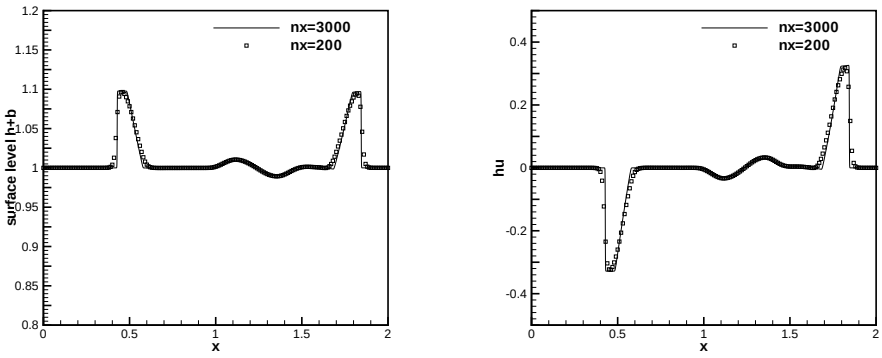

Figure 6 – FD5: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

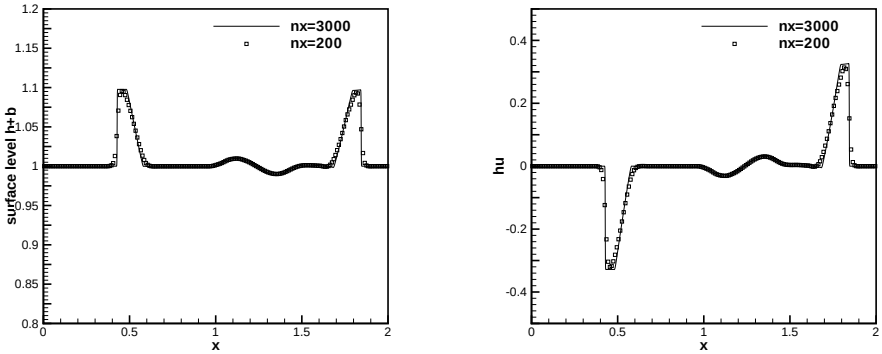

Figure 7 – FV5-D: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

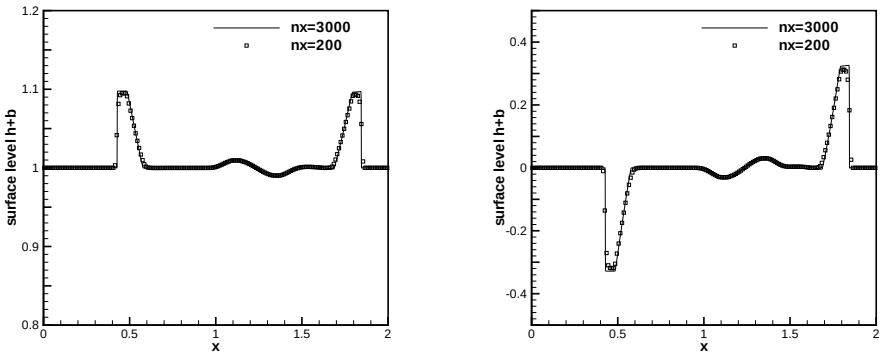

Figure 8 – RKDG3: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

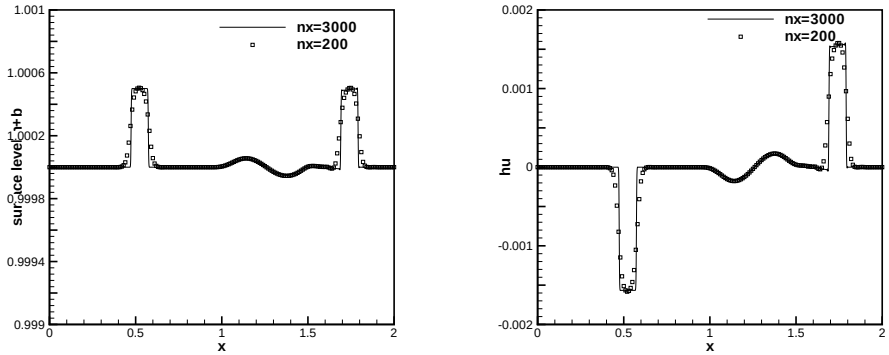

Figure 9 – FD5: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

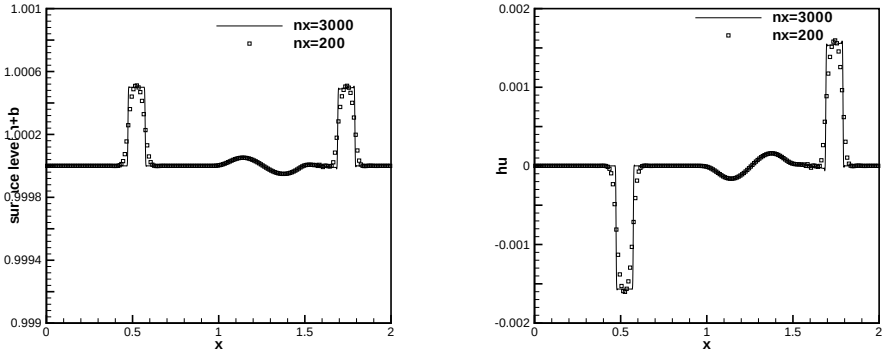

Figure 10 – FV5-D: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

