High order well-balanced numerical schemes for hyperbolic systems with source terms

by

Yulong Xing

B.Sc., University of Science and Technology of China, P.R.China, 2002

M.Sc., Brown University, Providence, RI, 2004

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2006

Abstract of “High order well-balanced numerical schemes for
hyperbolic systems with source terms,” by Yulong Xing, Ph.D., Brown University,
May 2006

Hyperbolic balance laws have steady state solutions in which the flux gradients are nonzero but are exactly balanced by the source term. This thesis contains several topics on constructing genuinely high order accurate well balanced numerical schemes, which can preserve exactly these steady state solutions.

In the first part, we start our investigation by designing high order well balanced WENO finite difference schemes for the still water solution of the shallow water equations, and then generalize our idea to a general class of balance laws with separable source terms. Well balanced high order finite volume weighted essentially non-oscillatory (WENO) schemes and Runge-Kutta discontinuous Galerkin (RKDG) finite element schemes, which are more suitable for computations in complex geometry and / or for using adaptive meshes, are also designed for the same class of balance laws. The key ingredient in our design is a special decomposition of the source term before discretization, which allows us to design specific approximations such that the resulting schemes satisfy the well balanced property, and at the same time maintain their original high order accuracy and essentially non-oscillatory property for general solutions.

In the second part, we present a different approach to design high order well-balanced finite volume WENO schemes and RKDG finite element methods. We make the observation that the traditional RKDG methods are capable of maintaining certain steady states exactly, if a small modification on either the initial condition or the flux is provided. The computational cost to obtain such a well balanced RKDG method is basically the same as the traditional RKDG method.

The third topic is related to the moving steady state solution of the shallow water equation, which cannot be preserved by the above methods. We introduce a new technique to obtain high order finite volume schemes for this problem, based on a special treatment of the flux and source term. Extensive numerical simulations are performed to verify high order accuracy, the well balanced property, and good resolution for smooth and discontinuous solutions.

This dissertation by Yulong Xing is accepted in its present form
by the Department of Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date________________ ______________________________
Chi-Wang Shu, Ph.D., Advisor

Recommended to the Graduate Council

Date________________ ______________________________
David Gottlieb,Ph.D., Reader

Date________________ ______________________________
Walter Strauss, Ph.D., Reader

Approved by the Graduate Council

Date________________ ______________________________
Sheila Bonde, Dean of the Graduate School

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

We are interested in numerically solving the following hyperbolic conservation laws with source terms, also referred to as hyperbolic balance laws:

$$u_t + f_1(u,x,y)_x + f_2(u,x,y)_y = g(u,x,y) \tag{1.1}$$

or in the one dimensional case

$$u_t + f(u,x)_x = g(u,x) \tag{1.2}$$

where u is the solution vector, $f_1(u,x,y)$ and $f_2(u,x,y)$ (or $f(u,x)$) are the fluxes and $g(u,x,y)$ (or $g(u,x)$) is the source term. Hyperbolicity refers to the fact that the Jacobians $\frac{\partial f_1(u,x,y)}{\partial u}$ and $\frac{\partial f_2(u,x,y)}{\partial u}$ (or $\frac{\partial f(u,x)}{\partial u}$) always have real eigenvalues and complete sets of eigenvectors. Often, this balance law would admit steady state solutions in which the source term is exactly balanced by the flux gradient. Notice that in such situations the solution u is typically a non-trivial function, hence a straightforward numerical scheme may fail to preserve exactly this balance. Many physical phenomena come from small perturbations of these steady state solutions, which are very difficult to capture numerically, unless the numerical schemes can preserve the unperturbed steady state at the discrete level. Schemes which can preserve the unperturbed steady state at the discrete level are the so called well balanced schemes. Our purpose is to design well balanced schemes without sacrificing the high order accuracy and non-oscillatory properties of the scheme when applied to general, non-steady state solutions.

A prototype example for the balance laws (1.2), which has been investigated extensively in the literature, is the shallow water equation with a non-flat bottom topology. The shallow water equations, also referred to as the Saint-Venant system, are widely used to model flows in rivers and coastal areas. It has wide applications in ocean and hydraulic engineering: tidal flows in estuary and coastal water region; bore wave propagation; and river, reservoir, and open channel flows, among others. This system describes the flow as a conservation law with additional source terms. We consider the system with a geometrical source term due to the bottom topology. In one space dimension, it takes the form

$$\begin{cases} h_t + (hu)_x = 0 \\ (hu)_t + \left(hu^2 + \dfrac{1}{2}gh^2\right)_x = -ghb_x, \end{cases} \tag{1.3}$$

where h denotes the water height, u is the velocity of the fluid, b represents the bottom topography and g is the gravitational constant. In the homogeneous case, the system is equivalent to that of the isentropic Euler system. However, the properties of the system change a lot due to the presence of the source term. The above system is quite simple in the sense that only the topography of the bottom is taken into account, but other terms could also be added in order to include effects such as friction on the bottom and on the surface as well as variations of the channel width.

Research on numerical methods for the solution of the shallow water system has attracted much attention in the past two decades. Many numerical schemes have been developed to solve this system. Similar to other balance laws, this system admits stationary solutions in which nonzero flux gradients are exactly balanced by the source terms in the steady state case. The steady state solutions are given by

$$hu = constant \qquad and \qquad \frac{1}{2}u^2 + g(h+b) = constant. \tag{1.4}$$

A special case is the still water stationary solution, denoted by

$$u = 0 \qquad and \qquad h + b = constant. \tag{1.5}$$

1.2 History

A significant result in maintaining the stationary solution (1.5) is given by Bermudez and Vazquez [5]. They have proposed the idea of the "exact C-property", which means that the scheme is "exact" when applied to the stationary case $h + b = constant$ and $hu = 0$. This property is necessary for maintaining the above balance. A good scheme for the shallow water system should satisfy this property. Also, they have introduced the first order Q-scheme and the idea of source term upwinding. After this pioneering work, many other schemes for the shallow water equations with such exact C-property have been developed. LeVeque [28] has introduced a quasi-steady wave propagation algorithm. A Riemann problem is introduced in the center of each grid cell such that the flux difference exactly cancels the source term. Zhou et al. [51] use the surface gradient method for the treatment of the source terms. They use $h+b$ for the reconstruction instead of using h. Russo [36] and Kurganov and Levy [27] apply finite volume central-upwind schemes to this system, keeping higher-order accuracy for the flux term and second order accuracy for the source term. Recently, Vukovic and Sopta [44] have used the essentially non-oscillatory (ENO) and weighted ENO (WENO) schemes for this problem. They applied WENO reconstruction not only to the flux but also to a combination of the flux and the source term. For related work, see also [2, 18, 23, 25, 27, 32, 34, 36, 50]. All of these works are for well balanced numerical schemes for the stationary solution (1.5). Little in the literature has been done for the more general steady state problem (1.4).

Most of the works mentioned above are for numerical schemes of at most second order accuracy. There is a fundamental difficulty in maintaining genuine high order accuracy for the general solutions and at the same time achieving the exact C-property. The work mentioned above which addresses this issue is [44], see also [45, 14]. They have applied the ENO and WENO schemes to the shallow water equation to maintain the steady state. First, they split the flux term into the original Q-scheme [5] and two modification terms (the WENO reconstruction of some function $w^{\pm}$). In order to obtain a well-balanced scheme, they also discretize the source term in a similar way: the sum of the source term in Q-scheme [5] and two modification terms (the WENO reconstruction of the function $v^{\pm}$, which is a discretization of the source term). In the steady state case, the flux term and source term can be exactly balanced one by one ($w^{\pm}$=$v^{\pm}$). Hence the well balanced property is obtained.

It is not easy to see whether the source term in [44] is discretized with high order directly. After checking the Taylor expansion of the source term discretization, we find the source term is approximated by: $-ghb' + \frac{g}{12}(h'b'' - h''b')\triangle x^2 + O(\triangle x^4)$, which shows that actually it has only second order accuracy. In [44], the authors did a convergence test to find the order of this scheme. The example they picked is a steady state solution, which should be preserved. But for that example, the maximum of the coefficient $\frac{g}{12}(h'b'' - h''b')$ is 1.8×10^{-4}. So the truncation error given by $\frac{g}{12}(h'b'' - h''b')\triangle x^2$ is only 4.5×10^{-7} if $n = 20$ grid points are used, which is much smaller than the computed error (around 10^{-3}, see TABLE I of [44]). Even for the case $n = 320$, the second order term is 1.8×10^{-9}, which is also smaller than the computed error 1.1×10^{-8}. So the truncation error is dominated by the high order term. This is why we can see high order property for that example. Also, they have extended the scheme to the one-dimensional elastic wave equations [45]. The order of this scheme applied to a Cauchy problem in linear acoustics is also computed.

The Taylor expansion for this problem is: $-u\rho' + \frac{\rho''u'-\rho'u''}{12}\triangle x^2 + O(\triangle x^4)$. We have recomputed their scheme to obtain the following results: (since final time t=0.001s is too small, we use t=0.1s instead. Numerical result at $n = 10240$ is used as a reference solution in order to compute the errors.)

Table 1.1: The convergence result for the linear acoustics problem

Number of cells	$\rho\epsilon$		ρu	
	L^1 error	order	L^1 error	order
10	5.578E-002		4.408E-002	
20	2.323E-002	1.2643	2.298E-002	0.9397
40	3.091E-003	2.9098	2.845E-003	3.0139
80	3.181E-004	3.2805	2.619E-004	3.4413
160	2.726E-005	3.5446	2.276E-005	3.5244
320	5.556E-006	2.2947	4.138E-006	2.4595
640	1.348E-006	2.0432	9.541E-007	2.1167
1280	3.327E-007	2.0185	2.294E-007	2.0563

Although it exhibits high order at first, the order approaches to 2 as we refine the mesh. This shows that the schemes in [44] and [45] are high order accurate for solutions of certain specific forms, but seem to be still only second order accurate for the general solutions based on truncation error analysis and numerical results.

1.3 Well balanced high order schemes

Our goal is to design finite difference, finite volume WENO and RKDG finite element schemes, which maintain the well balanced property and at the same time are genuinely high order accurate for the general solutions of hyperbolic systems with source terms. Several different approaches have been introduced in this thesis. We first start from a special decomposition of the source term before discretization. By applying WENO reconstruction on each component of the source term, we can then design a well balanced scheme. This idea has been successfully applied on finite difference, finite volume WENO and RKDG finite element schemes. We also observe that the traditional RKDG methods are capable of maintaining certain steady states exactly, if a small modification on either the initial condition or the flux is provided, which provides us a new approach to obtain well balanced schemes. At the end, we introduce a new well balanced scheme aimed for the moving steady state solution of the shallow water equations.

We will briefly review the traditional high order finite difference, finite volume WENO and RKDG finite element methods in Chapter 2, emphasizing the features of the methods which are important for the design of well balanced high order schemes.

In Chapter Three, we concentrate on the shallow water equations, and design a WENO finite difference scheme which maintains the exact C-property and at the same time is genuinely high order accurate for the general solutions of the shallow water equations. A key ingredient in our design is a special splitting of the source term into two parts which are discretized separately.

In Chapter Four, we extend this idea of decomposition of source terms introduced in Chapter Three to a general class of balance laws with separable source terms, allowing the design of well balanced high order finite difference WENO scheme for all balance laws falling into this category. This class is quite general, including the elastic wave equation, the hyperbolic model for a chemosensitive movement, the nozzle flow and a two phase flow model.

Well balanced high order finite volume WENO schemes and finite element RKDG schemes on general triangulations are designed for the same class of balance laws in Chapter Five. Compared with the finite difference schemes, they are more suitable for computations in complex geometry and / or for using adaptive meshes. Even though the detailed technical approaches are different, the framework of the algorithm construction in this chapter follows that in Chapter Four.

In Chapter Six, we present a completely different setup for well balanced finite volume WENO and RKDG methods, which can be considered as a generalization of a well balanced high order scheme recently developed by Noelle et al. [30]. Traditional RKDG methods with a special treatment of the flux are proven to be well balanced for certain steady state solutions. Very little additional computational cost is involved to obtain such property. Similar ideas are then applied to obtain well balanced finite volume WENO schemes. Comparing with the well balanced schemes developed in Chapter Five, the well balanced RKDG schemes here are simpler and involve less modification to the original RKDG methods, while the well balanced WENO finite volume schemes here and that in Chapter Five are comparable in computational cost.

Moving steady state (1.4) of the shallow water equation is more difficult to be maintained exactly, due to the presence of the nonlinear term $\frac{1}{2}u^2 + g(h+b)$ in the steady state. In Chapter Seven, we present a different technique to design high order well balanced finite volume WENO scheme for this problem.

All these works are based on a joint work with Professor Chi-Wang Shu, and Chapter Seven is also a joint work with Professor Sebastian Noelle. Some contents have previously appeared in [46, 47, 48, 49].

Chapter 2

Review of High Order Numerical Schemes

2.1 Finite difference WENO schemes

In this section we give a short overview of the finite difference WENO schemes. For more details, we refer to [29, 24, 4, 40, 41, 42].

First, we consider a scalar hyperbolic conservation law equation in one dimension

$$u_t + f(u)_x = 0,$$

with a positive wind direction $f'(u) \geq 0$. For a finite difference scheme, we evolve the point values u_i at mesh points x_i in time. We assume the mesh is uniform with mesh size Δx for simplicity. The spatial derivative $f(u)_x$ is approximated by a conservative flux difference

$$f(u)_x|_{x=x_i} \approx \frac{1}{\Delta x}\left(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}\right). \tag{2.1}$$

The numerical flux $\hat{f}_{i+\frac{1}{2}}$ is computed through the neighboring point values $f_j =$

$f(u_j)$. For a (2k-1)-th order WENO scheme, we first compute k numerical fluxes

$$\hat{f}^{(r)}_{i+\frac{1}{2}} = \sum_{j=0}^{k-1} c_{rj} f_{i-r+j}, \qquad r = 0, ..., k-1,$$

corresponding to k different candidate stencils $S_r(i) = \{x_{i-r}, ..., x_{i-r+k-1}\}$, $r = 0, ..., k-1$. Each of these k numerical fluxes is k-th order accurate. For example, when $k = 3$ (fifth order WENO scheme), the three third order accurate numerical fluxes are given by:

$$\begin{aligned}
\hat{f}^{(0)}_{i+1/2} &= \frac{1}{3}f_i + \frac{5}{6}f_{i+1} - \frac{1}{6}f_{i+2}, \\
\hat{f}^{(1)}_{i+1/2} &= -\frac{1}{6}f_{i-1} + \frac{5}{6}f_i + \frac{1}{3}f_{i+1}, \\
\hat{f}^{(2)}_{i+1/2} &= \frac{1}{3}f_{i-2} - \frac{7}{6}f_{i-1} + \frac{11}{6}f_i.
\end{aligned}$$

The (2k-1)-th order WENO flux is a convex combination of all these k numerical fluxes

$$\hat{f}_{i+\frac{1}{2}} = \sum_{r=0}^{k-1} w_r \hat{f}^{(r)}_{i+\frac{1}{2}}.$$

The nonlinear weights w_r satisfy $w_r \geq 0$, $\sum_{r=0}^{k-1} w_r = 1$, and are defined in the following way:

$$w_r = \frac{\alpha_r}{\sum_{s=0}^{k-1}\alpha_s}, \qquad \alpha_r = \frac{d_r}{(\varepsilon + \beta_r)^2}. \tag{2.2}$$

Here d_r are the linear weights which yield (2k-1)-th order accuracy, β_r are the so-called "smoothness indicators" of the stencil $S_r(i)$ which measure the smoothness of the function $f(u(x))$ in the stencil. ε is a small constant used to avoid the denominator to become zero and is typically taken as 10^{-6}. For example, when $k = 3$ (fifth order WENO scheme), the linear weights are given by

$$d_0 = \frac{3}{10}, \qquad d_1 = \frac{3}{5}, \qquad d_2 = \frac{1}{10},$$

and the smoothness indicators are given by

$$\begin{aligned}\beta_0 &= \frac{13}{12}\left(f_i-2f_{i+1}+f_{i+2}\right)^2+\frac{1}{4}\left(3f_i-4f_{i+1}+f_{i+2}\right)^2\\ \beta_1 &= \frac{13}{12}\left(f_{i-1}-2f_i+f_{i+1}\right)^2+\frac{1}{4}\left(f_{i-1}-f_{i+1}\right)^2\\ \beta_2 &= \frac{13}{12}\left(f_{i-2}-2f_{i-1}+f_i\right)^2+\frac{1}{4}\left(f_{i-2}-4f_{i-1}+3f_i\right)^2.\end{aligned}$$

The procedure for the case with $f'(u) \leq 0$ is mirror symmetric with respect to $i+\frac{1}{2}$. More details can be found in [24, 40].

An upwinding mechanism, essential for the stability of the scheme, can be realized by a global “flux splitting”. The simplest one is the Lax-Friedrichs splitting:

$$f^{\pm}(u) = \frac{1}{2}(f(u) \pm \alpha u), \tag{2.3}$$

where α is taken as $\alpha = \max_u |f'(u)|$. The WENO procedure is applied to $f^{\pm}$ individually with upwind biased stencils. Depending on whether the max is taken globally (along the line of computation) or locally, such schemes are referred to as the Lax-Friedrichs WENO scheme (WENO-LF) or the local Lax-Friedrichs WENO scheme (WENO-LLF).

For hyperbolic systems such as the shallow water equations, we use the local characteristic decomposition, which is more robust than a component by component version. First, we compute an average state $u_{i+\frac{1}{2}}$ between u_i and u_{i+1}, using either the simple arithmetic mean or a Roe's average [35]. The right eigenvectors r_m and the left eigenvectors l_m of the Jacobian $f'(u_{i+\frac{1}{2}})$ are needed for the local characteristic decomposition. The WENO procedure is used on

$$v_j^{\pm} = R^{-1} f_j^{\pm}, \qquad \text{j in a neighborhood of i.} \tag{2.4}$$

where $R = (r_1, ..., r_n)$ is the matrix whose columns are the right eigenvectors of $f'(u_{i+\frac{1}{2}})$. The numerical fluxes $\hat{v}^{\pm}_{i+\frac{1}{2}}$ thus computed are then projected back into the physical space by left multiplying with R, yielding finally the numerical fluxes in the physical space.

With the numerical fluxes $\hat{f}_{i+\frac{1}{2}}$, $f(u)_x$ is approximated by (2.1) to high order accuracy at $x = x_i$. Together with a TVD high order Runge-Kutta time discretization [41], this completes a high order WENO scheme. Multi-dimensional problems are handled in the same fashion, with each derivative approximated along the line of the relevant variable. Again, we refer to [24, 40] for further details.

2.2 Finite volume WENO schemes

In this section, we briefly review the basic ideas of finite volume WENO schemes. For further details, we refer to [29, 38, 22, 24, 4, 40, 41, 42].

First, we consider a scalar hyperbolic conservation law equation in one dimension

$$u_t + f(u)_x = 0, \tag{2.5}$$

and discretize the computational domain with cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $i = 1, \cdots, N$. We denote the size of the i-th cell by $\triangle x_i$ and the center of the cell by $x_i = \frac{1}{2}\left(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}\right)$. Let $\bar{u}(x_i, t) = \frac{1}{\triangle x_i}\int_{I_i} u(x,t)\,dx$ denote the cell average of $u(\cdot, t)$ over the cell I_i. In a finite volume scheme, our computational variables are $\bar{u}_i(t)$, which approximate the cell averages $\bar{u}(x_i, t)$.

For finite volume schemes, we solve an integrated version of (2.5):

$$\frac{d}{dt}\bar{u}(x_i, t) = -\frac{1}{\triangle x_i}\left(f\left(u(x_{i+\frac{1}{2}}), t\right) - f\left(u(x_{i-\frac{1}{2}}), t\right)\right).$$

This is approximated by the following conservative scheme:

$$\frac{d}{dt}\bar{u}_i(t) = -\frac{1}{\triangle x_i}\left(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}\right) \tag{2.6}$$

with $\hat{f}_{i+\frac{1}{2}} = F(u^-_{i+\frac{1}{2}}, u^+_{i+\frac{1}{2}})$ being the numerical flux. Here $u^-_{i+\frac{1}{2}}$ and $u^+_{i+\frac{1}{2}}$ are the high order pointwise approximations to $u(x_{i+\frac{1}{2}}, t)$, obtained from the cell averages by a high order WENO reconstruction procedure. Flux $F(a, b)$ is consistent if it reduces to the true flux f for the case of constant flow, i.e.

$$F(a, a) = f(a) \qquad \forall a \in \mathbb{R}. \tag{2.7}$$

In order to obtain a stable scheme, the numerical flux $F(a, b)$ needs to be a monotone flux, namely F is a nondecreasing function of its first argument a and a nonincreasing function of its second argument b. There are many choices of these fluxes, such as the Godunov flux, the Engquist-Osher flux and the Lax-Fridrichs (LF) flux. The difference among these fluxes is significant for low order schemes but becomes less significant for higher order reconstructions. The simplest and most inexpensive monotone flux is the Lax-Friedrichs flux:

$$F(a, b) = \frac{1}{2}(f(a) + f(b) - \alpha(b - a)), \tag{2.8}$$

where $\alpha = \max_u |f'(u)|$. Depending on whether the maximum is taken globally (along the line of computation) or locally, this flux is referred to as the Lax-Friedrichs (LF) or the local Lax-Friedrichs (LLF) flux.

The approximations $u^-_{i+\frac{1}{2}}$ and $u^+_{i+\frac{1}{2}}$ are computed through the neighboring cell average values $\bar{u}_j$. For a (2k-1)-th order WENO scheme, we first compute k reconstructed values

$$\hat{u}^{(r)}_{i+\frac{1}{2}} = \sum_{j=0}^{k-1} c_{rj} \bar{u}_{i-r+j}, \qquad r = 0, ..., k-1,$$

corresponding to k different candidate stencils

$$S_r(i) = \{x_{i-r}, ..., x_{i-r+k-1}\}, \qquad r = 0, ..., k-1. \tag{2.9}$$

The coefficients c_{rj} are chosen such that each of these k reconstructed values is k-th order accurate, see [40]. Also, we obtain the k reconstructed values $\tilde{u}^{(r)}_{i-\frac{1}{2}}$, of k-th order accuracy, using

$$\tilde{u}^{(r)}_{i-\frac{1}{2}} = \sum_{j=0}^{k-1} \tilde{c}_{rj}\bar{u}_{i-r+j}, \qquad r = 0, ..., k-1,$$

with

$$\tilde{c}_{rj} = c_{r-1,j},$$

based on the same stencils (2.9). The (2k-1)-th order WENO reconstruction is a convex combination of all these k reconstructed values

$$u^{-}_{i+\frac{1}{2}} = \sum_{r=0}^{k-1} w_r \hat{u}^{(r)}_{i+\frac{1}{2}}, \qquad u^{+}_{i-\frac{1}{2}} = \sum_{r=0}^{k-1} \tilde{w}_r \tilde{u}^{(r)}_{i-\frac{1}{2}}.$$

The nonlinear weights w_r satisfy $w_r \geq 0$, $\sum_{j=0}^{k-1} w_r = 1$, and are defined in the following way:

$$w_r = \frac{\alpha_r}{\sum_{s=0}^{k-1}\alpha_s}, \qquad \alpha_r = \frac{d_r}{(\varepsilon + \beta_r)^2}. \tag{2.10}$$

Here d_r are the linear weights which yield (2k-1)-th order accuracy, β_r are the so-called "smoothness indicators" of the stencil $S_r(i)$ which measure the smoothness of the function $u(x)$ in the stencil. ε is a small constant used to avoid the denominator to become zero and is typically taken as 10^{-6}. By symmetry, $\tilde{w}_r$ is computed by:

$$\tilde{w}_r = \frac{\tilde{\alpha}_r}{\sum_{s=0}^{k-1}\tilde{\alpha}_s}, \qquad \tilde{\alpha}_r = \frac{\tilde{d}_r}{(\varepsilon + \beta_r)^2}, \tag{2.11}$$

with

$$\tilde{d}_r = d_{k-1+r}. \tag{2.12}$$

The exact form of the smoothness indicators and other details about the WENO reconstruction can be found in [24, 40].