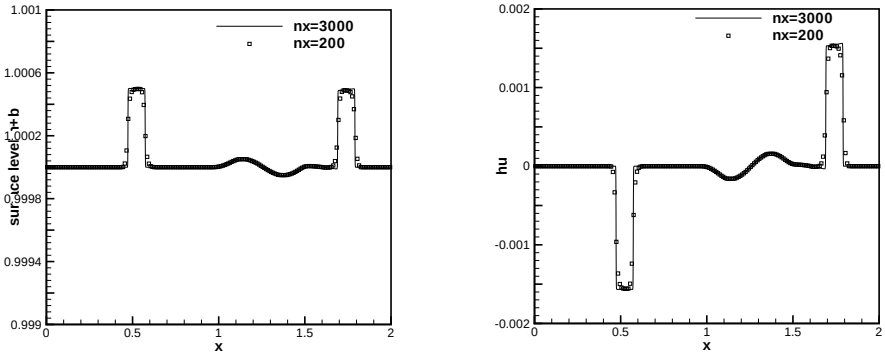

Figure 11 – RKDG3: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

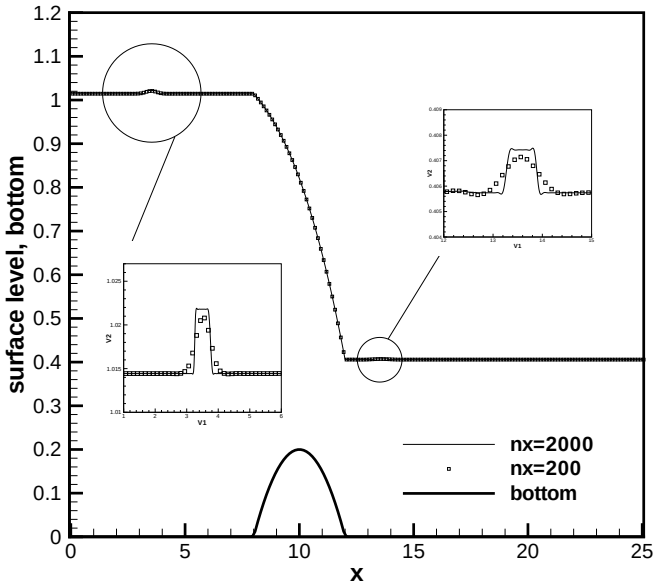

Figure 12 – FV4-Q: Small perturbation of the transcritical flow without a shock.

over discontinuous bottoms, where key modelling assumptions are violated. Should a user have to switch to the full Euler or Navier-Stokes equations, as soon as there is a step in the bottom? The present example shows that the SW model, and our algorithms, may provide stable and sharp computational results for discontinuous topography.

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leq 1500/8, \\ 0 & \text{otherwise,} \end{cases} \tag{5.9}$$

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

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 20 - b(x) & \text{if } x \leq 750, \\ 15 - b(x) & \text{otherwise.} \end{cases} \tag{5.10}$$

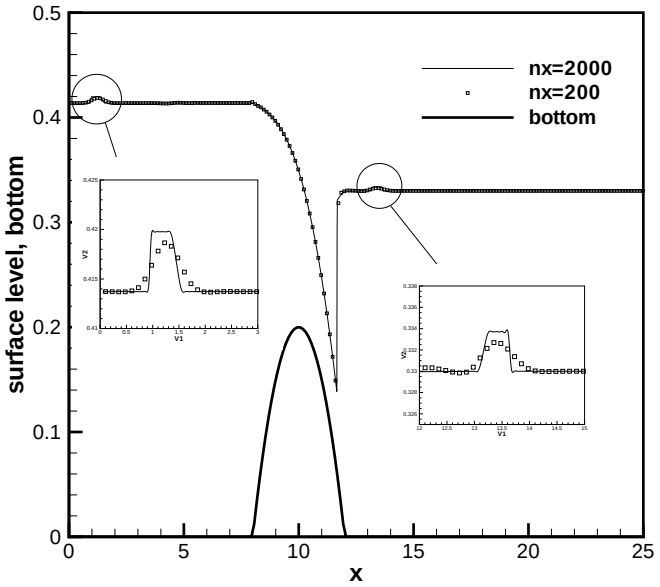

Figure 13 – FV4-Q: Small perturbation of the transcritical flow with a shock.

Figure 15 shows numerical results obtained by FD5 500 uniform cells (and a comparison with the results using 5000 uniform cells) with ending time t=60s. In this example, the water height $h(x)$ is discontinuous at the points x=562.5 and x=937.5, while the surface level $h(x) + b(x)$ is smooth there. All schemes FV4-Q, FD5, FV5-D, RKDG3 work well for this example, giving well resolved, non-oscillatory solutions using 400 cells which agree with the converged results using 4000 cells.

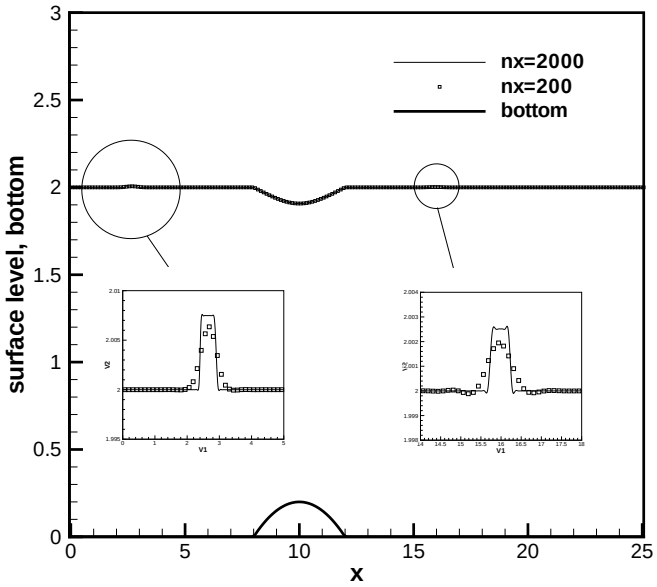

Figure 14 – FV4-Q: Small perturbation of the subcritical flow.

5.5. A two-dimensional example

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

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2} g h^2 \right)_x + (huv)_y = -ghb_x \\ (hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2} g h^2 \right)_y = -ghb_y \end{cases} \tag{5.11}$$

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. It is straightforward to generalize 1D schemes dimension by dimension to this 2D system, and usually one will maintain the 2D well-balancing of the lake at rest. It is also

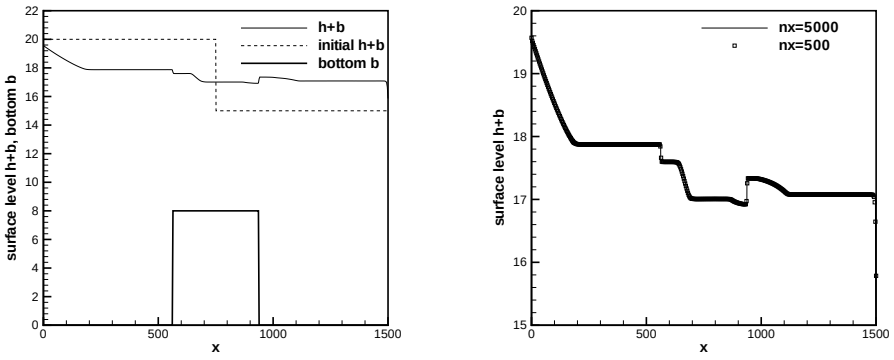

Figure 15 – FD5: The surface level $h+b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 500 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 500 and 5000 grid cells.

fairly straightforward to balance geostrophic jets, driven by the Coriolis force, which are aligned with the grid, see e.g. [4, 22, 25]. In general, however, there is an abundance of steady states, each being a solution of a mixed hyperbolic-elliptic boundary value problem in (x, y)-space. In particluar, there is no way we could well-balance general moving steady flows.