For hyperbolic systems such as the shallow water equations, we use the local characteristic decomposition, which is more robust than a component by component version. First, we compute an average state $\bar{u}_{i+\frac{1}{2}}$ between $\bar{u}_i$ and $\bar{u}_{i+1}$, using either the simple arithmetic mean or a Roe's average [35]. The WENO procedure is used on

$$\bar{v}_j = R^{-1}\bar{u}_j, \text{ j in a neighborhood of i.} \tag{2.13}$$

where $R = (r_1, ..., r_n)$ is the matrix whose columns are the right eigenvectors of $f'(\bar{u}_{i+\frac{1}{2}})$. The reconstructed values $v^{\pm}_{i+\frac{1}{2}}$ thus computed are then projected back into the physical space by left multiplying with R, yielding finally the reconstructed values in the physical space.

With the reconstructed values $u^{\pm}_{i+\frac{1}{2}}$, the right hand side of (2.6) can be computed through (2.8) to high order accuracy. Together with a TVD high order Runge-Kutta time discretization [41], this completes the description of a high order finite volume WENO scheme.

Finite volume WENO schemes in the two dimensional case have the same framework but are more complicated to implement. In this thesis, we consider only rectangular cells for simplicity, although the technique also works for general triangulations. Consider the two dimensional homogeneous conservation law

$$u_t + f_1(u, x, y)_x + f_2(u, x, y)_y = 0, \tag{2.14}$$

together with a spatial discretization of the computational domain with cells $I_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, $i = 1, \cdots, N_x$, $j = 1, \cdots, N_y$. As usual, we use the notations:

$$\triangle x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \qquad \triangle y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$$

to denote the grid sizes.

We integrate (2.14) over the interval I_{ij} to obtain:

$$\begin{aligned}\frac{d}{dt}\bar{u}(x_i,y_j,t) &= -\frac{1}{\triangle x_i \triangle y_j}\left(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f_1(u(x_{i+\frac{1}{2}},y,t))dy - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f_1(u(x_{i-\frac{1}{2}},y,t))dy\right.\\ &\left. + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f_2(u(x,y_{j+\frac{1}{2}},t))dx - \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f_2(u(x,y_{j-\frac{1}{2}},t))dx\right) \quad (2.15)\end{aligned}$$

where

$$\bar{u}(x_i,y_j,t) = \frac{1}{\triangle x_i \triangle y_j}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(\xi,\eta,t)\, d\xi d\eta$$

is the cell average. We approximate (2.15) by the conservative scheme

$$\frac{d}{dt}\bar{u}_{ij}(t) = -\frac{1}{\triangle x_i}\left((\hat{f}_1)_{i+\frac{1}{2},j} - (\hat{f}_1)_{i-\frac{1}{2},j}\right) - \frac{1}{\triangle y_j}\left((\hat{f}_2)_{i,j+\frac{1}{2}} - (\hat{f}_2)_{i,j-\frac{1}{2}}\right), \quad (2.16)$$

where the numerical flux $(\hat{f}_1)_{i+\frac{1}{2},j}$ is defined by

$$(\hat{f}_1)_{i+\frac{1}{2},j} = \sum_\alpha w_\alpha F\left(u^-_{x_{i+\frac{1}{2}},y_j+\beta_\alpha \triangle y_j}, u^+_{x_{i+\frac{1}{2}},y_j+\beta_\alpha \triangle y_j}\right) \quad (2.17)$$

where β_α and w_α are the Gaussian quadrature nodes and weights, to approximate the integration in y:

$$\frac{1}{\triangle y_j}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f_1(u(x_{i+\frac{1}{2}},y,t))dy.$$

The monotone flux $F(a,b)$ is the same as defined above (for example, Formula (2.8)). $u^\pm_{x_{i+\frac{1}{2}},y_j+\beta_\alpha \triangle y_j}$ are the $(2k-1)$-th order accurate reconstructed values obtained by a WENO reconstruction procedure. In this procedure, for rectangular meshes, if we use the tensor products of one dimensional polynomials, i.e. polynomials in Q^{k-1}, things can proceed as in one dimension. A practical way to perform the reconstruction in two dimensions is given as follows. We first perform an one dimensional reconstruction in one of the directions (e.g. the y-direction), obtaining one dimensional cell averages of the function u in the other direction (e.g. the x-direction).

We then perform a reconstruction in the other direction to obtain the approximated point values, see [40, 38].

Similarly, we can compute the flux $(\hat{f}_2)_{i,j+\frac{1}{2}}$ by

$$(\hat{f}_2)_{i,j+\frac{1}{2}} = \sum_{\alpha} w_\alpha F\left(u^-_{x_i+\beta_\alpha \triangle x_i, y_{j+\frac{1}{2}}}, u^+_{x_i+\beta_\alpha \triangle x_i, y_{j+\frac{1}{2}}}\right). \tag{2.18}$$

2.3 Discontinuous Galerkin methods

In this section, we give a short overview of another widely used high order scheme, namely the Runge-Kutta discontinuous Galerkin method, which was first introduced by Cockburn and Shu. We refer to [11, 10, 12, 8, 13] for more information.

Again, a scalar hyperbolic conservation law in one dimension is considered:

$$u_t + f(u)_x = 0, \qquad u(x,0) = u_0(x). \tag{2.19}$$

As before, we discretize the computational domain into cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, and denote the size of the i-th cell by $\triangle x_i$ and the maximum mesh size by $h = \max_i \triangle x_i$.

First, we multiply the equation (2.19) by an arbitrary smooth function v, integrate it over cell I_j and perform integration by parts to obtain

$$\begin{aligned} \int_{I_j} \partial_t u(x,t) v(x) dx &- \int_{I_j} f(u(x,t)) \partial_x v(x) dx \\ &+ f(u(x_{j+\frac{1}{2}}, t)) v(x_{j+\frac{1}{2}}) - f(u(x_{j-\frac{1}{2}}, t)) v(x_{j-\frac{1}{2}}) = 0, \end{aligned} \tag{2.20}$$

$$\int_{I_j} u(x,0) v(x) dx = \int_{I_j} u_0(x) v(x) dx.$$

The main difference between the DG method and a traditional finite element method lies in the choice of the test space and solution space. Here, we seek an approximation u_h to u which belongs to the finite dimensional space

$$V_h = V_h^k \equiv \{v : v|_{I_j} \in P^k(I_j), j = 1, ..., N\}, \tag{2.21}$$

where $P^k(I)$ denotes the space of polynomials in I of degree at most k. Notice that u_h can be discontinuous at the cell boundary $x_{j+\frac{1}{2}}$. In equation (2.20), we replace the smooth functions v by test functions v_h from the test space V_h, and u by the numerical solution u_h. Together with the replacement of the nonlinear flux $f(u(x_{j+\frac{1}{2}}, t))$ by a numerical flux $\hat{f}_{j+\frac{1}{2}} = F(u_h(x^-_{j+\frac{1}{2}}, t), u_h(x^+_{j+\frac{1}{2}}, t))$, we obtain the numerical scheme denoted by

$$\int_{I_j} \partial_t u_h(x,t) v_h(x) dx - \int_{I_j} f(u_h(x,t)) \partial_x v_h(x) dx + \hat{f}_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) = 0,$$

$$\int_{I_j} u_h(x,0) v_h(x) dx = \int_{I_j} u_0(x) v_h(x) dx. \tag{2.22}$$

As before, $F(a, b)$ is chosen as a monotone flux to recover a finite volume monotone scheme for the piecewise constant $k = 0$ case. We could, for example, again use the simple Lax-Friedrichs flux (2.8).

Another important ingredient for the RKDG method is that a slope limiter procedure should be performed after each inner stage in the Runge-Kutta time stepping. This is necessary for computing solutions with strong discontinuities. There are many choices for the slope limiters, see, e.g. [33]. In this thesis we use the total variation bounded (TVB) limiter in [39, 11, 9, 12]; we refer to these references for the details of this limiter.

Together with a TVD high order Runge-Kutta time discretization [41], we have then finished the description of the RKDG method.

Multi-dimensional problems can be handled in the same fashion. We also perform an integration by parts (Green's formula) first, and then replace the boundary values by numerical fluxes. The main difference is that the fluxes are now integrals along the cell boundary, which can be calculated by Gauss-quadrature rules. For more details, we refer to [9, 12, 8].

Chapter 3

High Order Finite Difference Well-balanced WENO Schemes for the Shallow Water Equations

In this chapter we design high order well balanced finite difference WENO scheme for the still water stationary solution (1.5) of the shallow water equations (1.3). In Section 3.1, the WENO scheme which maintains the exact C-property and at the same time is genuinely high order accurate for the general solutions of the shallow water equations is presented. Sections 3.2 and 3.3 contain extensive numerical simulation results to demonstrate the behavior of our WENO schemes for one and two dimensional shallow water equations, verifying high order accuracy, the exact C-property, and good resolution for smooth and discontinuous solutions.

3.1 A balance of the flux and the source term

In this section we design a finite difference high order WENO-LF scheme for the shallow water equation, with the objective of keeping the exact C-property without reducing the high order accuracy of the scheme. The scheme reduces to the original

WENO-LF scheme described in Section 2.1 when the bottom is flat ($b_x = 0$). We start with the description in the one dimensional case. First, we split the source term into two separate terms in the discretization and prove, if written in this form, any linear scheme can maintain the exact C-property. Next, we apply the nonlinear WENO procedure with a small modification and prove that it can also maintain the exact C-property without losing high order accuracy.

We now describe the details. For the shallow water equation (1.3), we split the source term $-ghb_x$ into two terms $\left(\frac{1}{2}gb^2\right)_x - g(h+b)b_x$. Hence the equations become

$$\begin{cases} h_t + (hu)_x = 0 \\ (hu)_t + \left(hu^2 + \dfrac{1}{2}gh^2\right)_x = \left(\dfrac{1}{2}gb^2\right)_x - g(h+b)b_x, \end{cases} \tag{3.1}$$

which will be denoted by

$$U_t + f(U)_x = G_1 + G_2 \tag{3.2}$$

where $U = (h, hu)^T$ with the superscript T denoting the transpose, f(U) is the flux term and G_1, G_2 are the two source terms.

As we will see below, the special splitting of the source term in (3.1) is crucial for the design of our high order schemes satisfying the exact C-property. It should be noted that the two derivative terms on the right hand side of (3.1) involve only known functions, not the solution h and u. It is important *not* to include any derivatives of the unknown solution h and u on the right hand side source term. Otherwise, conservation and convergence towards weak solutions will be problematic for discontinuous solutions.

As usual, 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 (3.1) is said to be a linear scheme if all the spatial derivatives are approximated by linear finite difference operators. For the still water stationary solution of (3.1), we have

$$h + b = constant \qquad \text{and} \qquad hu = 0. \tag{3.3}$$

For any consistent linear scheme, the first equation $(hu)_x = 0$ is satisfied exactly since $hu = 0$. The second equation has the truncation error

$$D_1\left(hu^2 + \frac{1}{2}gh^2\right) - D_2\left(\frac{1}{2}gb^2\right) + g(h+b)D_3(b),$$

where D_1, D_2 and D_3 are linear finite difference operators. Since $hu = 0$, this truncation error reduces to

$$D_1\left(\frac{1}{2}gh^2\right) - D_2\left(\frac{1}{2}gb^2\right) + g(h+b)D_3(b).$$

We further restrict our attention to linear schemes which satisfy

$$D_1 = D_2 = D_3 = D \tag{3.4}$$

for the still water stationary solutions. For such linear schemes we have

Proposition 3.1.1 *Linear schemes for the shallow water equation (3.1) satisfying (3.4) for the still water stationary solutions (3.3) can maintain the exact C-property.*

Proof. For still water stationary solutions (3.3), linear schemes satisfying (3.4) is exact for the first equation $(hu)_x = 0$, and the truncation error for the second equation reduces to

$$\begin{aligned}
& D\left(\frac{1}{2}gh^2\right) - D\left(\frac{1}{2}gb^2\right) + g(h+b)D(b) \\
= \;& D\left(\frac{1}{2}gh^2 - \frac{1}{2}gb^2 + g(h+b)b\right) \\
= \;& D\left(\frac{1}{2}g(h+b)^2\right) \\
= \;& 0
\end{aligned}$$

where the first equality is due to the linearity of D and the fact that $h+b = constant$; the second equality is just a simple regrouping of terms inside the parenthesis, and the last equality is due to the fact that $h+b = constant$ and the consistency of the finite difference operator D. This finishes the proof. □

Of course, the high order finite difference WENO schemes described in the Section 2.1 are nonlinear. The nonlinearity comes from the nonlinear weights, which in turn comes from the nonlinearity of the smooth indicators β_r measuring the smoothness of the functions f^+ and f^-. We would like to make minor modifications to these high order finite difference WENO schemes, so that the exact C-property is maintained and accuracy and nonlinear stability are not affected.

To present the basic ideas, we first consider the situation when the WENO scheme is used without the flux splitting and the local characteristic decomposition. In this case the smoothness indicators β_r measure the smoothness of each component of the flux function $f(U)$. The first equation in (3.1) does not cause a problem for the still water solution, as $hu = 0$ and the consistent WENO approximation to $(hu)_x$ is exact. For the second equation in (3.1), there are three derivative terms, $\left(hu^2 + \frac{1}{2}gh^2\right)_x$, $\left(\frac{1}{2}gb^2\right)_x$ and b_x, that must be approximated. The approximation to the flux derivative term $\left(hu^2 + \frac{1}{2}gh^2\right)_x$ proceeds as before using the WENO approximation. We notice that the WENO approximation to d_x where $d = hu^2 + \frac{1}{2}gh^2$ can be eventually written out as

$$d_x\big|_{x=x_i} \approx \sum_{k=-r}^{r} a_k d_{i+k} \equiv D_d(d)_i \tag{3.5}$$

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 difference operator D_d with $d = hu^2 + \frac{1}{2}gh^2$ *fixed*, namely to use the same coefficients a_k obtained through the smoothness indicators of $d = hu^2 + \frac{1}{2}gh^2$, and apply this difference operator D_d to approximate $\left(\frac{1}{2}gb^2\right)_x$ and b_x in the source terms. Thus

$$\left(\frac{1}{2}gb^2\right)_x\Bigg|_{x=x_i} \approx \sum_{k=-r}^{r} a_k \left(\frac{1}{2}gb^2\right)_{i+k} \equiv D_d\left(\frac{1}{2}gb^2\right)_i;$$

$$b_x\big|_{x=x_i} \approx \sum_{k=-r}^{r} a_k b_{i+k} \equiv D_d(b)_i.$$

Clearly, the finite difference operator D_d, obtained from the fifth order WENO procedure, is a fifth order accurate approximation to the first derivative on any grid function, thus our approximation to the source terms is also fifth order accurate. The approximation of $\left(\frac{1}{2}gb^2\right)_x$ can even be absorbed together with the approximation of the flux derivative term $\left(hu^2 + \frac{1}{2}gh^2\right)_x$ in actual implementation to save cost (of course, the smoothness indicators to determine the nonlinear weights in the approximation would still be based on $hu^2 + \frac{1}{2}gh^2$). A key observation is that the finite difference operator D_d, with the coefficients a_k based on the smoothness indicators of $d = hu^2 + \frac{1}{2}gh^2$ fixed, is a linear operator on any grid functions, i.e.

$$D_d(af_1 + bf_2) = aD_d(f_1) + bD_d(f_2)$$

for constants a, b and arbitrary grid functions f_1 and f_2. Thus the proof of Proposition 3.1.1 will go through and we can prove that the component-wise WENO scheme, without the flux splitting or local characteristic decomposition, and with the special handling of the source terms described above, maintains exactly the C-property.

Next, we look at the situation when the local characteristic decomposition is invoked in the WENO procedure. 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 fixed, and neighboring point values of the grid functions needed for computing the numerical flux are projected to the local characteristic fields determined by R^{-1}. Therefore, (3.5) still holds, with $d = \left(hu, hu^2 + \frac{1}{2}gh^2\right)^T$ now being a vector grid function and a_k are 2×2 matrices depending nonlinearly on the smoothness indicators involving the grid function d. The key idea is still to use the difference operator D_d with $d = \left(hu, hu^2 + \frac{1}{2}gh^2\right)^T$ fixed, and apply it to approximate $\left(0, \frac{1}{2}gb^2\right)_x^T$ and $(0, b)_x^T$ in the source terms. In actual implementation, we can still absorb the approximation of $\left(0, \frac{1}{2}gb^2\right)_x^T$ into that of the flux derivative term $\left(hu, hu^2 + \frac{1}{2}gh^2\right)_x^T$ to save computational cost. The remaining arguments stay the same as above, and we can prove that the WENO scheme with a local characteristic decomposition, but without the flux splitting, and with the special handling of the source terms described above, maintains exactly the C-property.

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

$$f^{\pm}(U) = \frac{1}{2}\left[\begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix} \pm \alpha_i \begin{pmatrix} h \\ hu \end{pmatrix}\right] \tag{3.6}$$

for the i-th characteristic field, where $\alpha_i = max_u|\lambda_i(u)|$ with $\lambda_i(u)$ being the i-th eigenvalue of the Jacobian $f'(U)$, see [24, 40] for more details. We now make a modification to this flux splitting, by replacing $\pm\alpha_i \begin{pmatrix} h \\ hu \end{pmatrix}$ in (3.6) with $\pm\alpha_i \begin{pmatrix} h+b \\ hu \end{pmatrix}$.

The flux splitting (3.6) now becomes

$$f^{\pm}(U)=\frac{1}{2}\left[\begin{pmatrix} hu \\ hu^2+\frac{1}{2}gh^2 \end{pmatrix} \pm \alpha_i \begin{pmatrix} h+b \\ hu \end{pmatrix}\right]. \tag{3.7}$$

This modification is justified since b does not depend on the time t, hence the first equation in (1.3) can also be considered as an evolution equation for $h+b$ instead of for h. Similar techniques are used in the surface gradient method by Zhou et al. [51]. Our motivation for using $\pm\alpha_i \begin{pmatrix} h+b \\ hu \end{pmatrix}$ instead of the original $\pm\alpha_i \begin{pmatrix} h \\ hu \end{pmatrix}$ in the flux splitting, is that the former becomes a constant vector for the still water stationary solution (3.3). Thus for this still water stationary solution, by the consistency of the WENO approximation, the effect of this viscosity term $\pm\alpha_i \begin{pmatrix} h+b \\ hu \end{pmatrix}$ towards the approximation of $f(U)_x$ is zero. Clearly, (3.5) can represent the flux splitting WENO approximation, with a simple splitting $f^{\pm}(U)=\frac{1}{2}f(U)$, with $d=\left(hu, hu^2+\frac{1}{2}gh^2\right)^T$ being a vector grid function and a_k being 2×2 matrices depending nonlinearly on the smoothness indicators involving the grid function $f^{\pm}(U)$ in (3.7). What we have shown above is that, *for the still water stationary solution*, this is also the flux splitting WENO approximation with the modified Lax-Friedrich flux splitting (3.7). As before, the key idea now is to use the difference operator D_d in (3.5) with smoothness indicators, hence the nonlinear weights obtained from $f^{\pm}(U)$ in (3.7) fixed, and apply it to approximate $\left(0,\frac{1}{2}gb^2\right)^T_x$ and $(0,b)^T_x$ in the source terms. This amounts to split also the two derivatives in the source terms as

$$\begin{pmatrix} 0 \\ \frac{1}{2}gb^2 \end{pmatrix}_x = \frac{1}{2}\begin{pmatrix} 0 \\ \frac{1}{2}gb^2 \end{pmatrix}_x + \frac{1}{2}\begin{pmatrix} 0 \\ \frac{1}{2}gb^2 \end{pmatrix}_x, \qquad \begin{pmatrix} 0 \\ b \end{pmatrix}_x = \frac{1}{2}\begin{pmatrix} 0 \\ b \end{pmatrix}_x + \frac{1}{2}\begin{pmatrix} 0 \\ b \end{pmatrix}_x, \tag{3.8}$$

and apply the same flux split WENO procedure to approximate them, namely, one half of each source term is approximated by the difference operator with coefficients a_k obtained from the computation of f^+, and the remaining part by the difference operator with coefficients a_k obtained from the computation of f^-. In actual implementation, we can still absorb the approximation of $\left(0, \frac{1}{2}gb^2\right)_x^T$ into that of the flux derivative term $\left(hu, hu^2 + \frac{1}{2}gh^2\right)_x^T$ to save computational cost. The remaining arguments stay the same as above, and we can prove that the WENO scheme with a local characteristic decomposition and a flux splitting (3.7), and with the special handling of the source terms described above, maintains exactly the C-property.

We now summarize the complete procedure of the high order finite difference WENO-LF or WENO-LLF scheme with a local characteristic decomposition and a flux splitting, for solving the shallow water equation (1.3):

1. Split the source term and write the equation in the form (3.1).

2. Perform the usual WENO-LF or WENO-LLF approximation on the flux derivative $\begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix}_x$, with a modified flux splitting (3.7) and using the local characteristic decomposition.

3. Split the two derivative terms in the source terms on the right hand side of (3.1) as in (3.8), and perform the same WENO approximation which is used in step 2 above, using the local characteristic decomposition and the same nonlinear weights, to approximate these two derivative terms. In actual implementation, the approximation of the first derivative term $\begin{pmatrix} 0 \\ \frac{1}{2}gb^2 \end{pmatrix}_x$ can be absorbed into the approximation of the flux derivative $\begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix}_x$, to save computational cost.

4. Add up the residues and forward in time.

With this computational procedure, we have already shown above the exact C-property and high order accuracy.

Proposition 3.1.2 *The WENO-LF or WENO-LLF schemes as stated above can maintain the exact C-property and their original high order accuracy.*

Even though we have described the algorithm using the WENO-LF and WENO-LLF flux splitting form, the algorithm can clearly also be defined with the same properties for other finite difference WENO schemes in [24, 40], such as WENO-Roe and WENO-Roe with an entropy fix.

3.2 One dimensional numerical results

In this section we present numerical results of our fifth order finite difference WENO-LF scheme satisfying the exact C-property for the one dimensional shallow water equations (1.3). In all the examples, time discretization is by the classical fourth order Runge-Kutta method, and the CFL number is taken as 0.6, except for the accuracy tests where smaller time step is taken to ensure that spatial errors dominate. The gravitation constant g is taken as $9.812 m/s^2$.

3.2.1 Test for the exact C-property

The purpose of the first test problem is to verify that the scheme indeed maintains the exact C-property over a non-flat bottom. We choose two different functions for the bottom topography given by $(0 \le x \le 10)$:

$$b(x) = 5\, e^{-\frac{2}{5}(x-5)^2}, \tag{3.9}$$

which is smooth, and

$$b(x) = \begin{cases} 4 & \text{if } 4 \le x \le 8 \\ 0 & \text{otherwise,} \end{cases} \tag{3.10}$$

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 mesh points. The computed surface level $h + b$ and the bottom b for (3.9) are plotted in Figure 3.1. In order to demonstrate that the exact C-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 3.1 and 3.2 for the two bottom functions (3.9) and (3.10) and 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 exact C-property.