However, the 1D techniques presented in this paper are already useful for some 2D flows. Note that river and channel flows have a pronounced direction of propagation (usually close to the downhill direction). Also the topography in rivers and channels (dams, barrages) is often essentially one-dimensional. In the present example, we will apply the 1D techniques of Section 4 (i.e. well-balancing non-stationary steady flows) only in the x-direction (the direction of the underlying unperturbed flow). This will be of great advantage when computing a fully 2D perturbation of this non-stationary steady flow.

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

$$b(x,y) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12, \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

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

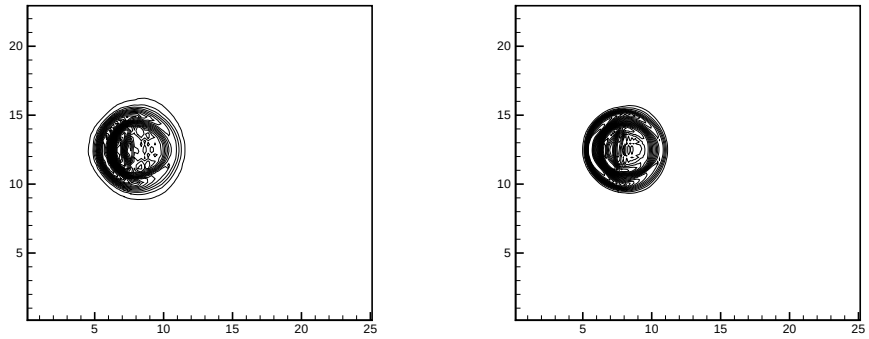

Figure 16 – The contours of the difference between the height h and the initial steady state (5.12) for the problem in Section 5.5 at time $t = 0.5$. 30 uniformly spaced contour lines from -0.009 to 0.012. Left: results with a 100×100 uniform mesh. Right: results with a 200×200 uniform mesh.

computed from:

$$\frac{1}{2}u^2 + g(h+b) = 22.06605, \qquad hu(x,y,0) = 4.42, \qquad hv(x,y,0) = 0. \tag{5.13}$$

These data correspond precisely to the one-dimensional subcritical steady state of (5.6), and the cross section of the unperturbed solution can be seen in Figure 4. 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 $6.5 \leq x \leq 7.5$, $12 \leq y \leq 13$. Figures 16 and 17 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 (5.12) is presented at different times $t = 0.5$ and $t = 1$. We also run the same numerical test with FV5-D. Note that FV5-D is not well-balanced for moving steady states. The comparison of the numerical results are presented in Figures 18 and 19. The results indicate that FV4-Q can resolve the complex small features of the flow very well, without spurious features which do appear in the results obtained with FV5-D.

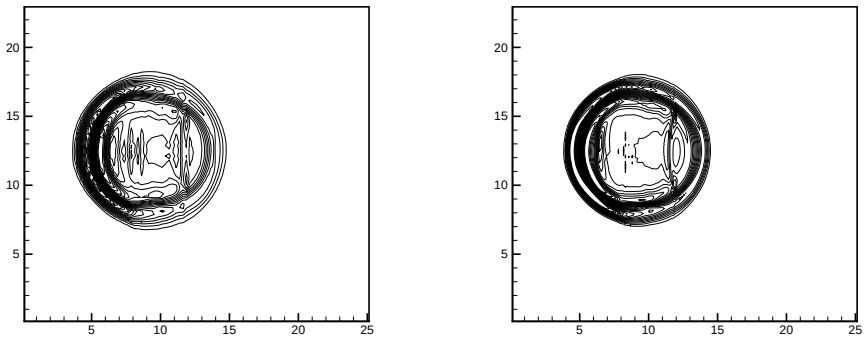

Figure 17 – The contours of the difference between the height h and the initial steady state (5.12) for the problem in Section 5.5 at time $t = 1$. 30 uniformly spaced contour lines from -0.005 to 0.008. Left: results with a 100×100 uniform mesh. Right: results with a 200×200 uniform mesh.