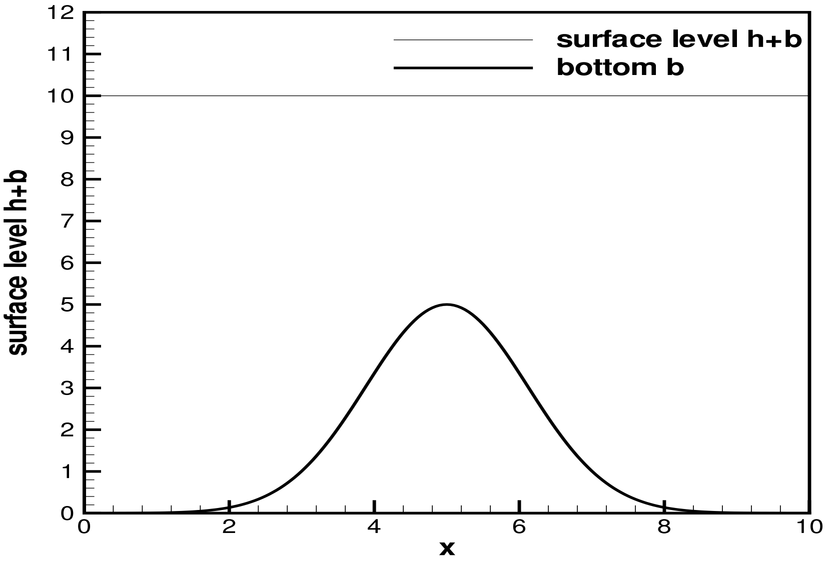


Figure 3.1: The surface level $h + b$ and the bottom b for the stationary flow over a smooth bump.

Table 3.1: L^1 and L^∞ errors for different precisions for the stationary solution with a smooth bottom (3.9).

precision	L^1 error		L^∞ error	
	h	hu	h	hu
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

Table 3.2: L^1 and L^∞ errors for different precisions for the stationary solution with a nonsmooth bottom (3.10).

precision	L^1 error		L^∞ error	
	h	hu	h	hu
single	2.28E-07	3.61E-06	1.91E-06	2.37E-05
double	1.14E-15	9.05E-15	3.55E-15	4.46E-14
quadruple	1.30E-33	1.40E-32	4.62E-33	5.64E-32

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 exact C-property.

3.2.2 Testing the orders of accuracy

In this example we will test the fifth order accuracy of our scheme for a smooth solution. There are some known exact solutions (in closed form) to the shallow water equation with non-flat bottom in the literature, e.g. [44], but these solutions have special properties, making the leading terms in the truncation errors of many schemes vanish, hence they are not generic test cases for accuracy. We have therefore chosen to use the following bottom function and initial conditions

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

with periodic boundary conditions. Since the exact solution is not known explicitly for this case, we use the same fifth order WENO scheme with $N = 25,600$ points to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. We compute up to $t = 0.1$ when the solution is still smooth (shocks develop later in time for this problem). Table 3.3 contains the L^1 errors and numerical orders of accuracy. We can clearly see that fifth order accuracy is achieved for this example. For comparison, we also list the L^1 errors and numerical orders of accuracy when the original fifth order WENO scheme [24] with the source term directly added to the residue as a point value at the grid x_i (hence not a C-property satisfying scheme) is used on the same problem. We can clearly see that the errors of the two schemes are comparable. For this problem, the solution is far from a still water stationary solution, hence our exact C-property satisfying WENO scheme is not expected to have an advantage in accuracy. Table 3.3 shows that it does not have a disadvantage either comparing with the traditional WENO scheme using point value treatment of source terms.

3.2.3 A small perturbation of a steady-state water

The following quasi-stationary test case was proposed by LeVeque [28]. 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.

The bottom topography consists of one hump:

$$b(x) = \begin{cases} 0.25(\cos(10\pi(x-1.5)) + 1) & \text{if } 1.4 \le x \le 1.6 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

Table 3.3: L^1 errors and numerical orders of accuracy for the example in Section 3.2.2. "Balanced WENO" refers to the WENO scheme with exact C-property and "original WENO" refers to the WENO scheme with the source terms directly added as point values at the grids.

No. of points	CFL	balanced WENO			
		h		hu	
		L^1 error	order	L^1 error	order
25	0.6	1.70E-002		1.06E-001	
50	0.6	2.17E-003	2.97	1.95E-002	2.45
100	0.6	3.33E-004	2.71	2.83E-003	2.78
200	0.6	2.36E-005	3.82	2.04E-004	3.80
400	0.6	9.67E-007	4.61	8.38E-006	4.61
800	0.6	3.38E-008	4.84	2.94E-007	4.83
1600	0.4	1.08E-009	4.97	9.34E-009	4.97
No. of points	CFL	original WENO			
		h		hu	
		L^1 error	order	L^1 error	order
25	0.6	1.96E-002		1.02E-001	
50	0.6	2.46E-003	2.99	1.82E-002	2.49
100	0.6	3.19E-004	2.95	2.79E-003	2.70
200	0.6	2.50E-005	3.67	2.18E-004	3.68
400	0.6	1.03E-006	4.59	9.07E-006	4.59
800	0.6	3.61E-008	4.84	3.15E-007	4.85
1600	0.4	1.15E-009	4.97	1.00E-008	4.98

The initial conditions are given with

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 1 - b(x) + \epsilon & \text{if } 1.1 \le x \le 1.2 \\ 1 - b(x) & \text{otherwise} \end{cases} \tag{3.12}$$

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 [28]. Both sets of initial conditions are shown in

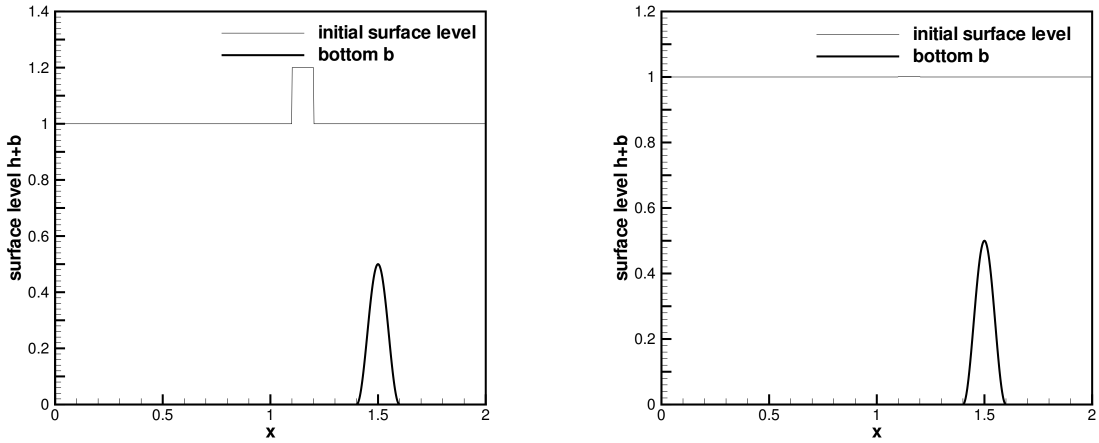


Figure 3.2: The initial surface level $h + b$ and the bottom b for a small perturbation of a steady-state water. Left: a big pulse ϵ=0.2; right: a small pulse ϵ=0.001.

Figure 3.2. The solution at time t=0.2s for the big pulse $\epsilon = 0.2$, obtained on a 200 cell uniform grid with simple transmissive boundary conditions, and compared with a 3000 cell solution, is shown in Figure 3.3. The one for the small pulse $\epsilon = 0.001$ is shown in Figure 3.4. For this small pulse problem, we take $\varepsilon = 10^{-9}$ in the WENO weight formula (2.2), such that it is smaller than the square of the perturbation. At this time, the downstream-traveling water pulse has already passed the bump. In the figures, we can clearly see that there are no spurious numerical oscillations, verifying the essentially non-oscillatory property of the modified WENO-LF scheme.

3.2.4 The dam breaking problem over a rectangular bump

In this example we 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 [44].

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leq 1500/8 \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

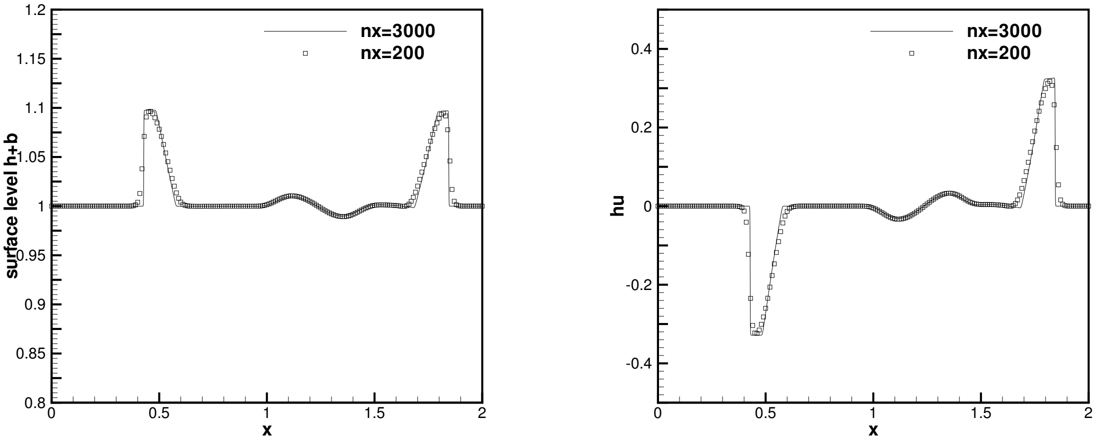


Figure 3.3: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

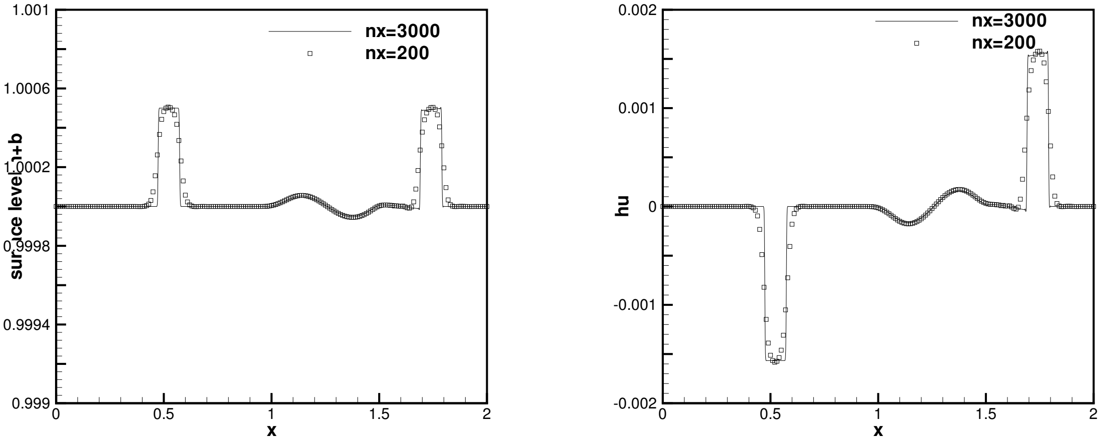


Figure 3.4: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

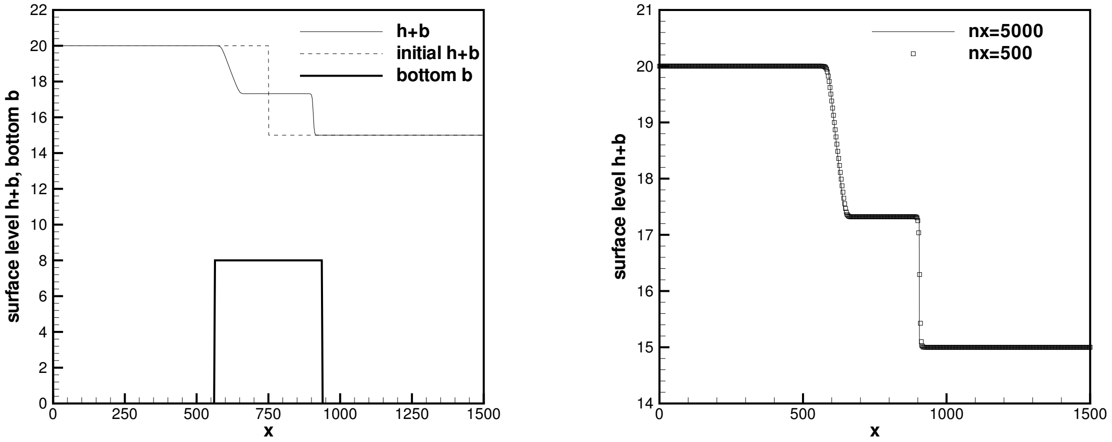


Figure 3.5: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 500 grid points, plotted with the initial condition and the bottom topography; Right: the numerical solution using 500 and 5000 grid points.

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

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 20 - b(x) & \text{if } x \leq 750 \\ 15 - b(x) & \text{otherwise} \end{cases} \tag{3.14}$$

The numerical results with 500 uniform points (and a comparison with the results using 5000 uniform points) are shown in Figures 3.5 and 3.6, with two different ending time t=15s and 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. Our scheme works well for this example, giving well resolved, non-oscillatory solutions using 500 points which agree with the converged results using 5000 points.

3.2.5 Steady flow over a hump

The purpose of this test case is to study the convergence in time towards steady flow over a bump. These are classical test problems for transcritical and subcritical flows, and they are widely used to test numerical schemes for shallow water equations. For

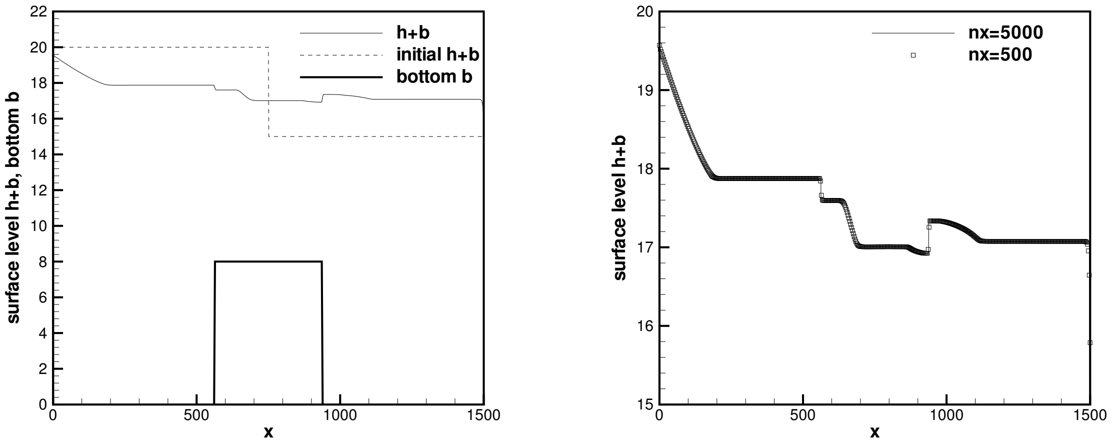


Figure 3.6: The surface level $h + b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 500 grid points, plotted with the initial condition and the bottom topography; Right: the numerical solution using 500 and 5000 grid points.

example, they have been considered by the *working group on dam break modeling* [17], and have been used as a test case in, e.g. [43].

The bottom function is given by:

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12 \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

for a channel of length $25m$. The initial conditions are taken as

$$h(x,0) = 0.5 - b(x) \quad \text{and} \quad u(x,0) = 0.$$

Depending on different boundary conditions, the flow can be subcritical or transcritical with or without a steady shock. The computational parameters common for all three cases are: uniform mesh size $\Delta x = 0.125$ m, ending time t= 200 s. Analytical solutions for the various cases are given in Goutal and Maurel [17].

a): Transcritical flow without a shock.

- upstream: The discharge hu=1.53 m^3/s is imposed.

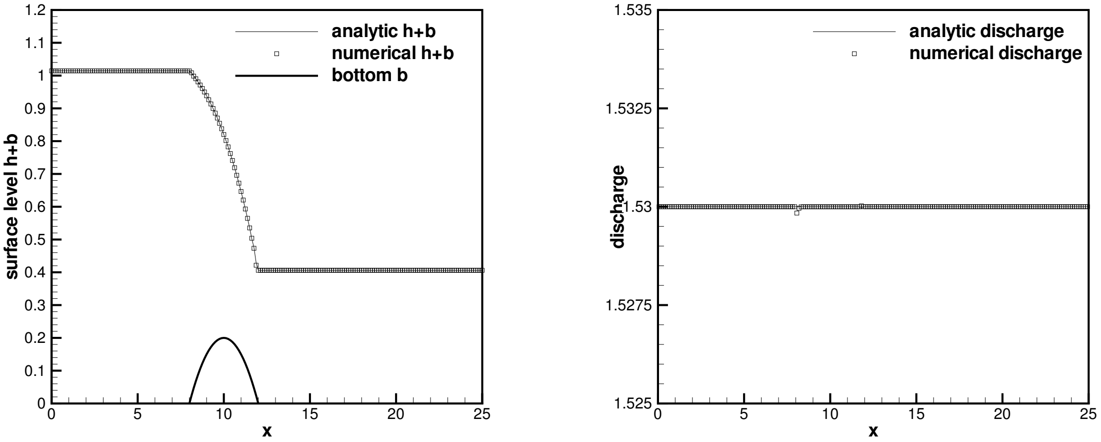


Figure 3.7: Steady transcritical flow over a bump without a shock. Left: the surface level $h+b$; right: the discharge hu as the numerical flux for the water height h.

- downstream: The water height h=0.66 m is imposed when the flow is subcritical.

The surface level h+b and the discharge hu, as the numerical flux for the water height h in equation (1.3), are plotted in Figure 3.7, which show very good agreement with the analytical solution. The correct capturing of the discharge hu is usually more difficult than the surface level $h+b$, as noticed by many authors. In Figure 3.8, we compare the pointwise errors of the numerical solutions obtained with 200 and 400 uniform grid points. The convergence history, measured by the L^1 norm of the residue, is given in Figure 3.9, left.

b): Transcritical flow with a shock.

- upstream: The discharge hu=0.18 m^3/s is imposed.

- downstream: The water height h=0.33 m is imposed.

In this case, the Froude number $Fr = u/\sqrt{gh}$ increases to a value larger than one above the bump, and then decreases to less than one. A stationary shock can appear on the surface. The surface level $h+b$ and the discharge hu, as the numerical flux

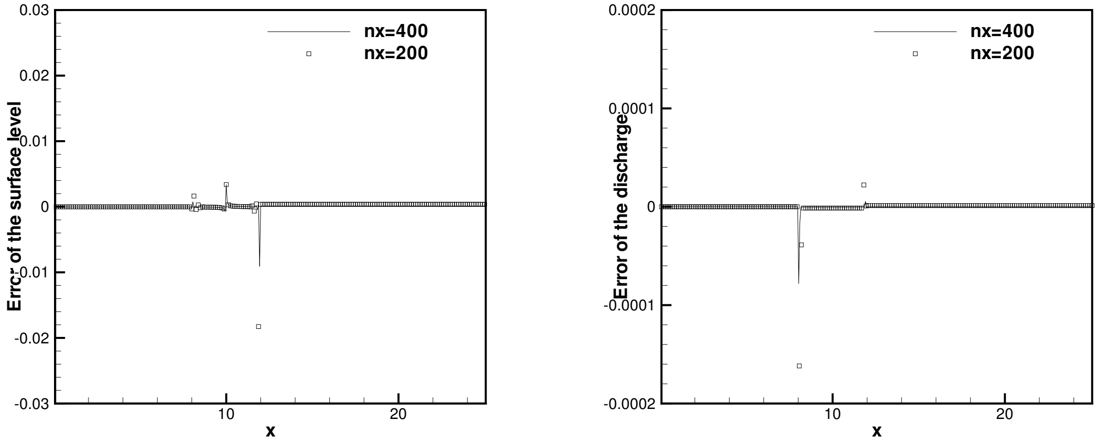


Figure 3.8: Steady transcritical flow over a bump without a shock. Pointwise error comparison between numerical solutions using 200 and 400 grid points. Left: the surface level $h+b$; right: the discharge hu as the numerical flux for the water height h.

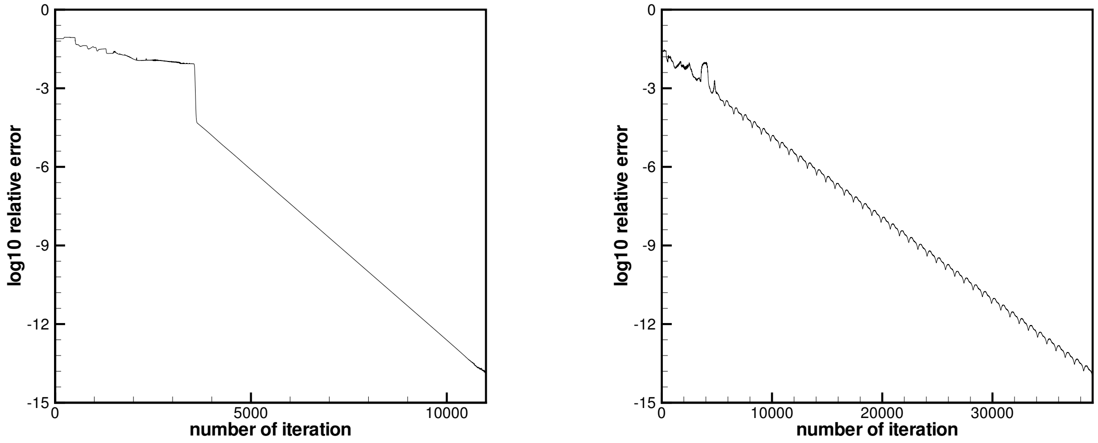


Figure 3.9: Convergence history in L^1 residue. Left: steady transcritical flow over a bump without a shock; right: steady transcritical flow over a bump with a shock.

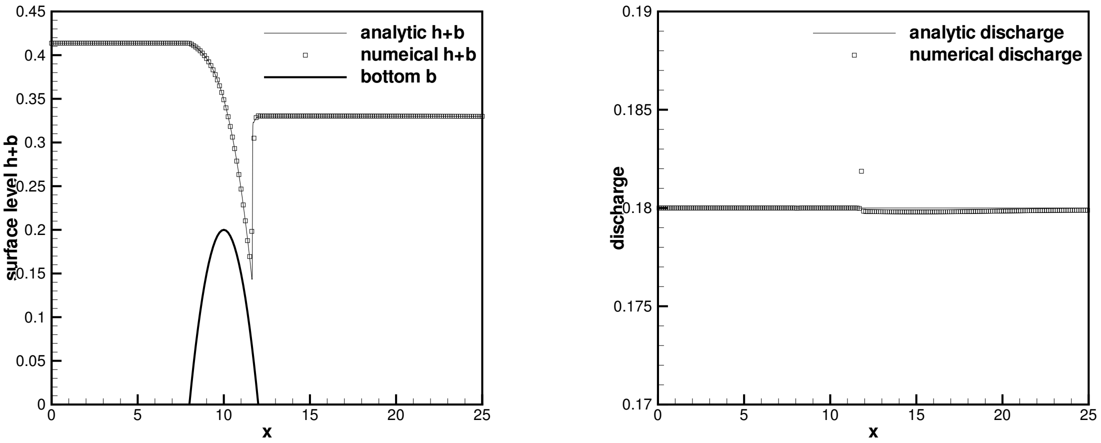


Figure 3.10: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

for the water height h in equation (1.3), are plotted in Figure 3.10, which show non-oscillatory results in good agreement with the analytical solution. In Figure 3.11, we compare the pointwise errors of the numerical solutions obtained with 200 and 400 uniform grid points. The convergence history, measured by the L^1 norm of the residue, is given in Figure 3.9, right.

c): Subcritical flow.

- upstream: The discharge hu=4.42 m^3/s is imposed.

- downstream: The water height h=2 m is imposed.

This is a subcritical flow. The surface level $h + b$ and the discharge hu, as the numerical flux for the water height h in equation (1.3), are plotted in Figure 3.12, which are in good agreement with the analytical solution. In Figure 3.13, we compare the pointwise errors of the numerical solutions obtained with 200 and 400 uniform grid points.

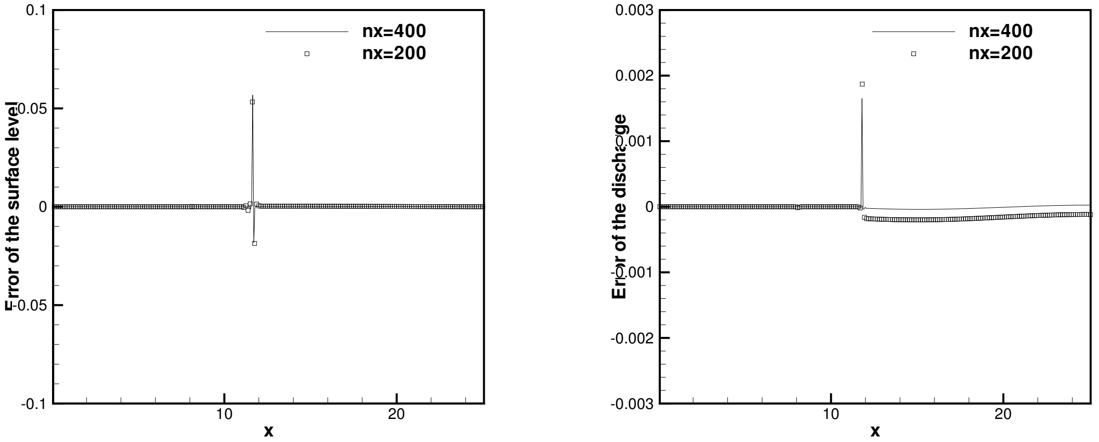


Figure 3.11: Steady transcritical flow over a bump with a shock. Pointwise error comparison between numerical solutions using 200 and 400 grid points. Left: the surface level $h+b$; right: the discharge hu as the numerical flux for the water height h.

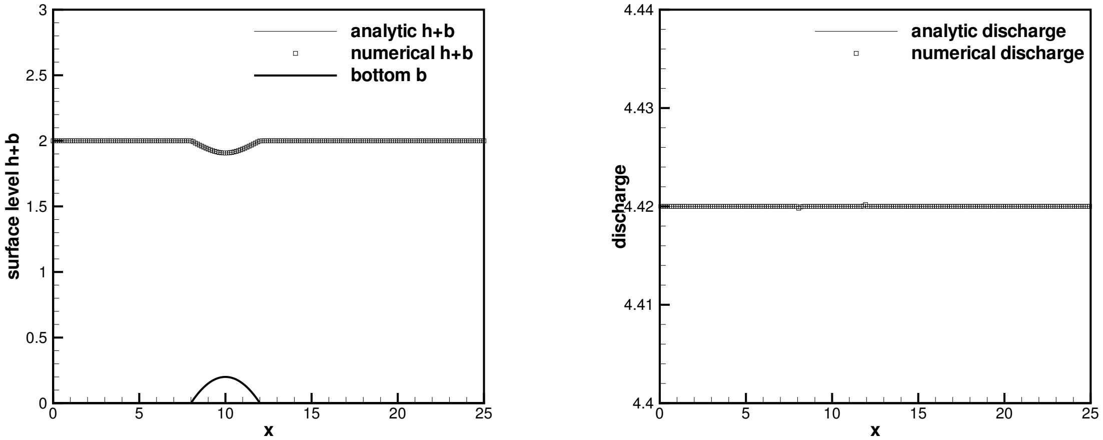


Figure 3.12: Steady subcritical flow over a bump. Left: the surface level $h+b$; right: the discharge hu as the numerical flux for the water height h.

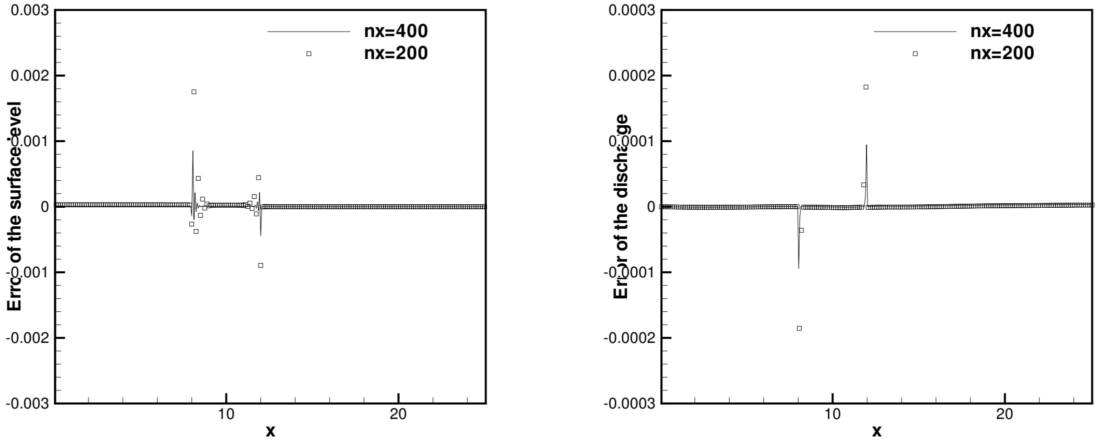


Figure 3.13: Steady subcritical flow over a bump. Pointwise error comparison between numerical solutions using 200 and 400 grid points. Left: the surface level $h+b$; right: the discharge hu as the numerical flux for the water height h.

3.2.6 The tidal wave flow

This example was used in [5], in which an almost exact solution (a very good asymptotically derived approximation) was given. We use this example to further test our scheme.

The bottom is defined by:

$$b(x) = 10 + \frac{40x}{L} + 10 \sin\left(\pi\left(\frac{4x}{L} - \frac{1}{2}\right)\right)$$

where L=14,000 m is the channel length. If we take the initial and boundary conditions as:

$$h(x,0) = 60.5 - b(x), \qquad (hu)(x,0) = 0$$

$$h(0,t) = 64.5 - 4\sin\left(\pi\left(\frac{4t}{86,400} + \frac{1}{2}\right)\right), \qquad (hu)(L,t) = 0,$$

a very accurate approximate solution, based on the asymptotic analysis, can be given by [5]

$$h(x,t) = 64.5 - b(x) - 4\sin\left(\pi\left(\frac{4t}{86,400} + \frac{1}{2}\right)\right)$$

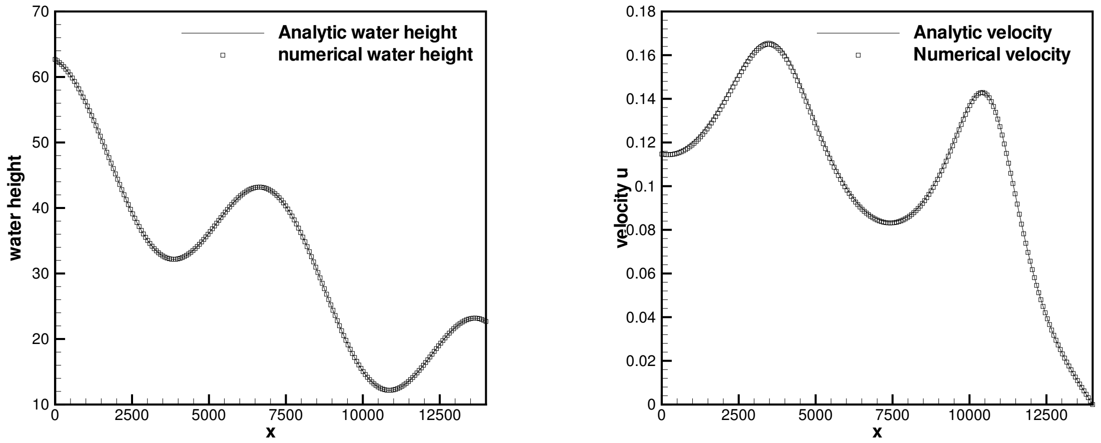


Figure 3.14: Numerical and analytic solutions for the tidal wave flow. Left: water height h; right: velocity u.

and

$$(hu)(x,t) = \frac{(x-L)\pi}{5400}\cos\left(\pi\left(\frac{4t}{86,400}+\frac{1}{2}\right)\right).$$

We use a uniform mesh size Δx =70 m. A comparison of the numerical and analytical results at $t = 7552.13$ s is shown in Figure 3.14. Their agreements are very good.

3.3 Two dimensional shallow water systems

A major advantage of the high order finite difference WENO schemes is that it is straightforward to extend them to multiple space dimensions, by simply approximating each spatial derivative along the relevant coordinate. It turns out that it is also straightforward to extend the high order finite difference WENO schemes with the exact C-property developed in Section 3.1 to two dimensions. The shallow water system in two space dimensions takes the form:

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0 \\ (hu)_t + \left(hu^2 + \dfrac{1}{2}gh^2\right)_x + (huv)_y = -ghb_x \\ (hv)_t + (huv)_x + \left(hv^2 + \dfrac{1}{2}gh^2\right)_y = -ghb_y \end{cases} \tag{3.16}$$

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

Finite difference WENO schemes are very easy to be extended to multidimensional cases. The conservative approximation to the derivative from point values is as simple in multi dimensions as in one dimension. In fact, for fixed j, if we take $w(x) = f(u(x, y_j))$, then we only need to perform the one dimensional WENO approximation to $w(x)$ to obtain an approximation to $w'(x_i) = f_x(u(x_i, y_j))$. See again [24, 40] for more details.

The source term is again split as in the one dimensional case

$$-ghb_x = \left(\frac{1}{2}gb^2\right)_x - g(h+b)b_x, \qquad -ghb_y = \left(\frac{1}{2}gb^2\right)_y - g(h+b)b_y,$$

and the one dimension procedure described in Section 3.1 is followed in each of the x and y directions. The residues are then summed up and advancement in time is again by a Runge-Kutta method.

All results proved in the one dimensional case, such as high order accuracy and the exact C-property, are still valid in the two dimensional case.

We now give numerical experiment results for the exact C-property satisfying fifth order WENO-LF scheme in two dimensions. Similar to the one dimensional case, we use the classical fourth order Runge-Kutta time discretization and a CFL number 0.6, except for the accuracy test problem where smaller time step is taken to guarantee that spatial errors dominate.

3.3.1 Test for the exact C-property in two dimensions

This example is used to check that our scheme indeed maintains the exact C-property over a non-flat bottom. The two-dimensional hump

$$b(x,y) = 0.8e^{-50((x-0.5)^2+(y-0.5)^2)}, \qquad x,y \in [0,1] \tag{3.17}$$

is chosen to be the bottom. $h(x,y,0) = 1 - b(x,y)$ is the initial depth of the water. Initial velocity is set to be zero. This surface should remain flat. The computation is performed to $t = 0.1$ using single, double and quadruple precisions with a 100×100 uniform mesh. Table 3.4 contains the L^1 errors for the water height h (which is not a constant function) and the discharges hu and hv. We can clearly see that the L^1 errors are at the level of round-off errors for different precisions, verifying the exact C-property.

Table 3.4: L^1 errors for different precisions for the stationary solution in Section 3.3.1.

	L^1 error		
precision	h	hu	hv
single	2.18E-08	2.32E-07	2.32E-07
double	7.71E-17	9.36E-16	9.36E-16
quadruple	7.64E-34	9.33E-34	9.33E-34

3.3.2 Testing the orders of accuracy

In this example we check the numerical orders of accuracy when the WENO schemes are applied to the following two dimensional problem. The bottom topography and the initial data are given by:

$$b(x,y) = \sin(2\pi x) + \cos(2\pi y), \qquad h(x,y,0) = 10 + e^{\sin(2\pi x)}\cos(2\pi y),$$

$$(hu)(x,y,0) = \sin(\cos(2\pi x))\sin(2\pi y), \qquad (hv)(x,y,0) = \cos(2\pi x)\cos(\sin(2\pi y))$$

defined over a unit square, with periodic boundary conditions. The terminal time is taken as t=0.05 to avoid the appearance of shocks in the solution. Since the exact solution is not known explicitly for this case, we use the same fifth order WENO scheme with an extremely refined mesh consisting of 1600 × 1600 grid points to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. Table 3.5 contains the L^1 errors and orders of accuracy. We can clearly see that fifth order accuracy is achieved in this two dimensional test case.

Table 3.5: L^1 errors and numerical orders of accuracy for the example in Section 3.3.2.

Number of cells	CFL	h		hu		hv	
		L^1 error	order	L^1 error	order	L^1 error	order
25	0.6	1.08E-002		3.23E-002		8.92E-002	
50	0.6	1.30E-003	3.06	2.47E-003	3.70	1.19E-002	2.90
100	0.6	1.06E-004	3.61	1.47E-004	4.07	9.06E-004	3.72
200	0.4	4.82E-006	4.46	6.25E-006	4.56	3.98E-005	4.51
400	0.3	1.79E-007	4.75	2.31E-007	4.76	1.41E-006	4.82
800	0.2	6.30E-009	4.83	8.19E-009	4.82	4.70E-008	4.91

3.3.3 A small perturbation of a two dimensional steady-state water

This is a classical example to show the capability of the proposed scheme for the perturbation of the stationary state, given by LeVeque [28]. It is analogous to the test done previously in Section 3.2.3 in one dimension.

We solve the system in the rectangular domain $[0,2] \times [0,1]$. The bottom topography is an isolated elliptical shaped hump:

$$b(x,y) = 0.8\,e^{-5(x-0.9)^2-50(y-0.5)^2}. \tag{3.18}$$

The surface is initially given by:

$$\begin{aligned} h(x,y,0) &= \begin{cases} 1-b(x,y)+0.01 & \text{if } 0.05 \le x \le 0.15 \\ 1-b(x,y) & \text{otherwise} \end{cases} \\ hu(x,y,0) &= hv(x,y,0) = 0 \end{aligned} \tag{3.19}$$

So the surface is almost flat except for $0.05 \le x \le 0.15$, where h is perturbed upward by 0.01. Figure 3.15 displays the right-going disturbance as it propagates past the hump, on two different uniform meshes with 200×100 points and 600×300 points for comparison. The surface level $h+b$ is presented at different time. The results indicate that our scheme can resolve the complex small features of the flow very well.

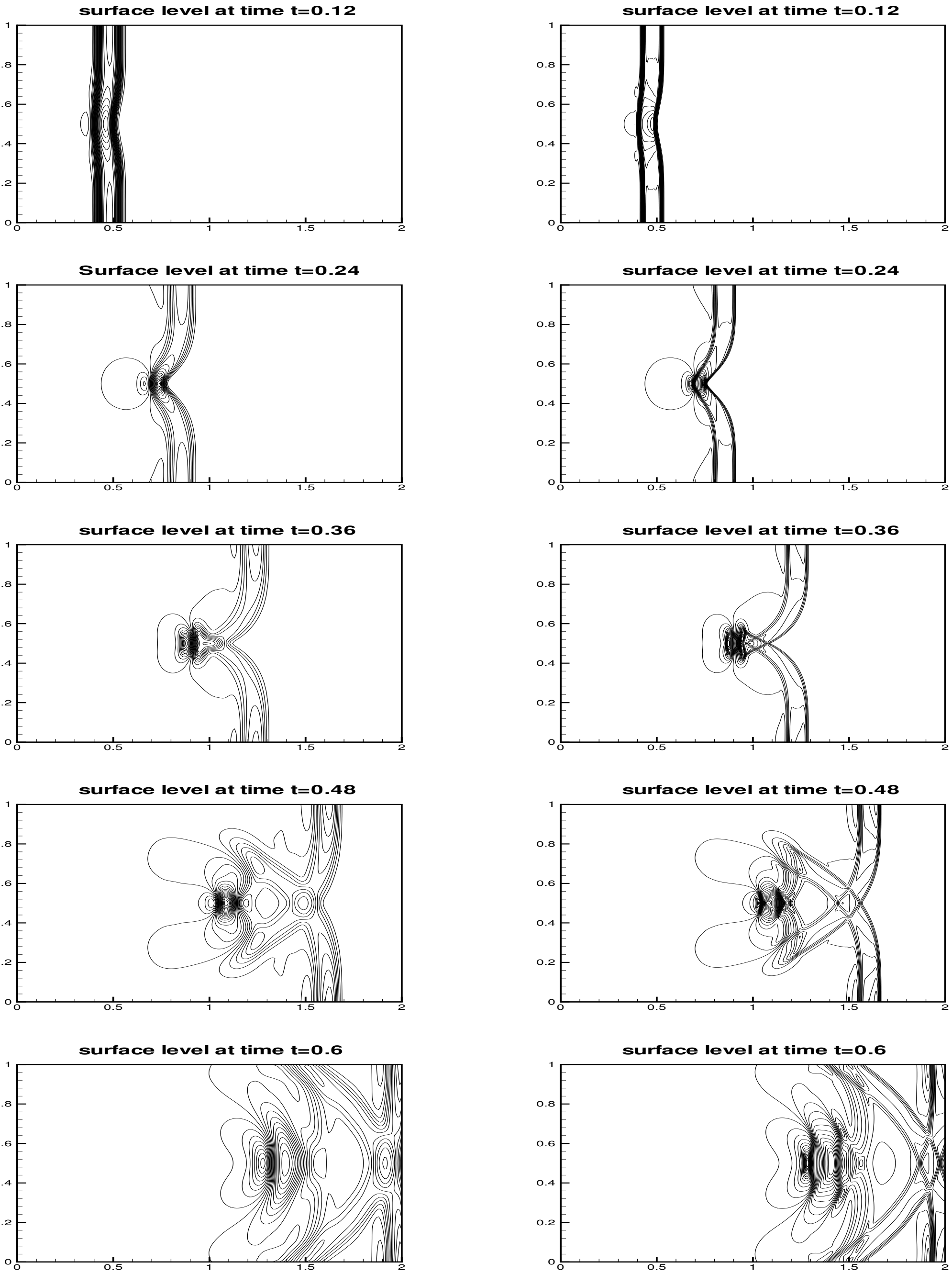


Figure 3.15: The contours of the surface level $h + b$ for the problem in Section 3.3.3. 30 uniformly spaced contour lines. From top to bottom: at time $t = 0.12$ from 0.999703 to 1.00629; at time $t = 0.24$ from 0.994836 to 1.01604; at time $t = 0.36$ from 0.988582 to 1.0117; at time $t = 0.48$ from 0.990344 to 1.00497; and at time $t = 0.6$ from 0.995065 to 1.0056. Left: results with a 200×100 uniform mesh. Right: results with a 600×300 uniform mesh.

Chapter 4

High Order Finite Difference Well-balanced WENO Schemes for a Class of Hyperbolic Systems with Source Terms

In this chapter, we extend the idea designed in the previous chapter to a general class of balance laws with separable source terms, and design well balanced high order finite difference WENO scheme for all balance laws falling into this category. In section 4.1, we describe the class of balance laws under consideration and develop well balanced finite difference WENO schemes for such balance laws. In section 4.2, we give several examples in applications which fall into the category of balance laws discussed in section 4.1, and show selective numerical results to demonstrate the behavior of our schemes.

4.1 A general class of balance laws

The main idea in Chapter Three to design a high order finite difference WENO scheme for the shallow water equation is to decompose its source term into a sum of two 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. In this section, we generalize this idea to a class of general balance laws (1.2). We first consider the case that (1.2) 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 (1.2):

$$f(u,x)_x = g(u,x). \tag{4.1}$$

We make two assumptions on the equation (1.2) and the steady state solution u of (4.1) that we are interested to preserve exactly:

Assumption 4.1.1 *The steady state solution u of (4.1) that we are interested to preserve satisfies*

$$a(u,x) = \textit{constant} \tag{4.2}$$

for a known function $a(u,x)$.

Assumption 4.1.2 *The source term $g(u,x)$ in (1.2) can be decomposed as*

$$g(u,x) = \sum_i s_i(a(u,x))\, t_i'(x) \tag{4.3}$$

for some functions s_i and t_i.

Before proceeding further, let us comment on Assumption 4.1.1. We consider a special case of (1.2):

$$u_t + f(u)_x = g(u)\, z'(x) \tag{4.4}$$

i.e. when the flux f does not depend explicitly on x and the source term $g(u,x)$ in (1.2) is separable as a product of a function in u and a function in x. Notice that the case $g(u,x)$=$g(u)$ falls into this category with $z(x) = x$. The steady state solution of the equation (4.4) is given by:

$$f(u)_x = g(u)\, z'(x).$$

Clearly

$$\int \frac{f'(u)}{g(u)} du = \int \frac{f'(u)}{g(u)} u_x dx = \int \frac{f(u)_x}{g(u)} dx = \int z'(x) dx = z(x) + constant.$$

Hence we have

$$a(u,x) \equiv b(u) + z(x) = constant \tag{4.5}$$

if we denote $b(u) = -\int \frac{f'(u)}{g(u)} du$. This is an example of (4.2).

We would like to preserve exactly the steady state solutions u which satisfy Assumption 4.1.1, for a balance law (1.2) with a source term satisfying Assumption 4.1.2. Following the ideas in Chapter Three, we will first consider a linear scheme with an identical finite difference operator for the flux derivative and the derivatives in the decomposed source terms. As usual, 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 (1.2) with a source term given by (4.3) 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

$$D_0(f(u,x)) - \sum_i s_i(a(u,x))\, D_i(t_i(x)),$$

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

$$D_0 = D_1 = \cdots = D \tag{4.6}$$

for the steady state solution. For such linear schemes we have

Proposition 4.1.3 *For the balance law (1.2) with its source term given by (4.3), linear schemes with (4.6) for the steady state solutions satisfying (4.2) can preserve these steady state solutions exactly.*

Proof. For the steady state solution u satisfying (4.2), the truncation error for such linear schemes with (4.6) reduces to

$$D\left(f(u,x)\right) - \sum_i s_i(a(u,x))\, D\left(t_i(x)\right) = D\left(f(u,x) - \sum_i s_i(a(u,x))\, t_i(x)\right) \tag{4.7}$$

where the linearity of D and the fact that $a(u,x) = \textit{constant}$ for the steady state solution u are used. Clearly, for such steady state solution u,

$$\begin{aligned} & \frac{d}{dx}\left(f(u,x) - \sum_i s_i(a(u,x))\, t_i(x)\right) \\ = \quad & f(u,x)_x - \sum_i s_i(a(u,x))\, t_i'(x) = f(u,x)_x - g(u,x) = 0, \end{aligned}$$

that is, $f(u,x) - \sum_i s_i(a(u,x))\, t_i(x)$ is a constant. Hence the truncation error (4.7) is 0 for any consistent finite difference operator D. This finishes the proof. □

We now consider high order nonlinear finite difference WENO schemes [24, 4], in which the nonlinearity comes from the nonlinear weights and the smooth indicators. We follow the procedures described in Chapter Three for the shallow water equations, to treat the general balance laws (1.2) to obtain well balanced high order finite difference WENO schemes.