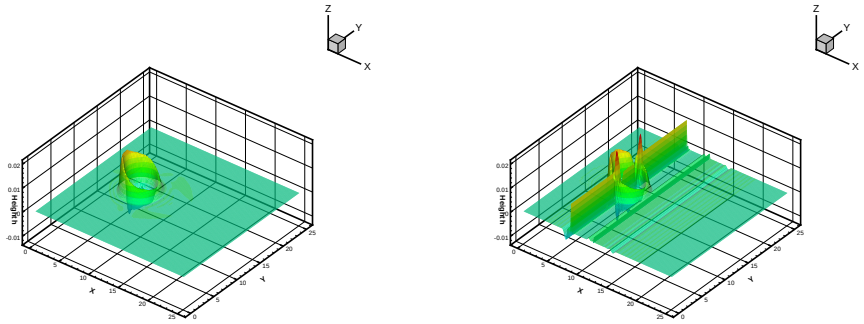

Figure 18 – The 3D figure of the difference between the height h and the initial steady state (5.12) for the problem in Section 5.5 at time $t = 0.5$ with a 200×200 uniform mesh. Left: results based on FV4-Q. Right: results based on FV5-D.

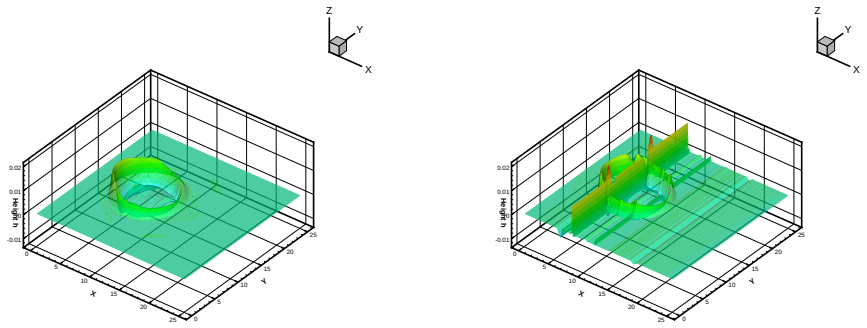

Figure 19 – The 3D figure of the difference between the height h and the initial steady state (5.12) for the problem in Section 5.5 at time $t = 1$ with a 200×200 uniform mesh. Left: results based on FV4-Q. Right: results based on FV5-D.

6. Conclusion

In this paper we gave an overview of some recently developed high-order well-balanced schemes, including fourth and fifth order schemes. The excellent resolution of the schemes is demonstrated by a number of challenging experiments for the shallow water equations. The presentation and discussion of the construction principles should enable the reader to implement them and develop them further for an application at hand. The constructions were either based on well-balanced, high-order accurate, non-oscillatory finite difference operators, or an well-balanced and accurate quadrature for the regular and singular parts of the cell-averaged residuals. The finite difference construction may be somewhat simpler and faster, which would play an even greater role in several space dimensions. But the quadrature approach can already handle moving water steady states and is in this sense more general.

References

[1] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein and B. Perthame: *A fast and stable well-balanced scheme with hydrostatic re-*

construction for shallow water flows, SIAM J. Sci. Comput. 25 (2004), 2050-2065.

[2] D.S. Bale, R.J. LeVeque, S. Mitran, and J.A. Rossmanith: *A wave-propagation method for conservation laws with spatially varying flux functions*, SIAM J. Sci. Comput. 24 (2002), 955-978.

[3] D.S. Balsara and C.-W. Shu: *Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy*, J. Comput. Phys. 160 (2000), 405-452.

[4] F. Bouchut, J. Le Sommer and V. Zeitlin: *Frontal geostrophic nonlinear wave phenomena in one dimensional rotating shallow water. Numerical simulations.* J. Fluid Mech. 514 (2004),35-63.

[5] M. Castro, J.M. Gallardo, J.A. Lopez, C. Parés: *Well-balanced high order extensions of Godunov's method for semi-linear balance laws*, SIAM J. Num. Anal. 46 (2008), 1012-1039.

[6] M. Castro, A. Pardo, C. Parés: *Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique*, Math. Mod. Meth. App. Sci. (M3AS) 17 (2007), 2055-2113.

[7] A. Chinnayya, A.Y. LeRoux and N. Seguin: *A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon*, Int. J. Fin. Vol. 1 (2004), electronic.

[8] B. Cockburn: *Discontinuous Galerkin methods for convection-dominated problems*, in *High-Order Methods for Computational Physics*, T.J. Barth and H. Deconinck, editors, Lecture Notes in Computational Science and Engineering, volume 9, Springer, 1999, 69-224.

[9] B. Cockburn, S.-Y. Lin and C.-W. Shu: *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems*, J. Comput. Phys. 84 (1989), 90-113.

[10] B. Cockburn and C.-W. Shu: *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Math. Comp. 52 (1989), 411-435.

[11] B. Cockburn, C.-W. Shu: *The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems*, J. Comput. Phys. 141 (1998), 199-224.

[12] B. Cockburn and C.-W. Shu: *Runge-Kutta Discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput. 16 (2001), 173-261.

[13] G. Dal Maso, P. Lefloch, F. Murat: *Definition and weak stability of nonconservative products*, J. Math. Pures Appl. 74 (1995), 483-548.

[14] L1. Gascón and J.M. Corberán: *Construction of second-order TVD schemes for nonhomogeneous hyperbolic conservation laws*, J. Comput. Phys. 172 (2001), 261-297.

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

[16] T. Hillen: *Hyperbolic models for chemosensitive movement*, Math. Mod. Meth. App. Sci. (M3AS) 12 (2002), 1007-1034.

[17] G. Jiang and C.-W. Shu: *Efficient implementation of weighted ENO schemes*, J. Comput. Phys. 126 (1996), 202-228.

[18] 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. Available on http://www-gm3.univ-mrs.fr/ leroux/publications/ay.le_roux.html (in French).

[19] R.J. LeVeque: *Balancing source terms and flux gradients on high-resolution Godunov methods: the quasi-steady wave-propagation algorithm*, J. Comput. Phys. 146 (1998), 346-365.

[20] R. LeVeque: *Finite volume methods for hyperbolic problems*, Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[21] X.-D. Liu, S. Osher and T. Chan: *Weighted essentially nonoscillatory schemes*, J. Comput. Phys. 115 (1994), 200-212.

[22] M. Lukáčová-Medvid'ová, S. Noelle and M. Kraft: *Well-balanced finite volume evolution Galerkin methods for the shallow water equations*, J. Comput. Phys. Vol. 221 (2007), 122-147.

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

[24] S. Noelle, Y. Xing and C.-W. Shu: *High order well-balanced Finite Volume WENO schemes for shallow water equation with moving water*, J. Comput. Phys. 226 (2007), 29-58.

[25] N. Pankratz, J. Natvig, B. Gjevik and S. Noelle: *High-order well-balanced finite-volume schemes for barotropic flows. Development and numerical comparisons*, Ocean Mod. 18 (2007), 53-79.

[26] C. Parés: *Numerical methods for nonconservative hyperbolic systems. A theoretical framework*, SIAM J. Num. Anal. 44 (2006), 300-321.

[27] J. Shi, C. Hu and C.-W. Shu: *A technique of treating negative weights in WENO schemes*, J. Comput. Phys. 175 (2002), 108-127.

[28] C.-W. Shu: *TVB uniformly high-order schemes for conservation laws*, Math. Comp. 49 (1987), 105-121.

[29] C.-W. Shu: *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput. 9 (1988), 1073–1084.

[30] C.-W. Shu: *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor (Editor: A. Quarteroni), Lecture Notes in Mathematics, volume 1697, Springer, 1998, 325-432.

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

[32] 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.

[33] S. Vukovic and L. Sopta: *ENO and WENO schemes with the exact conservation property for one-dimensional shallow water equations*, J. Comput. Phys. 179 (2002), 593-621.

[34] Y. Xing and C.-W. Shu: *High order finite difference WENO schemes with the exact conservation property for the shallow water equations*, J. Comput. Phys. 208 (2005), 206-227.

[35] Y. Xing and 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.

[36] Y. Xing and C.-W. Shu: *High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms*, J. Comput. Phys. 214 (2006), 567-598.

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