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 [24]). We notice that the WENO approximation to d_x where $d = f(u,x)$ can be eventually written out as

$$d_x|_{x=x_j} \approx \sum_{k=-r}^{r} a_k d_{k+j} \equiv D_d(d)_j \tag{4.8}$$

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. As explained in [46], the key idea now is to use the finite difference operator D_d with $d = f(u, x)$ *fixed*, and apply it to approximate $t_i'(x)$ in the source terms. 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 $d = f(u, x)$ is fixed, is a high order accurate *linear* approximation to the first derivative for any grid function. Therefore the proof for Propositions 2.3 will go through 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 [24]. Here the flux $f(u, x)$ is written as a sum of $f^+(u, x)$ and $f^-(u, x)$, defined by

$$f^{\pm}(u, x) = \frac{1}{2} \left[f(u, x) \pm \alpha u\right] \tag{4.9}$$

where $\alpha = max_u \left|\frac{\partial f(u,x)}{\partial u}\right|$ with the maximum being taken over either a local region (WENO-LLF) or a global region (WENO-LF), see [24, 40] for more details. We now make a modification to this flux splitting, by replacing $\pm\alpha u$ in (4.9) with $\pm\alpha \, \mathrm{sign}\left(\frac{\partial a(u,x)}{\partial u}\right) a(u, x)$. We would need to assume here that $\frac{\partial a(u,x)}{\partial u}$ does not change sign. The constant α should be suitably adjusted by the size of $\frac{\partial a(u,x)}{\partial u}$ in order to maintain enough artificial viscosity. The term $a(u, x)$ can also be replaced by $p(a(u, x))$ for any function p, whose choice should be such that $p(a(u, x))$ 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 $f(u, x) = f^+(u, x) + f^-(u, x)$. For the steady state solution satisfying (4.2), the artificial viscosity term $\pm\alpha\,\mathrm{sign}\left(\frac{\partial a(u,x)}{\partial u}\right) a(u,x)$ (or $\pm\alpha\,\mathrm{sign}\left(\frac{\partial p(a(u,x))}{\partial u}\right) p(a(u,x))$) 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 $f(u,x)_x$ is zero. The flux splitting WENO approximation in this situation becomes simply $f^{\pm}(u,x) = \frac{1}{2}f(u,x)$, hence the steady state solution is preserved as before, if we simply split the derivatives in the source term as:

$$t_i'(x) = \frac{1}{2}t_i'(x) + \frac{1}{2}t_i'(x), \tag{4.10}$$

and apply the same flux splitting WENO procedure to approximate them with the nonlinear coefficients a_k coming from the WENO approximations to $f^{\pm}(u,x)$ respectively. This will guarantee (4.6). We have thus proved that

Proposition 4.1.4 *The WENO-Roe, WENO-LF and WENO-LLF schemes as implemented above are exact for steady state solutions satisfying (4.2) 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 (4.2). Typically, for a system with m equations, we would have m relationships in the form of (4.2):

$$a_1(u,x) = constant, \qquad \cdots \qquad a_m(u,x) = constant \tag{4.11}$$

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 (4.3), where s_i could be arbitrary functions of $a_1(u,x), \cdots, a_m(u,x)$, 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. Examples of such systems will be given in next section. 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 [46] for more details.

4.2 Applications

In this section we give several examples from applications which fall into the category of balance laws considered in the previous section, and present well balanced high order finite difference WENO schemes for them. Due to page limitation, only selected numerical results are shown to give a glimpse of how these methods work. In the numerical tests, time discretization is by the classical fourth order Runge-Kutta method, and the CFL number is taken as 0.6.

4.2.1 Shallow water equations

The shallow water equations have wide applications in ocean and hydraulic engineering and river, reservoir, and open channel flows, among others. We consider the system with a geometrical source term due to the bottom topology. In one space dimension, the equations take the form

$$\begin{cases} h_t + (hu)_x = 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x = -ghb_x, \end{cases} \tag{4.12}$$

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

We are interested in preserving the still water solution, which satisfies (4.11) in the form

$$a_1 \equiv h + b = constant, \qquad a_2 \equiv u = 0.$$

The first component of the source term is 0. A decomposition of the second component of the source term in the form of (4.3) is

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

i.e. $s_1 = s_1(a_1) = -g\,(h+b)$, $s_2 = \frac{1}{2}g$, $t_1(x) = b(x)$, and $t_2(x) = b^2(x)$.

More details of the high order finite difference WENO scheme applied to this system, and extensive numerical results, can be found in [46].

4.2.2 Elastic wave equation

We consider the propagation of compressional waves [3, 45] in an one-dimensional elastic rod with a given media density $\rho(x)$. The equation of motion in a Lagrangian frame are given by the balance laws:

$$\begin{cases} (\rho\varepsilon)_t + (-\rho u)_x = -u\dfrac{d\rho}{dx} \\ (\rho u)_t + (-\sigma)_x = 0, \end{cases} \tag{4.13}$$

where $\varepsilon = \varepsilon(x,t)$ is the strain, $u = u(x,t)$ is the velocity and σ is a given stress-strain relationship $\sigma(\varepsilon, x)$. The equation of linear acoustics can be obtained from the elasticity problem if the stress-strain relationship is linear,

$$\sigma(\varepsilon, x) = K(x)\,\varepsilon$$

where $K(x)$ is the given bulk modulus of compressibility.

The steady state we are interested to preserve for this problem is characterized by

$$a_1 \equiv \sigma(\varepsilon, x) = constant, \qquad a_2 \equiv u = constant$$

which is of the form (4.11). The second component of the source term is 0. The first component of the source term is already in the form of (4.3) with $s_1 = s_1(a_2) = -u$ and $t_1 = \rho(x)$.

We now show two numerical examples to demonstrate the fifth order well balanced finite difference WENO scheme for (4.13). The first example, from [45], is to test the fifth order accuracy for smooth solutions, for which we take the initial conditions as

$$\rho\varepsilon(x,0) = \frac{-1 - 1.5\, e^{-(8x)^2}}{(1 - 0.5\sin(\pi x))^2}, \qquad u(x,0) = 0$$

with the density $\rho(x)$ and bulk modulus of compressibility $K(x)$ given by:

$$\rho(x) = \frac{1}{1 - 0.5\sin(\pi x)}, \qquad K(x) = 1 - 0.5\sin(\pi x).$$

The computational domain is [-1,1] and periodic boundary condition is used. The exact solution is unknown in this case, hence we use the same fifth order well balanced WENO scheme with $N = 5120$ grid points to compute a reference solution and use this reference solution as the exact solution in computing the numerical errors at $t = 0.1s$. Table 4.1 contains the L^1 errors and numerical orders of accuracy. We can clearly see that fifth order accuracy is achieved for this example.

Next, we present the numerical result for a linear acoustic test [3]. The properties of the media are given by

$$c(x) = \sqrt{\frac{K(x)}{\rho(x)}} = 1 + 0.5\sin(10\pi x), \qquad Z(x) = \rho(x)c(x) = 1 + 0.25\cos(10\pi x)$$

Table 4.1: L^1 errors and numerical orders of accuracy for the example in section 4.2.2.

Number of points	balanced WENO			
	$\rho\varepsilon$		ρu	
	L^1 error	order	L^1 error	order
20	2.33E-002		2.80E-002	
40	3.21E-003	2.86	3.50E-003	3.00
80	3.75E-004	3.10	2.30E-004	3.93
160	1.59E-005	4.56	1.10E-005	4.38
320	5.20E-007	4.93	3.92E-007	4.81
640	1.65E-008	4.97	1.25E-008	4.97

and are shown in Figure 4.1. The initial conditions are given by

$$\rho\varepsilon(x,0) = \begin{cases} \dfrac{-1.75 + 0.75\cos(10\pi x)}{c^2(x)}, & \text{if } 0.4 < x < 0.6 \\ \dfrac{-1}{c^2(x)}, & \text{otherwise} \end{cases}, \qquad u(x,0) = 0.$$

It is a test case where the impedance $Z(x)$ and hence the eigenvectors are both spatially varying. We perform the computation with 200 uniform cells, with the ending time $t = 0.4s$. An “exact” reference solution is computed with the same scheme over a 2000 grid point uniform cells. The simulation results are shown in Figure 4.2. The numerical resolution shows very good agreement with the “exact” reference solution.

4.2.3 Chemosensitive movement

Originated from biology, chemosensitive movement [21, 15] is a process by which cells change their direction reacting to the presence of a chemical substance, approaching chemically favorable environments and avoiding unfavorable ones. Hyperbolic

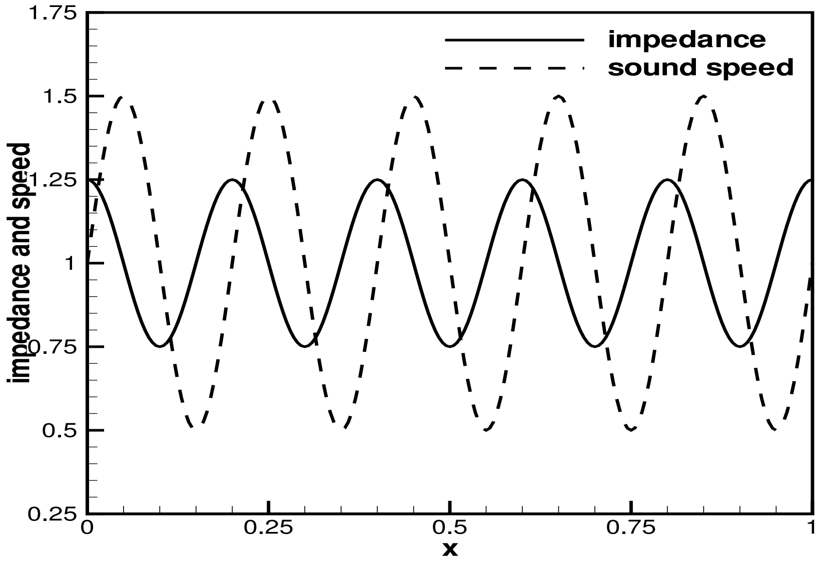


Figure 4.1: The impedance $Z(x)$ and the sound speed $c(x)$ for the smooth media.

models for chemotaxis are recently introduced [21] and take the form

$$\begin{cases} n_t + (nu)_x = 0 \\ (nu)_t + (nu^2 + n)_x = n\chi'(c)\dfrac{\partial c}{\partial x} - \sigma nu \end{cases} \tag{4.14}$$

where the chemical concentration $c = c(x, t)$ is given by the parabolic equation

$$\frac{\partial c}{\partial t} - D_c \triangle c = n - c.$$

Here, $n(x, t)$ is the cell density, $nu(x, t)$ is the population flux and σ is the friction coefficient. In [15], a well balanced WENO scheme is constructed based on a different approach, which can maintain the steady state solutions with zero population flux to the size of a small parameter ε in the nonlinear WENO weights. Here we construct well balanced WENO schemes based on the framework in section 4.1, which does

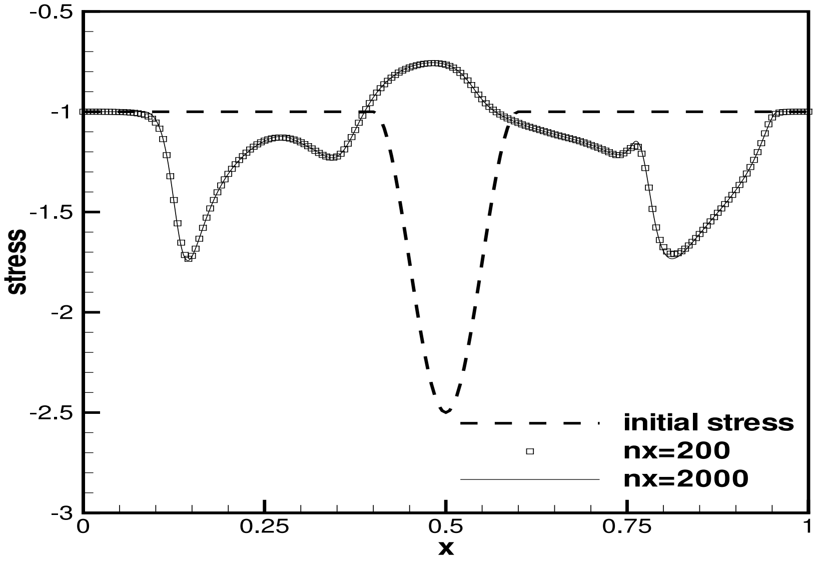


Figure 4.2: The numerical (symbols) and the "exact" reference (solid line) stress $\sigma(x)$ at time $t = 0.4s$.

not have this restriction.

We would like to preserve the steady state solution to (4.14) with a zero population flux, which satisfies

$$n\chi'(c)c_x - n_x = 0, \qquad nu = 0. \tag{4.15}$$

where $c = c(x)$ does not depend on t in steady state. The first equality above does not seem to be of the form (4.11). However, (4.15) is equivalent to

$$a_1 \equiv \log(n) - \chi(c) = constant, \qquad a_2 \equiv nu = 0,$$

which is clearly in the form of (4.11). The first component of the source term is 0.

A decomposition of the second component of the source term in the form of (4.3) is

$$n\chi'(c)\frac{\partial c}{\partial x} - \sigma nu = e^{\log(n)-\chi(c)}\,\frac{d}{dx}e^{\chi(c)} - \sigma nu$$

i.e. $s_1 = s_1(a_1) = e^{\log(n)-\chi(c)}$, $s_2 = s_2(a_2) = \sigma nu$, $t_1(x) = e^{\chi(c(x))}$, and $t_2(x) = x$.

We now show two numerical examples to demonstrate the fifth order well balanced finite difference WENO scheme for (4.14). The first example is to test the well balancedness property of the scheme. We take the initial conditions as

$$n(x,0) = \frac{1}{10}(1 + c(x)), \qquad nu(x,0) = 0.$$

with

$$c(x) = \begin{cases} 1 & \text{if } |x| \leq 0.5 \\ 0.125 & \text{otherwise} \end{cases}, \qquad \chi(c) = \log(1+c), \qquad \sigma = 1.$$

The initial condition is a steady state solution which should be exactly preserved. We compute the solution until $t = 2.0$s using $N = 500$ uniform mesh points. In order to demonstrate that the steady state 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 errors for the cell density n (note: n in this case is a discontinuous function!) and the population flux nu in Table 4.2 for these different precisions. We can clearly see that the L^1 errors are at the level of round-off errors for different precisions, verifying the steady state conservation.

The second example is to test the fifth order accuracy for smooth solutions, for which we take the initial conditions as

$$n(x,0) = 1 + 0.2\,\cos(\pi x), \qquad u(x,0) = 0, \qquad x \in [-1,1]$$

with

$$c(x) = e^{-16x^2}, \qquad \chi(c) = \log(1+c), \qquad \sigma = 0$$

Table 4.2: L^1 errors for different precisions for the stationary solution in section 4.2.3.

precision	L^1 error	
	n	nu
single	7.34E-07	3.16E-07
double	1.02E-15	3.96E-16
quadruple	9.13E-34	2.32E-34

with a periodic boundary condition. Since the exact solution is not known explicitly for this problem, we use the same fifth order WENO scheme with $N = 5120$ points to compute a reference solution and treat it as the exact solution when computing the numerical errors. Final time $t = 0.5s$ is used to avoid the development of shocks. Table 4.3 contains the L^1 errors and numerical orders of accuracy. We can clearly see that fifth order accuracy is achieved for this example.

Table 4.3: L^1 errors and numerical orders of accuracy for the example in section 4.2.3.

Number of points	balanced WENO			
	n		nu	
	L^1 error	order	L^1 error	order
20	1.02E-002		5.99E-003	
40	1.05E-003	3.29	7.70E-004	2.96
80	1.31E-004	3.00	1.07E-004	2.85
160	6.57E-006	4.32	5.49E-006	4.29
320	2.44E-007	4.75	2.06E-007	4.73
640	7.58E-009	5.01	6.43E-009	5.00

4.2.4 Nozzle flow

In this subsection we consider the balance laws for a quasi one-dimensional nozzle flow [16]. The governing equations for the quasi-one-dimensional unsteady flow through a duct of varying cross-section can be written in conservation form as:

$$\begin{cases} (\rho A)_t + (\rho u A)_x = 0 \\ (\rho u A)_t + \left((\rho u^2 + p)A\right)_x = pA'(x) \\ (EA)_t + ((E+p)uA)_x = 0 \end{cases} \tag{4.16}$$

where the quantities ρ, u, p and $E = \frac{1}{2}\rho u^2 + \frac{p}{\gamma - 1}$ represent the density, velocity, pressure and total energy, respectively. $A = A(x)$ denotes the area of the cross section. γ is the ratio of specific heats.

As in [16], we are interested in preserving the steady state solution

$$\rho(x,t) = \bar{\rho}(x), \qquad p(x,t) = \bar{p}, \qquad \text{and} \qquad u(x,t) = 0 \tag{4.17}$$

where $\bar{\rho}(x)$ is an arbitrary function in x and $\bar{p}$ is a constant. The second condition in (4.17)

$$a_1 \equiv p = \bar{p}$$

is of the form (4.11). The first and third components of the source term are 0. The second component of the source term is already in the form of (4.3) with $s_1 = s_1(a_1) = p$ and $t_1 = A(x)$.

We now show two numerical examples to demonstrate the fifth order well balanced finite difference WENO scheme for (4.16). The first example is to test the fifth order accuracy for smooth solutions, for which we take the cross section area and the initial conditions as

$$A(x) = 1 + \sin^2(\pi x), \qquad \rho(x) = \cos(\sin(2\pi x)), \qquad u(x) = 0, \qquad E(x) = e^{\sin(2\pi x)}$$

with periodic boundary conditions. As before, we compute a reference solution using the same fifth order WENO scheme with $N = 10240$ points. Final time is chosen as $t = 0.25s$ when the solution is still smooth. Table 4.4 contains the L^1 errors and numerical orders of accuracy. We can clearly see that fifth order accuracy is achieved for this example.

Table 4.4: L^1 errors and numerical orders of accuracy for the example in section 4.2.4.

Number of points	balanced WENO					
	ρA		$\rho u A$		EA	
	L^1 error	order	L^1 error	order	L^1 error	order
20	6.13E-003		3.90E-003		4.58E-003	
40	2.14E-003	1.52	9.46E-004	2.04	8.90E-004	2.38
80	2.10E-004	3.35	9.72E-005	3.28	8.49E-005	3.37
160	1.01E-005	4.38	4.79E-006	4.34	4.11E-006	4.37
320	3.44E-007	4.88	1.60E-007	4.91	1.40E-007	4.87
640	1.04E-008	5.04	5.08E-009	4.98	4.29E-009	5.03

The purpose of the second test case is to study the convergence in time towards steady flow. Proposed by Anderson in [1], it is concerned with a convergent-divergent nozzle flow with a parabolic area distribution, which is given by

$$A(x) = 1 + 2.2(x - 1.5)^2, \qquad 0 \leq x \leq 3. \tag{4.18}$$

The shape of this section is illustrated in Figure 4.3.

The initial conditions are taken as

$$\rho(x,0) = 1, \qquad u(x,0) = 0 \qquad \text{and} \qquad p(x,0) = 1.$$

The boundary conditions are taken as 1 bar of pressure at the left, 0.6784 bar of pressure at the right, and 300°K of temperature at both ends. A shock is established inside the pipe, and the exact solution for this, a steady state, can be easily calculated. In this case, the Froude number $Fr = u/c$ increases to a value larger than one, and then decreases to less than one.

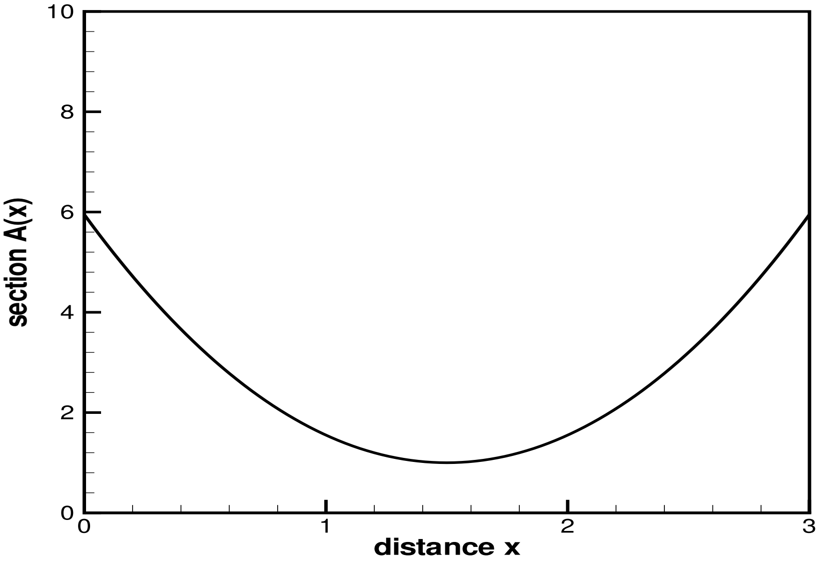


Figure 4.3: The shape of a convergent-divergent nozzle.

The computation is performed using $N = 100$ points. The pressure $p(x)$ is plotted in Figure 4.4, which shows very good agreement with the analytical solution. The numerical resolution is very good without oscillations, verifying the essentially non-oscillatory property of the modified WENO-LF scheme.

4.2.5 Two phase flow

The dynamics of fluids consisting of several fluid components is of great interest in a wide range of physical flows. In this subsection we are interested in a flow model for the compressible 2-velocity 2-pressure system [37, 26], which is suitable to describe liquid suspensions and bubbly flows. The balance equations are written for

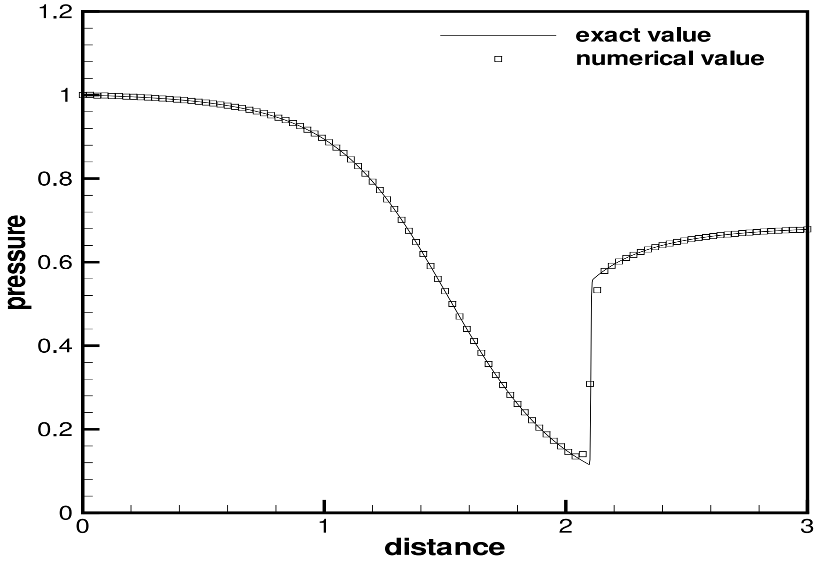


Figure 4.4: Steady state pressure for the nozzle flow.

the individual phases:

$$\begin{pmatrix} a_g\rho_g \\ a_g\rho_g u_g \\ a_g E_g \\ a_l\rho_l \\ a_l\rho_l u_l \\ a_l E_l \end{pmatrix}_t + \begin{pmatrix} a_g\rho_g u_g \\ a_g(\rho_g u_g^2 + p_g) \\ u_g a_g(E_g + p_g) \\ a_l\rho_l u_l \\ a_l(\rho_l u_l^2 + p_l) \\ u_l a_l(E_l + p_l) \end{pmatrix}_x = \tag{4.19}$$

$$\begin{pmatrix} 0 \\ p_i(a_g)_x + \quad \lambda(u_l - u_g) + \qquad\qquad a_g\rho_g g \\ u_i p_i(a_g)_x + \lambda u_i(u_l - u_g) + \mu p_i(p_l - p_g) + a_g\rho_g u_g g \\ 0 \\ p_i(a_l)_x - \quad \lambda(u_l - u_g) + \qquad\qquad a_l\rho_l g \\ u_i p_i(a_l)_x - \lambda u_i(u_l - u_g) - \mu p_i(p_l - p_g) + a_l\rho_l u_l g \end{pmatrix},$$

coupled with an additional equation for the volume fraction

$$(a_g)_t + u_i(a_g)_x = -\mu(p_l - p_g) \tag{4.20}$$

and the algebraic relation for the volume fractions

$$a_g + a_l = 1.$$

Here, a_k, $k \in \{l, g\}$ is the volume fraction of the k-th phase, and ρ_k, u_k, $E_k = \frac{1}{2}\rho_k u_k^2 + \frac{p_k}{\gamma_k - 1}$ denote its density, velocity and energy, respectively. λ is a velocity relaxation parameter and μ is a pressure relaxation parameter. The closure equations for the interface pressure p_i and the interface velocity u_i are

$$p_i = a_g p_g + a_l p_l, \qquad u_i = \frac{a_g \rho_g u_g + a_l \rho_l u_l}{a_g \rho_g + a_l \rho_l}.$$

If the gravitation effect is ignored, one stationary solution for (4.19) is given by:

$$\rho_g = \bar{\rho}_1(x), \qquad \rho_l = \bar{\rho}_2(x), \qquad u_g = u_l = 0, \qquad p_g = p_l = \bar{p} \tag{4.21}$$

where $\bar{\rho}_1(x)$ and $\bar{\rho}_2(x)$ are arbitrary functions of x and $\bar{p}$ is a constant. We would like to preserve this steady state solution exactly. (4.21) can clearly be written in the form of (4.11):

$$a_1 \equiv p_i = \bar{p}, \qquad a_2 \equiv u_l - u_g = 0, \qquad a_3 \equiv p_l - p_g = 0, \qquad a_4 \equiv u_i = 0$$

and hence the source terms are already in the form of (4.3). For example, the second component of the source term is of the form (4.3) with $s_1 = s_1(a_1) = p_i$, $t_1 = a_g(x)$, $s_2 = s_2(a_2) = \lambda(u_l - u_g)$, $t_2 = x$; the third component of the source term is of the form (4.3) with $s_1 = s_1(a_4, a_1) = u_i p_i$, $t_1 = a_g(x)$, $s_2 = s_2(a_4, a_2) = \lambda u_i(u_l - u_g)$, $t_2 = x$, $s_3 = s_3(a_1, a_3) = \mu p_i(p_l - p_g)$, $t_3 = x$; etc.

Chapter 5

High Order Well-balanced Finite Volume WENO Schemes and RKDG Methods for a Class of Hyperbolic Systems with Source Terms

In this chapter, we design well balanced finite volume WENO and RKDG finite element methods for the same class of balance laws as in Chapter Four. In Section 5.1, we describe the class of balance laws under consideration and develop well-balanced finite volume WENO schemes, which at the same time are genuinely high order accurate for the general solutions. The well-balanced generalization of the RKDG schemes is presented in Section 5.2. In Section 5.3, we give several examples in applications which fall into the category of balance laws discussed in Section 5.1, and show selective numerical results in one and two dimensions to demonstrate the behavior of our well balanced finite volume WENO schemes and RKDG schemes, verifying high order accuracy, the well balanced property, and good resolution for smooth and discontinuous solutions.

5.1 Construction of well balanced finite volume WENO schemes

In this section, we design a genuine high order finite volume WENO scheme for a class of general balance laws (1.1). We will concentrate our discussion on the one dimensional case (1.2). Generalization to the multi-dimensional case (1.1) can be done in some situations, for example the cases discussed in [46, 47]; we present the details for the two dimensional shallow water equations in Section 5.3.2.

Our main objective is to preserve certain steady state solutions while maintaining high order accuracy for general solutions. The main idea in Chapter Three and Four 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 (1.2) 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 (1.2):

$$f(u,x)_x = g(u,x). \tag{5.1}$$

As in [47], we make some assumptions on the equation (1.2) and the steady state solution u of (4.1) that we are interested to preserve exactly:

Assumption 5.1.1 *The steady state solution u of (5.1) that we are interested to* *preserve satisfies*

$$a(u,x) \equiv \frac{u+p(x)}{q(x)} = constant \tag{5.2}$$

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

Assumption 5.1.2 *The source term $g(u,x)$ in (1.2) can be decomposed as*

$$g(u,x) = \sum_j s_j(a(u,x))\, t_j'(x) \tag{5.3}$$

for some known functions s_j and t_j.

Note that Assumption 5.1.1 given here is more restrictive than that in Chapter Four. This is due to the additional difficulties related to the finite volume formulation.

We would like to preserve exactly the steady state solutions u which satisfy Assumption 5.1.1, for a balance law (1.2) with a source term satisfying Assumption 5.1.2.

Now let us describe the details of the algorithm. We consider the semi-discrete formulation of the balance law

$$\frac{d}{dt}\bar{u}_i(t) = -\frac{1}{\triangle x_i}(f(u(x_{i+\frac{1}{2}}),t) - f(u(x_{i-\frac{1}{2}}),t)) + \frac{1}{\triangle x_i}\int_{I_i} g(u,x)dx. \tag{5.4}$$

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 explained in Section 2.2, which are high order accurate approximations to the exact value $u(x_{i+\frac{1}{2}})$. We use the smoothness indicators β_r to measure the smoothness of the variable u. The WENO reconstruction can be eventually written out as

$$u^+_{i+\frac{1}{2}} = \sum_{k=-r+1}^{r} w_k \bar{u}_{i+k} \equiv S^+_{\bar{u}}(\bar{u})_i, \qquad u^-_{i+\frac{1}{2}} = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{u}_{i+k} \equiv S^-_{\bar{u}}(\bar{u})_i. \tag{5.5}$$

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 $\bar{u}$, following (2.10)-(2.11). 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 $\bar{u}$. Once again, our purpose is to find a high order finite volume scheme for a class of conservation laws which can preserve the steady state solution (5.2). The 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

$$\begin{aligned} p^+_{i+\frac{1}{2}} &= S^+_{\bar{u}}(\bar{p})_i = \sum_{k=-r+1}^{r} w_k \bar{p}_{i+k}, & p^-_{i+\frac{1}{2}} &= S^-_{\bar{u}}(\bar{p})_i = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{p}_{i+k} \\ q^+_{i+\frac{1}{2}} &= S^+_{\bar{u}}(\bar{q})_i = \sum_{k=-r+1}^{r} w_k \bar{q}_{i+k}, & q^-_{i+\frac{1}{2}} &= S^-_{\bar{u}}(\bar{q})_i = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{q}_{i+k}. \end{aligned} \tag{5.6}$$

With the reconstructed values $p^{\pm}_{i+\frac{1}{2}}$ and $q^{\pm}_{i+\frac{1}{2}}$, we obtain the pointwise value of a(u,x) by $a(u,x)^{\pm}_{i+\frac{1}{2}} = \frac{u^{\pm}_{i+\frac{1}{2}} + p^{\pm}_{i+\frac{1}{2}}}{q^{\pm}_{i+\frac{1}{2}}}$. 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, $a(u,x)^{\pm}_{i+\frac{1}{2}}$ is a high order approximation to $a(u(x_{i+\frac{1}{2}}), x_{i+\frac{1}{2}})$.

Now assume that u is the steady state solution satisfying (5.2), namely

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

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,

$$a(u,x)^{\pm}_{i+\frac{1}{2}} = c \tag{5.7}$$

for the same constant c.

Clearly, for a steady state solution u satisfying Assumptions 5.1.1 and 5.1.2,

$$\begin{aligned} & \frac{d}{dx}\left(f(u,x) - \sum_j s_j(a(u,x))\, t_j(x) \right) \\ = \quad & f(u,x)_x - \sum_j s_j(a(u,x))\, t'_j(x) = f(u,x)_x - g(u,x) = 0. \end{aligned}$$

Therefore, $f(u,x) - \sum_j s_j(a(u,x))\, t_j(x)$ 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

$$f(u^{\pm}_{i+\frac{1}{2}}) - \sum_j s_j(a(u,x)^{\pm}_{i+\frac{1}{2}})\, (t_j)^{\pm}_{i+\frac{1}{2}} = constant \tag{5.8}$$

for a steady state solution u satisfying Assumptions 5.1.1 and 5.1.2. In the applications stated later in Section 5.3, we will specify the choices of $(t_j)^{\pm}_{i+\frac{1}{2}}$ in each case.

The integral of the source term takes the form

$$\int_{I_i} g(u,x)dx = \sum_j \int_{I_i} s_j(a(u,x))t_j'(x)dx.$$

We need to decompose it further in the following way in order to obtain a well-balanced scheme

$$\begin{aligned}
&\sum_j \int_{I_i} s_j(a(u,x))t_j'(x)dx \\
= &\sum_j \left(\frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right) \int_{I_i} t_j'(x)dx \right. \\
&\left. + \int_{I_i} \left(s_j(a(u,x)) - \frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right) \right) t_j'(x)dx \right) \\
= &\sum_j \left(\frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right) (t_j(x_{i+\frac{1}{2}}) - t_j(x_{i-\frac{1}{2}})) \right. \\
&\left. + \int_{I_i} \left(s_j(a(u,x)) - \frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right) \right) t_j'(x)dx \right). \quad (5.9)
\end{aligned}$$

The purpose of this decomposition is to ensure the balance with the flux difference term on the right hand side of (5.4), see the proof of Proposition 5.1.3 below. We remark that $\frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right)$ can also be replaced by $s_j\left(\frac{\overline{u + p(x)}}{\overline{q(x)}} \right)$ where as usual the overbar denotes the cell average over the cell I_i, which could be used when there is a singularity at the boundary, for example, in the application in Section 5.3.5.

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

$$\frac{d}{dt}\bar{u}_i(t) = -\frac{1}{\triangle x_i}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}) + \frac{1}{\triangle x_i}\hat{g}_i, \quad (5.10)$$

with

$$\hat{g}_i = \sum_j \left(\frac{1}{2}\left(s_j(a(u,x)_{i-\frac{1}{2}}^{+}) + s_j(a(u,x)_{i+\frac{1}{2}}^{-}) \right) \left((\hat{t}_j)_{i+\frac{1}{2}} - (\hat{t}_j)_{i-\frac{1}{2}} \right) + g_{i,j} \right) \quad (5.11)$$

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 $g_{i,j}$ is any high order approximation to the integral

$$\int_{I_i} \left(s_j(a(u,x)) - \frac{1}{2} \left(s_j(a(u,x)^+_{i-\frac{1}{2}}) + s_j(a(u,x)^-_{i+\frac{1}{2}}) \right) \right) t_j'(x)\, dx. \tag{5.12}$$

Comparing with (5.9), it is clear that $\hat{g}_i$ is a high order approximation to the source term in (5.4).

The numerical flux $\hat{f}_{i+\frac{1}{2}}$ is defined by a monotone flux such as the Lax-Friedrichs flux (2.8)

$$F(u^-_{i+\frac{1}{2}}, u^+_{i+\frac{1}{2}}) = \frac{1}{2} \left[f(u^-_{i+\frac{1}{2}}) + f(u^+_{i+\frac{1}{2}}) - \alpha(u^+_{i+\frac{1}{2}} - u^-_{i+\frac{1}{2}}) \right]. \tag{5.13}$$

We need to make a modification to this flux, by replacing $\alpha(u^+_{i+\frac{1}{2}} - u^-_{i+\frac{1}{2}})$ in (5.13) with $\alpha\,\mathrm{sign}(q(x))(a(u,x)^+_{i+\frac{1}{2}} - a(u,x)^-_{i+\frac{1}{2}})$. The numerical flux now becomes

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2} \left[f(u^-_{i+\frac{1}{2}}) + f(u^+_{i+\frac{1}{2}}) - \alpha\,\mathrm{sign}(q(x))(a(u,x)^+_{i+\frac{1}{2}} - a(u,x)^-_{i+\frac{1}{2}}) \right]. \tag{5.14}$$

We would need to assume here that $q(x)$ in (5.2) 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 (5.2),

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

because of (5.7). Hence, the effect of these viscosity terms becomes zero and the numerical flux turns out to be in a simple form

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2} \left[f(u^-_{i+\frac{1}{2}}) + f(u^+_{i+\frac{1}{2}}) \right]. \tag{5.15}$$

Following this, we treat the approximation $(\hat{t}_j)_{i+\frac{1}{2}}$ in (5.11) in a similar way:

$$(\hat{t}_j)_{i+\frac{1}{2}} = \frac{1}{2}\left[(t_j)^-_{i+\frac{1}{2}} + (t_j)^+_{i+\frac{1}{2}}\right] \tag{5.16}$$

where, as mentioned before, $(t_j)^\pm_{i+\frac{1}{2}}$ are high order approximations to $t_j(x_{i+\frac{1}{2}})$ satisfying (5.8). Note that we implement (5.16) for the general case, not only for the steady solution. There is no viscosity term in the source term, compared with the numerical flux (5.14).

For the remaining source term $g_{i,j}$, 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, non-oscillatory property is maintained.

We now formulate the preservation of the steady state solution (5.2) by our numerical scheme.

Proposition 5.1.3 *The WENO-LF schemes as implemented above with (5.10), (5.11), (5.14) and (5.16) are exact for steady state solutions satisfying (5.2) and can maintain the original high order accuracy for general solutions.*

Proof. The high order accuracy is straightforward to observe. We only prove the well balanced property here. First, for the steady state solution $a(u,x) = c$ for some constant c, the reconstructed values $a(u,x)^\pm_{i-\frac{1}{2}}$ are also equal to the same constant c, see (5.7). Hence, we notice that the source term $g_{i,j}$, which is a high order numerical approximation of the integral in (5.12) by a Gauss quadrature, is simply zero since $a(u,x)$ is equal to $a(u,x)^\pm_{i-\frac{1}{2}}$ at each Gauss point. Furthermore, in this case the flux terms take the form (5.15) and (5.16). Therefore, the truncation error reduces to

$$\begin{aligned} & -\hat{f}_{i+\frac{1}{2}} + \hat{f}_{i-\frac{1}{2}} + \sum_j \frac{1}{2}\left(s_j(a(u,x)^+_{i-\frac{1}{2}}) + s_j(a(u,x)^-_{i+\frac{1}{2}})\right)\left((\hat{t}_j)_{i+\frac{1}{2}} - (\hat{t}_j)_{i-\frac{1}{2}}\right) \\ = & -\hat{f}_{i+\frac{1}{2}} + \sum_j s_j(c)(\hat{t}_j)_{i+\frac{1}{2}} + \hat{f}_{i-\frac{1}{2}} - \sum_j s_j(c)(\hat{t}_j)_{i-\frac{1}{2}} \\ = & \ 0 \end{aligned}$$

where we have used (5.7) for the first equality, and (5.8), (5.15) and (5.16) for the second equality. This finishes the proof. □

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 (5.2). Typically, for a system with m equations, we would have m relationships in the form of (5.2):

$$a_1(u,x) = constant, \qquad \cdots \qquad a_m(u,x) = constant \tag{5.17}$$

for the steady state solutions that we would like to preserve exactly. Here we require that, for the steady state solution (5.17), $a_j(u,x) = \frac{\sum_k b_k u_k + p_j(x)}{q_j(x)}$, where $u = (u_1, \cdots, u_m)$, b_k are arbitrary constants, and $p_j(x)$ and $q_j(x)$ are arbitrary known functions of x. We would then still aim for decomposing each component of the source term in the form of (5.3), where s_j could be arbitrary functions of $a_1(u,x), \cdots, a_m(u,x)$, and the functions s_j and t_j 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. Examples of such systems will be given in Section 5.3. 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 reconstructing the point value 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.

5.2 Construction of well balanced discontinuous Galerkin schemes

In this section, we generalize the idea used in Section 5.1 to RKDG schemes. A well-balanced high order RKDG scheme will be designed for a class of conservation laws satisfying Assumptions 5.1.1 and 5.1.2. 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. We start with the description in the scalar case.

Consider now the equation (1.2). Following the description in Section 2.3, the semi-discrete DG scheme for (1.2) is

$$\begin{aligned}\int_{I_j} \partial_t u_h(x,t) v_h(x) dx - \int_{I_j} f(u_h(x,t)) \partial_x v_h(x) dx + \hat{f}_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) \\ = \int_{I_j} g(u_h(x,t),t) v_h(x) dx \qquad (5.18)\end{aligned}$$

$$\int_{I_j} u_h(x,0) v_h(x) dx = \int_{I_j} u_0(x) v_h(x) dx. \qquad (5.19)$$

First, we define a high order approximation $a_h(u_h, x) = \frac{u_h+p_h}{q_h}$ to $a(u_h, x)$, where p_h and q_h are L^2 projections of p and q into V_h, see (5.19) for such a projection. Now assume that u is the steady state solution satisfying (5.2), 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

$$a_h(u_h, x) = \frac{u_h(x) + p_h(x)}{q_h(x)} = c$$

for the same constant c.

For such steady state solution u satisfying Assumptions 5.1.1 and 5.1.2, we have

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

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

$$f(u_h(x)) - \sum_j s_j(a_h(u_h(x),x))(t_j)_h(x) = constant \qquad (5.20)$$

for all x. The construction of $(t_j)_h$ follows a similar procedure as that for the construction of $(t_j)^{\pm}_{i+\frac{1}{2}}$ for the finite volume well balanced scheme in Section 5.1. We will describe in detail the construction of $(t_j)_h$ for each application case in Section 5.3.

Similar to the decomposition of the source term in the well balanced finite volume schemes (5.9), we decompose the integral of the source term on the right hand side of (5.18) as:

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

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

$$
\begin{aligned}
&\sum_j \left(\frac{1}{2}\left(s_j(a_h(u_h,x)^+_{i-\frac{1}{2}}) + s_j(a_h(u_h,x)^-_{i+\frac{1}{2}})\right)\right. \\
&\left((\hat{t}_j)_{h,i+\frac{1}{2}}v_h(x^-_{i+\frac{1}{2}}) - (\hat{t}_j)_{h,i-\frac{1}{2}}v_h(x^+_{i-\frac{1}{2}}) - \int_{I_i}(t_j)_h(x)v_h'(x)dx\right) \\
&\left. + \int_{I_i}\left(s_j(a_h(u_h,x)) - \frac{1}{2}\left(s_j(a_h(u_h,x)^+_{i-\frac{1}{2}}) + s_j(a_h(u_h,x)^-_{i+\frac{1}{2}})\right)\right) t_j'(x)v_h dx\right)
\end{aligned}
\tag{5.21}
$$

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

To deal with the "hat" terms (numerical fluxes and approximations to $t_j(x_{i+\frac{1}{2}})$), we use the relation between the finite volume schemes and the DG schemes. If we simply take the test function v_h in the DG scheme as the constant function 1, we obtain the evolution of the cell averages similar to that for a finite volume scheme. We have already explained the construction of the "hat" terms for well balanced finite volume schemes. Here we simply copy those definitions (5.14) and (5.16) from

Section 5.1 without further explanation

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

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

A combination of the above equations gives the final version of our well-balanced high order RKDG schemes if one more modification on the slope limiter procedure is provided. 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 $a(u,x) = constant$. 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 $a_h(u_h,x) = constant$. We therefore propose to first check whether any limiting is needed based on the function $a_h(u_h,x)$ in each Runge-Kutta stage, where the cell averages of $a_h(u_h,x)$ (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 $a_h(u_h,x)$. When the limiting procedure is implemented this way, if the steady state u satisfying $a(u,x) = constant$ is reached, no cell will be flagged as requiring limiting since $a_h(u_h,x)$ is equal to the same constant, hence u_h will not be limited and therefore the steady state is preserved.

It is easy to compute the remaining integrals because u_h, $(t_j)_h$ and v_h are all piecewise polynomials in the space V_h. This finishes the description of the RKDG schemes. We can clearly observe that the accuracy is maintained. We also state below the proposition claiming the exact preservation of the steady state solution (5.2). The proof is similar to that of Proposition 5.1.3 for the finite volume schemes, and is therefore omitted.

Proposition 5.2.1 *The RKDG schemes as stated above are exact for steady state solutions satisfying (5.2) and can maintain the original high order accuracy for gen-*

eral 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.

5.3 Applications

In this section we give several examples from applications which fall into the category of balance laws considered in the previous sections, and present well balanced high order finite volume WENO and discontinuous Galerkin schemes for them. Due to page limitation, only selected numerical results are shown to give a glimpse of how these methods work. Fifth order finite volume WENO scheme and third order finite element RKDG scheme are implemented as examples. In all numerical tests, time discretization is by the third order TVD Runge-Kutta method in [41]. For finite volume 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 [39, 11] for its definition) is taken as 0 in most numerical examples, unless otherwise stated.

5.3.1 One dimensional shallow water equations

The shallow water equations have wide applications in ocean and hydraulic engineering and river, reservoir, and open channel flows, among others. We consider the system with a geometrical source term due to the bottom topology. In one space dimension, the equations take the form

$$\begin{cases} h_t + (hu)_x = 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x = -ghb_x, \end{cases} \tag{5.22}$$

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

The steady state solution we are interested in preserving satisfies (5.17) in the form

$$a_1 \equiv h + b = constant, \qquad a_2 \equiv u = 0.$$

The first component of the source term is 0. A decomposition of the second component of the source term in the form of (5.3) is

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

i.e. $s_1 = s_1(a_1) = -g\,(h+b)$, $s_2 = \frac{1}{2}g$, $t_1(x) = b(x)$, and $t_2(x) = b^2(x)$. For the finite volume schemes, we apply the WENO reconstruction to the function $(b(x),0)^T$, with coefficients computed from $(h,hu)^T$, to obtain $b^{\pm}_{i+\frac{1}{2}}$. We 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 $h+b=c$, $u=0$ is reached, we have

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

which is a constant. A similar manipulation leads to

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

For the RKDG method, we define

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

where $b_h(x)$ is the L^2 projection of $b(x)$ to the finite element space V_h. A similar manipulation as in the finite volume case leads to

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

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

Next, we provide numerical results to demonstrate the good properties of the well balanced finite volume WENO and finite element RKDG schemes when applied to the one dimensional shallow water equations. The gravitation constant g is taken as $9.812 m/s^2$ during the computation.

5.3.1.1 Test for the exact C-property

The purpose of the first test problem is to verify that the schemes indeed maintain the exact C-property over a non-flat bottom. We choose two different functions for the bottom topography given by $(0 \le x \le 10)$:

$$b(x) = 5\, e^{-\frac{2}{5}(x-5)^2}, \tag{5.23}$$

which is smooth, and

$$b(x) = \begin{cases} 4 & \text{if } 4 \le x \le 8 \\ 0 & \text{otherwise,} \end{cases} \tag{5.24}$$

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 exact C-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 5.1 and 5.2 for the two bottom functions (5.23) and (5.24) and different precisions. For the RKDG 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 exact C-property.

Table 5.1: L^1 and L^∞ errors for different precisions for the stationary solution with a smooth bottom (5.23).

	precision	L^1 error		L^∞ error	
		h	hu	h	hu
FV	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
RKDG	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

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 exact C-property.

5.3.1.2 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

Table 5.2: L^1 and L^∞ errors for different precisions for the stationary solution with a nonsmooth bottom (5.24).

		L^1 error		L^∞ error	
	precision	h	hu	h	hu
FV	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
RKDG	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

bottom function and initial conditions

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

with periodic boundary conditions, see [46]. 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). Table 5.3 contains the L^1 errors for the cell averages and numerical orders of accuracy for the finite volume and RKDG schemes, respectively. We can clearly see that fifth order accuracy is achieved for the WENO scheme, and third order accuracy is achieved for the RKDG scheme. 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 Table 5.3. For the RKDG method, the CFL number is fixed at 0.18.

Table 5.3: L^1 errors and numerical orders of accuracy for the example in Section 5.3.1.2.

No. of cells	FV schemes				
	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
No. of cells	RKDG schemes				
		h		hu	
		L^1 error	order	L^1 error	order
25		2.35E-03		2.12E-02	
50		1.15E-04	4.36	1.01E-03	4.39
100		1.24E-05	3.20	1.09E-04	3.21
200		1.02E-06	3.59	8.97E-06	3.60
400		1.11E-07	3.19	9.79E-07	3.19
800		1.30E-08	3.09	1.14E-07	3.08

5.3.1.3 A small perturbation of a steady-state water

The following quasi-stationary test case was proposed by LeVeque [28]. 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.

The bottom topography consists of one hump:

$$b(x) = \begin{cases} 0.25(\cos(10\pi(x-1.5))+1) & \text{if } 1.4 \le x \le 1.6 \\ 0 & \text{otherwise} \end{cases} \tag{5.25}$$

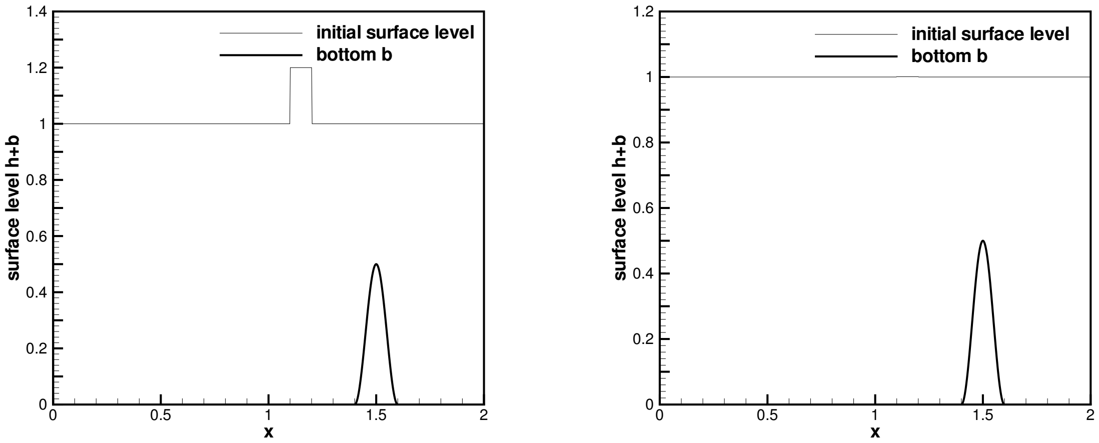


Figure 5.1: The initial surface level $h+b$ and the bottom b for a small perturbation of a steady-state water. Left: a big pulse ϵ=0.2; right: a small pulse ϵ=0.001.

The initial conditions are given with

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 1 - b(x) + \epsilon & \text{if } 1.1 \leq x \leq 1.2 \\ 1 - b(x) & \text{otherwise} \end{cases} \tag{5.26}$$

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 [28]. Both sets of initial conditions are shown in Figure 5.1. The solution at time t=0.2s for the big pulse $\epsilon = 0.2$, obtained on a 200 cell uniform grid with simple transmissive boundary conditions, and compared with a 3000 cell solution, is shown in Figure 5.2 for the FV scheme and in Figure 5.4 for the RKDG scheme. The results for the small pulse $\epsilon = 0.001$ are shown in Figures 5.3 and 5.5. For this small pulse problem, we take $\varepsilon = 10^{-9}$ in the WENO weight formula (2.10), such that it is smaller than the square of the perturbation. At this time, the downstream-traveling water pulse has already passed the bump. We can clearly see that there are no spurious numerical oscillations.

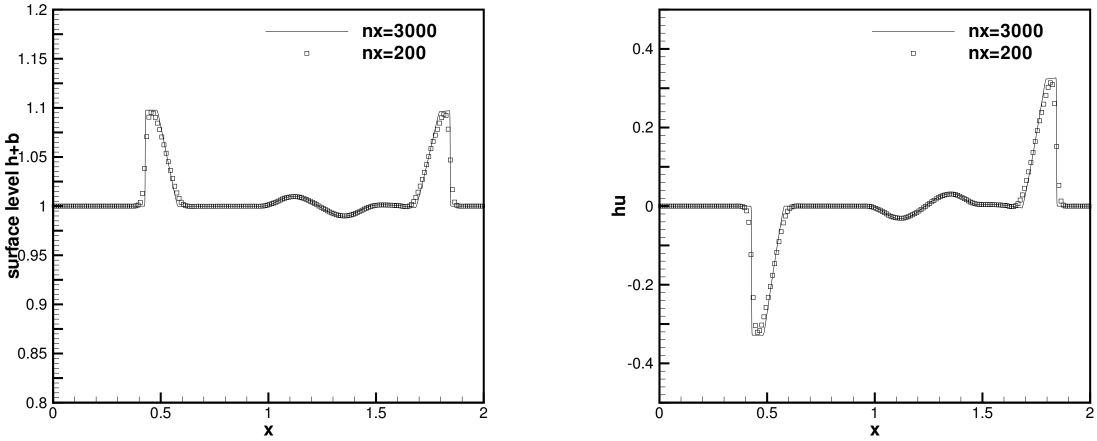


Figure 5.2: FV scheme: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

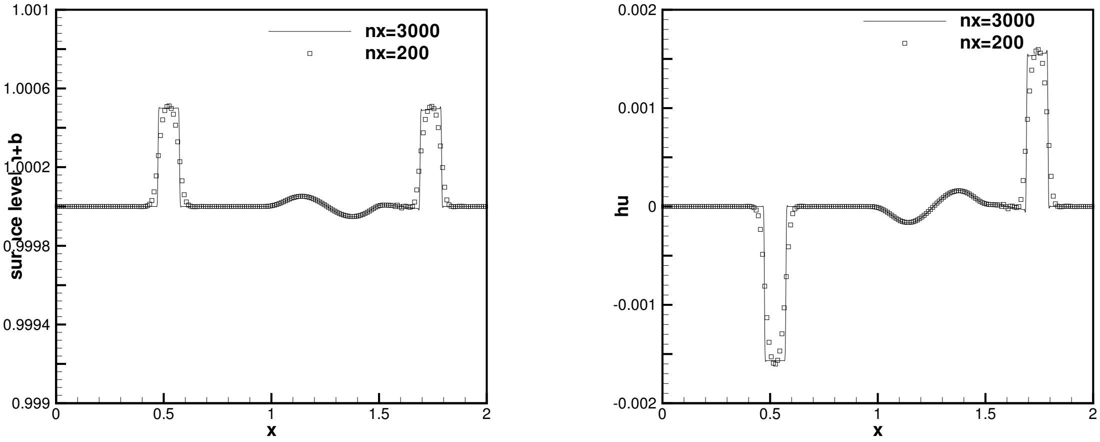


Figure 5.3: FV scheme: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

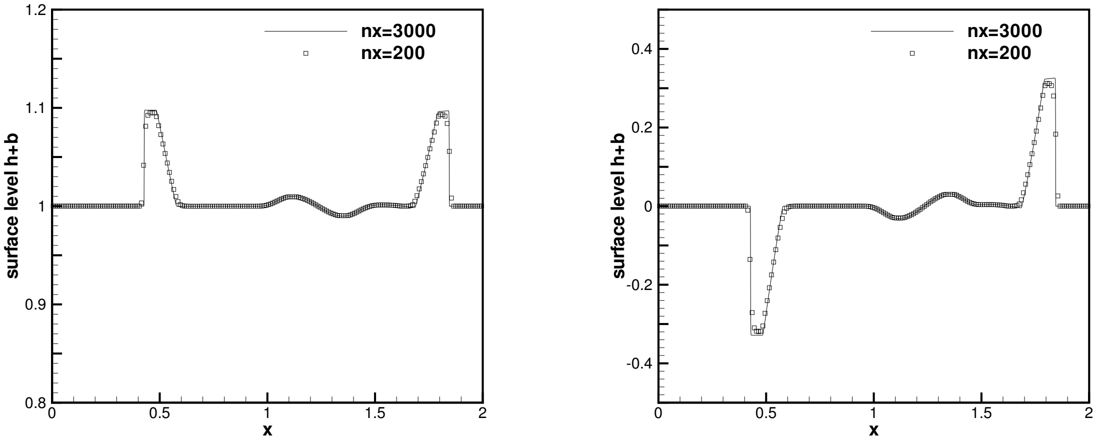


Figure 5.4: RKDG scheme: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

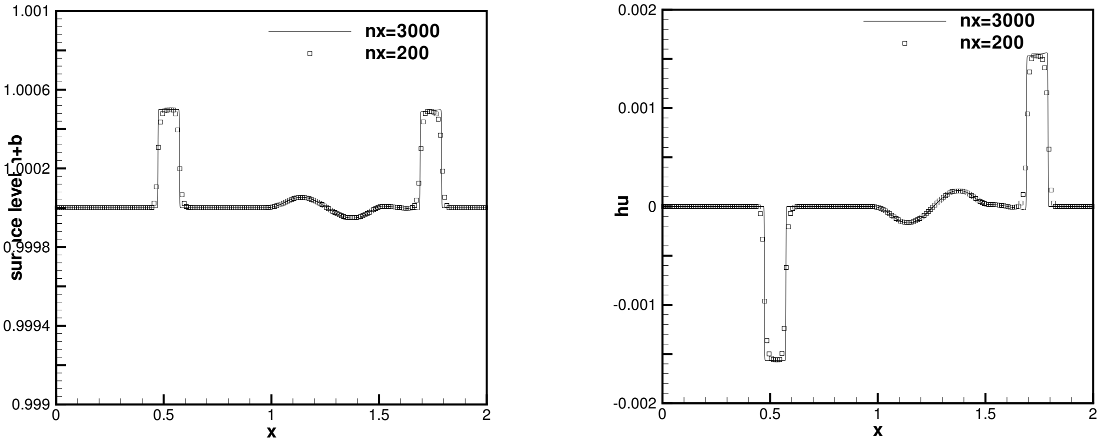


Figure 5.5: RKDG scheme: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h+b$; right: the discharge hu.

5.3.1.4 The dam breaking problem over a rectangular bump

In this example we 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 [44].

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leq 1500/8 \\ 0 & \text{otherwise} \end{cases} \tag{5.27}$$

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

$$(hu)(x, 0) = 0 \quad \text{and} \quad h(x, 0) = \begin{cases} 20 - b(x) & \text{if } x \leq 750 \\ 15 - b(x) & \text{otherwise} \end{cases} \tag{5.28}$$

The numerical results obtained by the FV scheme with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Figures 5.6 and 5.7, with two different ending time t=15s and t=60s. Figures 5.8 and 5.9 demonstrate the numerical results by the RKDG scheme, with the same number of uniform cells. 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. Both schemes work well for this example, giving well resolved, non-oscillatory solutions using 400 cells which agree with the converged results using 4000 cells.

5.3.1.5 Steady flow over a hump

The purpose of this test case is to study the convergence in time towards steady flow over a bump. These are classical test problems for transcritical and subcritical flows, and they are widely used to test numerical schemes for shallow water equations. For example, they have been considered by the *working group on dam break modelling* [17], and have been used as test cases in, e.g. [43].

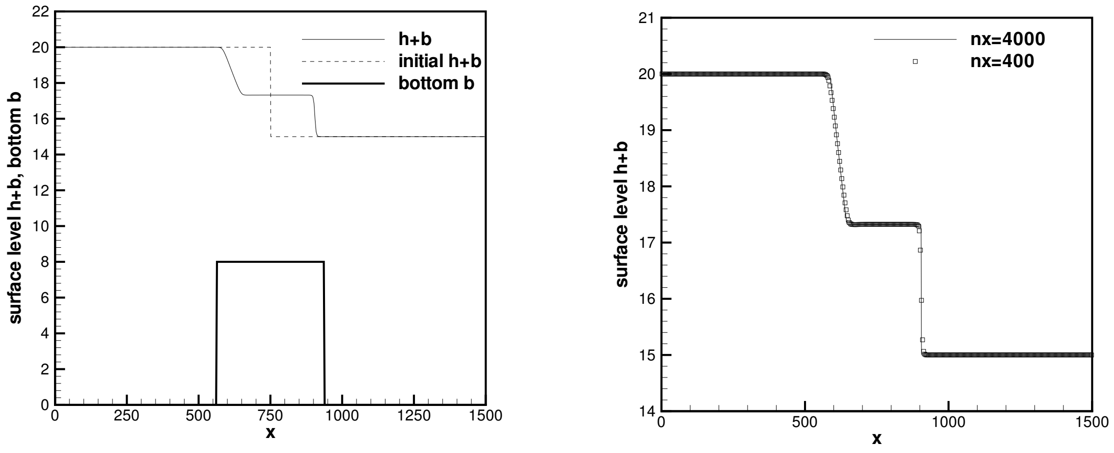


Figure 5.6: FV scheme: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

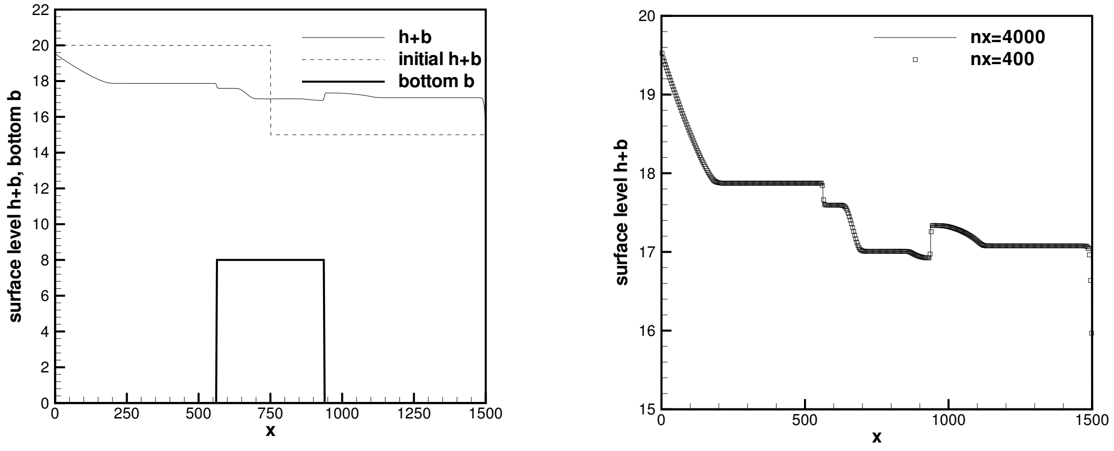


Figure 5.7: FV scheme: The surface level $h + b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

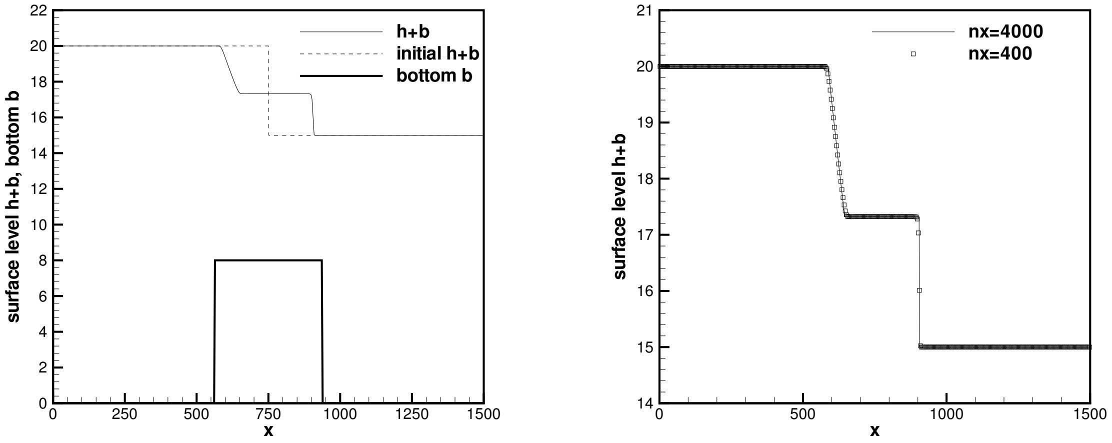


Figure 5.8: RKDG scheme: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

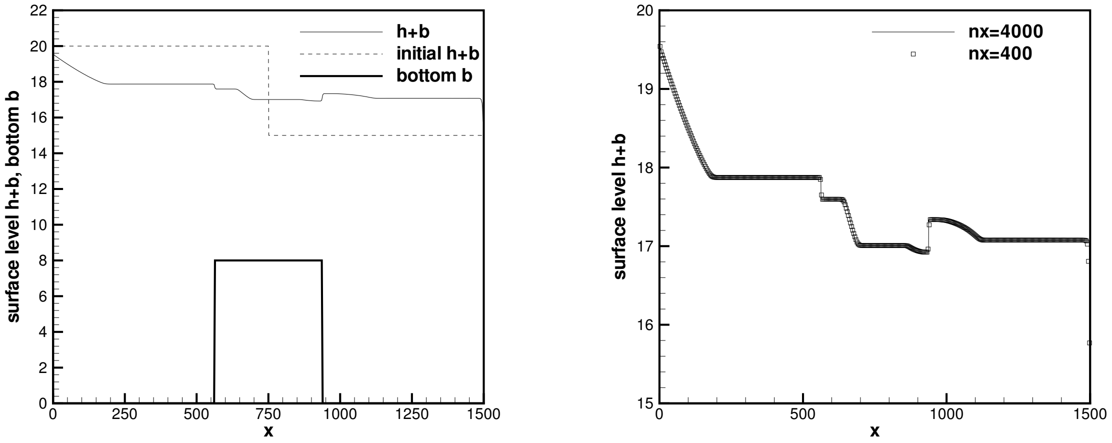


Figure 5.9: RKDG scheme: The surface level $h + b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

The bottom function is given by:

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12 \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

for a channel of length $25m$. The initial conditions are taken as

$$h(x,0) = 0.5 - b(x) \quad \text{and} \quad u(x,0) = 0.$$

Depending on different boundary conditions, the flow can be subcritical or transcritical with or without a steady shock. The computational parameters common for all three cases are: uniform mesh size $\Delta x = 0.125\ m$, ending time t= 200 s. Analytical solutions for the various cases are given in Goutal and Maurel [17].

a): Transcritical flow without a shock.

- 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.

The surface level $h+b$ and the discharge hu, as the numerical flux for the water height h in equation (5.22), are plotted in Figures 5.10 and 5.11, which show very good agreement with the analytical solution. The correct capturing of the discharge hu is usually more difficult than the surface level $h+b$, as noticed by many authors. The numerical errors for the discharge hu of our well-balanced finite volume WENO and RKDG schemes are both very small.

b): Transcritical flow with a shock.

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

In this case, the Froude number $Fr = u/\sqrt{gh}$ increases to a value larger than one above the bump, and then decreases to less than one. A stationary shock can appear

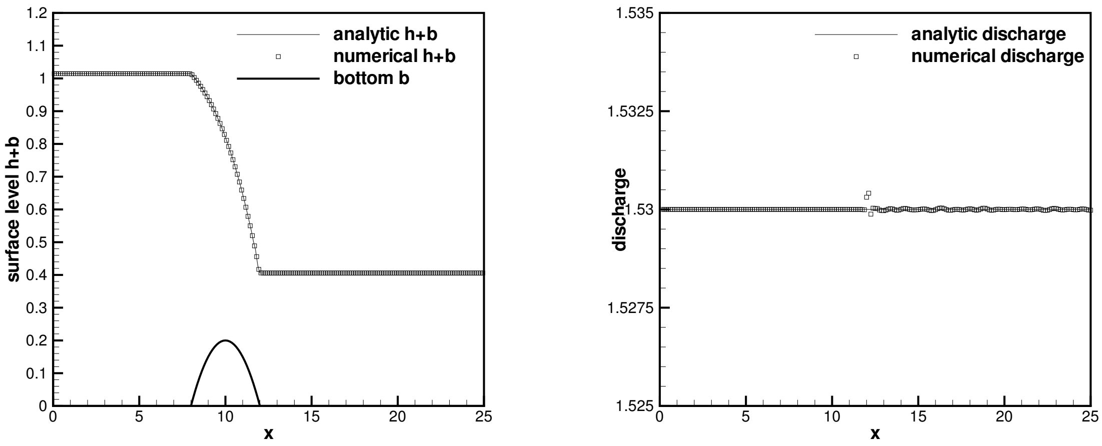


Figure 5.10: FV scheme: Steady transcritical flow over a bump without a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

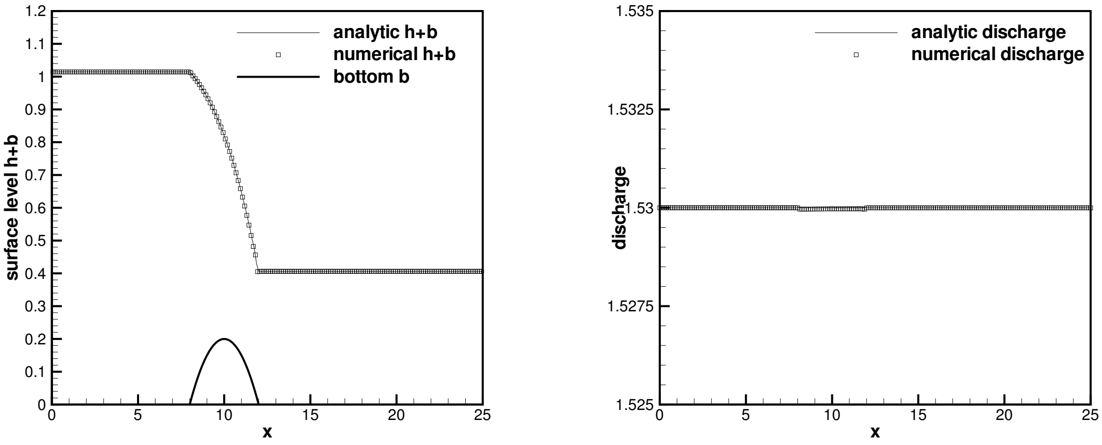


Figure 5.11: RKDG scheme: Steady transcritical flow over a bump without a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

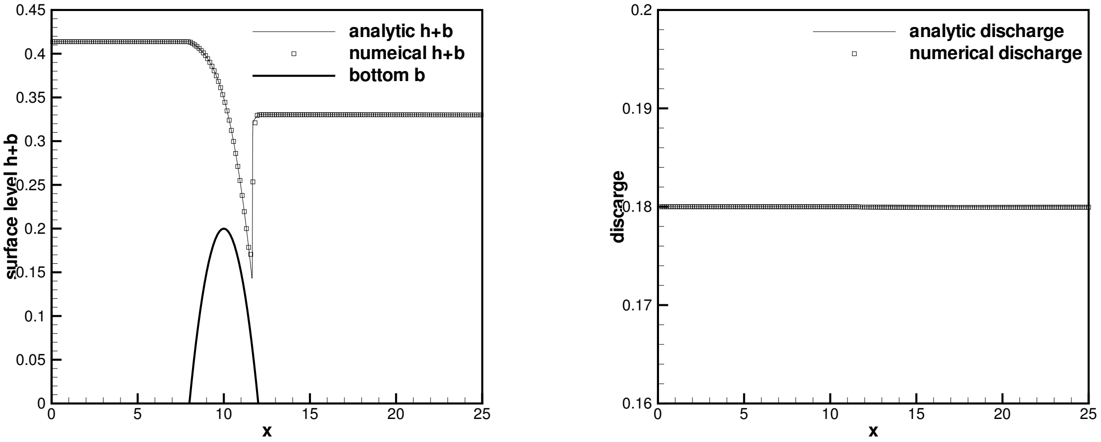


Figure 5.12: FV scheme: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

on the surface. The surface level $h+b$ and the discharge hu, as the numerical flux for the water height h in equation (5.22), are plotted in Figure 5.12 and 5.14, which show non-oscillatory results in good agreement with the analytical solution. In Figure 5.13 and 5.15, we compare the pointwise errors of the numerical solutions obtained with 200 and 400 uniform cells. We have also performed such error comparisons for the cases of the transcritical flow without a shock and of the subcritical flow, obtaining qualitatively similar results. We have therefore omitted them to save space.

c): Subcritical flow.

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

This is a subcritical flow. The surface level $h + b$ and the discharge hu, as the numerical flux for the water height h in equation (5.22), are plotted in Figure 5.16 and 5.17, which are in good agreement with the analytical solution.

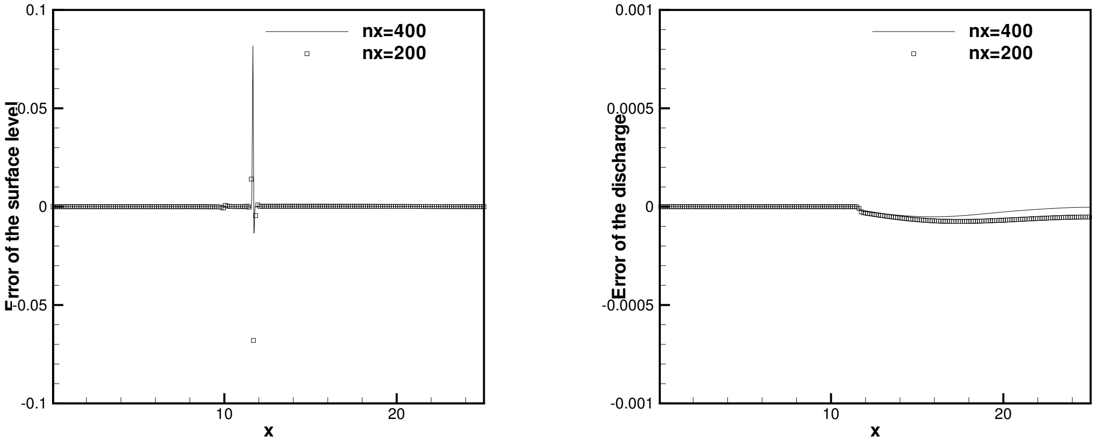


Figure 5.13: FV scheme: Steady transcritical flow over a bump with a shock. Pointwise error comparison between numerical solutions using 200 and 400 cells. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

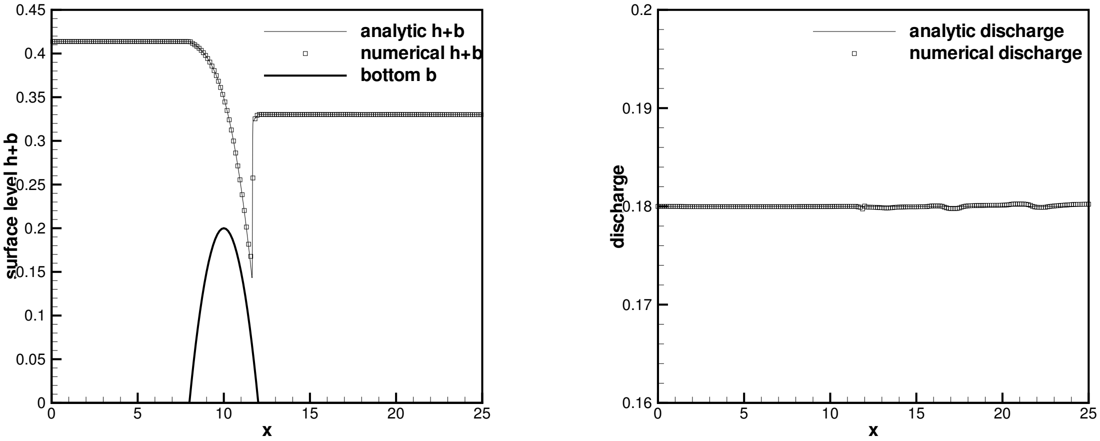


Figure 5.14: RKDG scheme: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

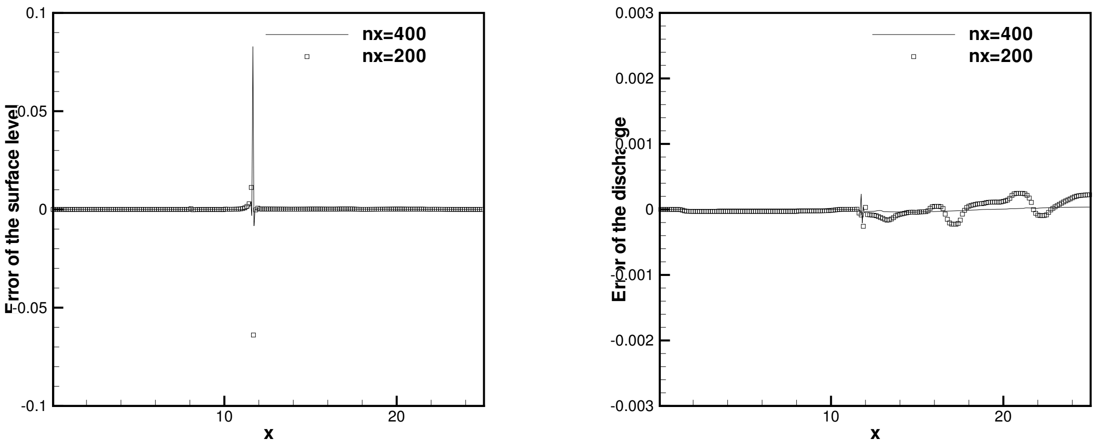


Figure 5.15: RKDG scheme: Steady transcritical flow over a bump with a shock. Pointwise error comparison between numerical solutions using 200 and 400 cells. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

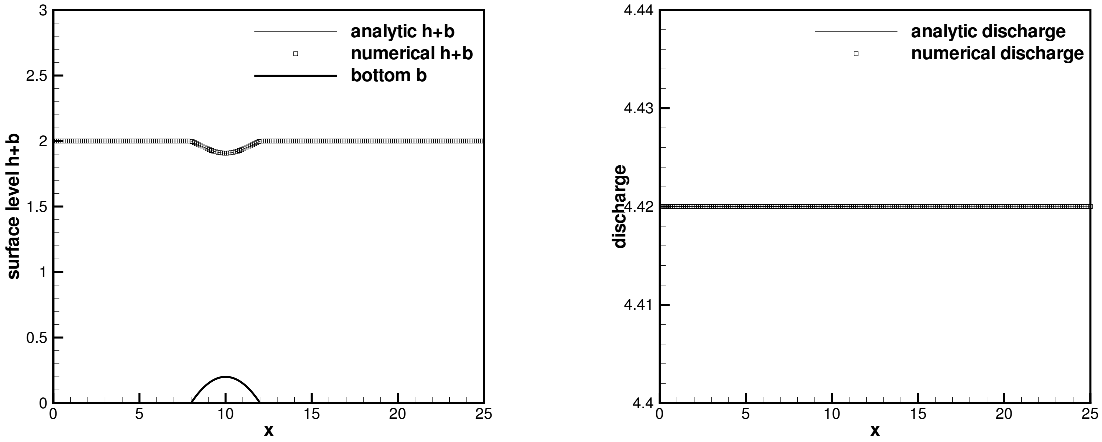


Figure 5.16: FV scheme: Steady subcritical flow over a bump. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

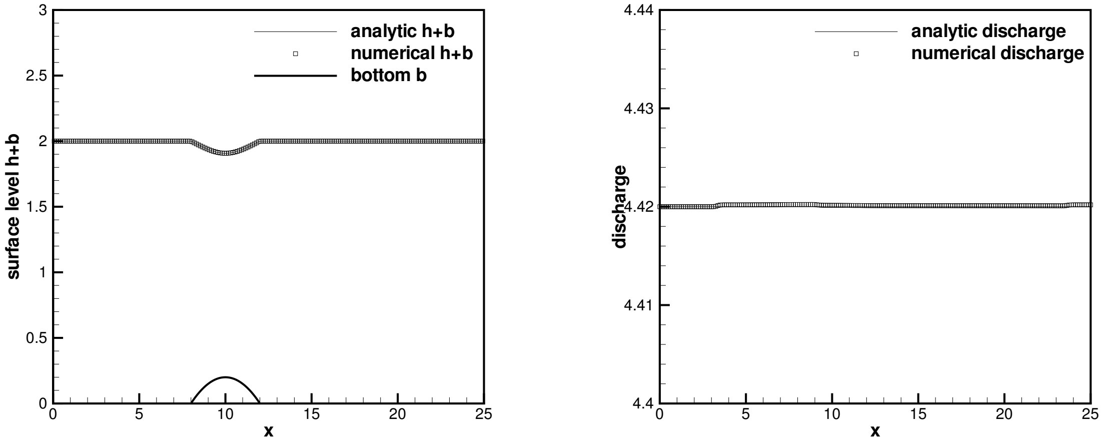


Figure 5.17: RKDG scheme: Steady subcritical flow over a bump. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

5.3.2 Two dimensional shallow water equations

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

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x + (huv)_y = -ghb_x \\ (hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2}gh^2\right)_y = -ghb_y \end{cases} \tag{5.30}$$

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.

We are interested in preserving the still water solution, which takes the form (satisfying (5.17))

$$a_1 \equiv h + b = constant, \qquad a_2 \equiv u = 0, \qquad a_3 \equiv v = 0.$$

The first component of the source term is 0. Similarly as in one dimensional case, we decompose the second and third components of the source term as

$$-ghb_x = -g\,(h+b)\,b_x + \frac{1}{2}g\,\left(b^2\right)_x, \qquad -ghb_y = -g\,(h+b)\,b_y + \frac{1}{2}g\,\left(b^2\right)_y,$$

i.e. $s_1 = s_1(a_1) = -g\,(h+b)$, $s_2 = \frac{1}{2}g$, $t_1(x) = b(x)$, $t_2(x) = b^2(x)$ for the second component, and $s_1 = s_1(a_1) = -g\,(h+b)$, $s_2 = \frac{1}{2}g$, $t_1(x) = b(x)$, $t_2(x) = b^2(x)$ for the third component.

For the finite volume scheme, we apply the WENO reconstruction to the function $(b(x),0,0)^T$, with coefficients computed from $(h,hu,hv)^T$, to obtain $b^{\pm}_{i+\frac{1}{2},j}$ and $b^{\pm}_{i,j+\frac{1}{2}}$. We define, for the source term of the second equation,

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

and, for the source term of the third equation,

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

We can verify, similar to the one dimensional case, that these choices of $t^{\pm}_j$ will maintain the requirement for the steady state solution satisfying $h+b=c$, $u=v=0$.

For the RKDG method, we define

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

where $b_h(x,y)$ is the L^2 projection of $b(x,y)$ to the finite element space V_h, for the source terms of both the second and the third equations.

We now show numerical examples to demonstrate the behavior of our well balanced schemes for the two dimensional shallow water equations.

5.3.2.1 Test for the exact C-property in two dimensions

This example is used to check that our schemes indeed maintain the exact C-property over a non-flat bottom. The two-dimensional hump

$$b(x,y) = 0.8e^{-50((x-0.5)^2+(y-0.5)^2)}, \qquad x, y \in [0,1] \tag{5.31}$$

is chosen to be the bottom. $h(x,y,0) = 1 - b(x,y)$ is the initial depth of the water. Initial velocity is set to be zero. This surface should remain flat. The computation is performed to $t = 0.1$ using single, double and quadruple precisions with a 100×100 uniform mesh. Table 5.4 contains the L^1 errors for the water height h (which is not a constant function) and the discharges hu and hv for both schemes. We can clearly see that the L^1 errors are at the level of round-off errors for different precisions, verifying the exact C-property.

Table 5.4: L^1 errors for different precisions for the stationary solution in Section 5.3.2.1.

	precision	L^1 error		
		h	hu	hv
FV	single	1.09E-06	8.87E-07	8.87E-07
	double	8.16E-16	9.31E-16	8.47E-16
	quadruple	7.30E-34	7.31E-34	7.34E-34
RKDG	single	9.40E-08	3.58E-07	3.60E-07
	double	6.20E-17	1.14E-15	1.16E-15
	quadruple	5.87E-34	8.35E-34	8.36E-34

5.3.2.2 Testing the orders of accuracy

In this example we check the numerical orders of accuracy when the schemes are applied to the following two dimensional problem. The bottom topography and the initial data are given by:

$$b(x,y) = \sin(2\pi x) + \cos(2\pi y), \qquad h(x,y,0) = 10 + e^{\sin(2\pi x)}\cos(2\pi y),$$

$$(hu)(x,y,0) = \sin(\cos(2\pi x))\sin(2\pi y), \qquad (hv)(x,y,0) = \cos(2\pi x)\cos(\sin(2\pi y))$$

defined over a unit square, with periodic boundary conditions. The terminal time is taken as t=0.05 to avoid the appearance of shocks in the solution. Since the exact solution is also not known explicitly for this case, we use the same fifth order WENO scheme with an extremely refined mesh consisting of 1600 × 1600 cells to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. The TVB constant M in the limiter for the RKDG scheme is taken as 40 here. Tables 5.5 and 5.6 contain the L^1 errors and orders of accuracy for the cell averages. We can clearly see that, in this two dimensional test case, fifth order accuracy is achieved for the finite volume WENO scheme and third order accuracy is achieved for the RKDG scheme.

Table 5.5: FV scheme: L^1 errors and numerical orders of accuracy for the example in Section 5.3.2.2.

Number		h		hu		hv	
of cells	CFL	L^1 error	order	L^1 error	order	L^1 error	order
25 × 25	0.6	7.91E-03		2.12E-02		6.52E-02	
50 × 50	0.6	1.13E-03	2.81	2.01E-03	3.40	9.23E-03	2.82
100 × 100	0.6	8.89E-05	3.66	1.25E-04	4.00	7.19E-04	3.68
200 × 200	0.4	4.07E-06	4.45	5.19E-06	4.59	3.30E-05	4.45
400 × 400	0.3	1.42E-07	4.84	1.84E-07	4.82	1.16E-06	4.84
800 × 800	0.2	4.38E-09	5.02	5.99E-09	4.94	3.63E-08	4.99

Table 5.6: RKDG scheme: L^1 errors and numerical orders of accuracy for the example in Section 5.3.2.2.

Number of cells	h		hu		hv	
	L^1 error	order	L^1 error	order	L^1 error	order
25×25	2.45E-03		1.36E-02		2.05E-02	
50×50	5.73E-04	2.10	2.92E-03	2.22	4.75E-03	2.11
100×100	1.06E-04	2.43	5.31E-04	2.46	8.51E-04	2.48
200×200	1.71E-05	2.63	8.82E-05	2.59	1.39E-04	2.61
400×400	2.52E-06	2.76	1.32E-05	2.74	2.10E-05	2.73
800×800	3.52E-07	2.84	1.89E-06	2.80	3.01E-06	2.81

5.3.2.3 A small perturbation of a two dimensional steady-state water

This is a classical example to show the capability of the proposed scheme for the perturbation of the stationary state, given by LeVeque [28]. It is analogous to the test done previously in Section 5.3.1.3 in one dimension.

We solve the system in the rectangular domain $[0, 2] \times [0, 1]$. The bottom topography is an isolated elliptical shaped hump:

$$b(x, y) = 0.8\, e^{-5(x-0.9)^2-50(y-0.5)^2}. \tag{5.32}$$

The surface is initially given by:

$$\begin{aligned} h(x, y, 0) &= \begin{cases} 1 - b(x, y) + 0.01 & \text{if } 0.05 \le x \le 0.15 \\ 1 - b(x, y) & \text{otherwise} \end{cases} \\ hu(x, y, 0) &= hv(x, y, 0) = 0 \end{aligned} \tag{5.33}$$

So the surface is almost flat except for $0.05 \le x \le 0.15$, where h is perturbed upward by 0.01. Figures 5.18 and 5.19 display the right-going disturbance as it propagates past the hump, on two different uniform meshes with 200×100 cells and 600×300 cells for comparison. The surface level $h + b$ is presented at different times. The results indicate that both schemes can resolve the complex small features of the flow very well.

5.3.3 Elastic wave equation

We consider the propagation of compressional waves [3, 45] in an one-dimensional elastic rod with a given media density $\rho(x)$. The equations of motion in a Lagrangian frame are given by the balance laws:

$$\begin{cases} (\rho\varepsilon)_t + (-\rho u)_x = -u\dfrac{d\rho}{dx} \\ (\rho u)_t + (-\sigma)_x = 0, \end{cases} \tag{5.34}$$

where $\varepsilon = \varepsilon(x,t)$ is the strain, $u = u(x,t)$ is the velocity and σ is a given stress-strain relationship $\sigma(\varepsilon, x)$. The equation of linear acoustics can be obtained from the elasticity problem if the stress-strain relationship is linear,

$$\sigma(\varepsilon, x) = K(x)\,\varepsilon$$

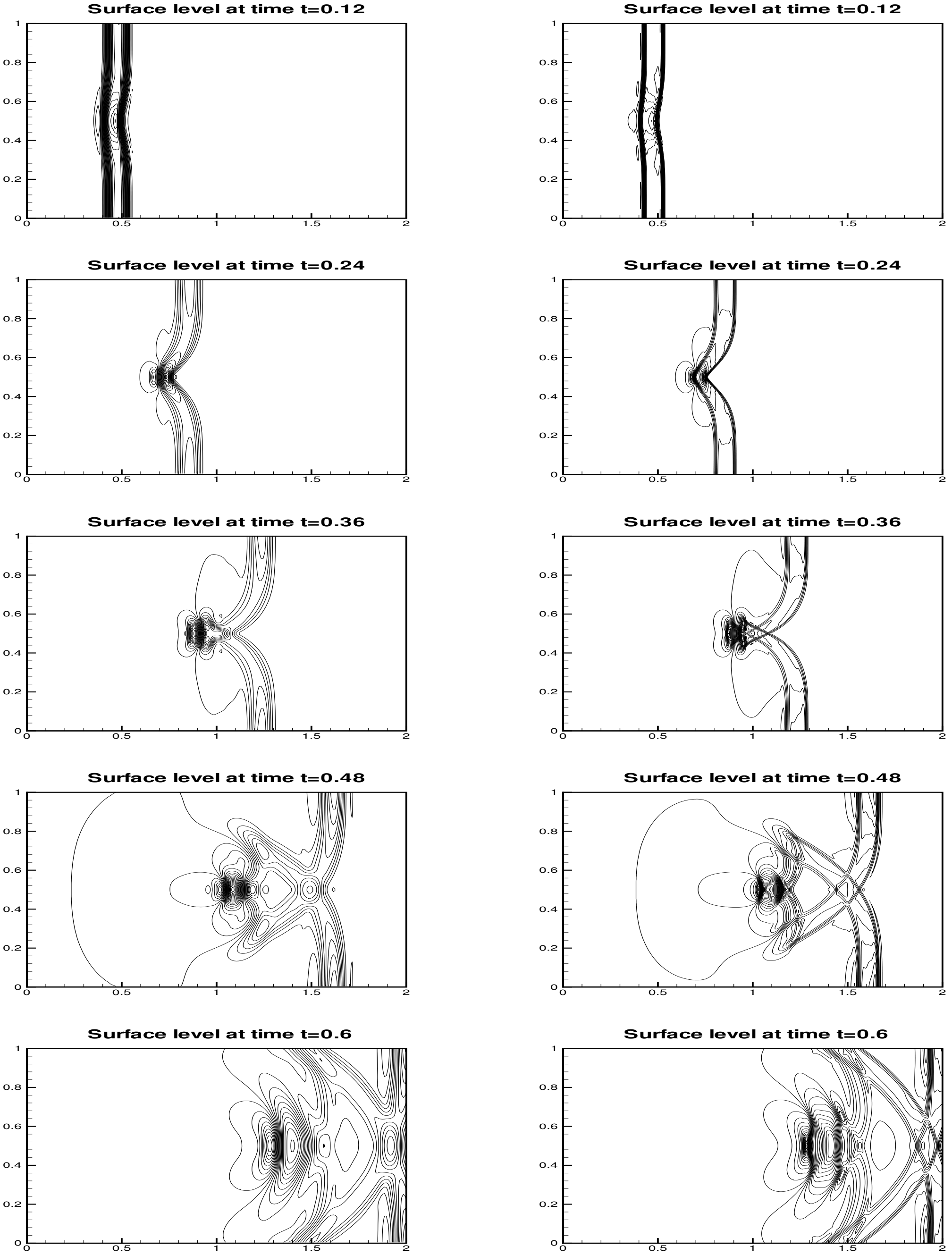
where $K(x)$ is the given bulk modulus of compressibility.

The steady state we are interested to preserve for this problem is characterized by

$$a_1 \equiv \sigma(\varepsilon, x) = constant, \qquad a_2 \equiv u = constant$$

which is of the form (5.17). The second component of the source term is 0. The first component of the source term is already in the form of (5.3) with $s_1 = s_1(a_2) = -u = -\frac{\rho u}{\rho}$ and $t_1 = \rho(x)$.

For finite volume schemes, we apply the WENO reconstruction to the function $(0, \rho(x))^T$, with coefficients computed from $(\rho\varepsilon, \rho u)^T$, to obtain $\rho^{\pm}_{i+\frac{1}{2}}$. We then define



Surface level at time t=0.12

Surface level at time t=0.12

Surface level at time t=0.24

Surface level at time t=0.24

Surface level at time t=0.36

Surface level at time t=0.36

Surface level at time t=0.48

Surface level at time t=0.48

Surface level at time t=0.6

Surface level at time t=0.6

Figure 5.18: FV scheme: The contours of the surface level $h + b$ for the problem in Section 5.3.2. 30 uniformly spaced contour lines. From top to bottom: at time $t = 0.12$ from 0.99942 to 1.00656; at time $t = 0.24$ from 0.99318 to 1.01659; at time $t = 0.36$ from 0.98814 to 1.01161; at time $t = 0.48$ from 0.99023 to 1.00508; and at time $t = 0.6$ from 0.99514 to 1.00629. Left: results with a 200×100 uniform mesh. Right: results with a 600×300 uniform mesh.

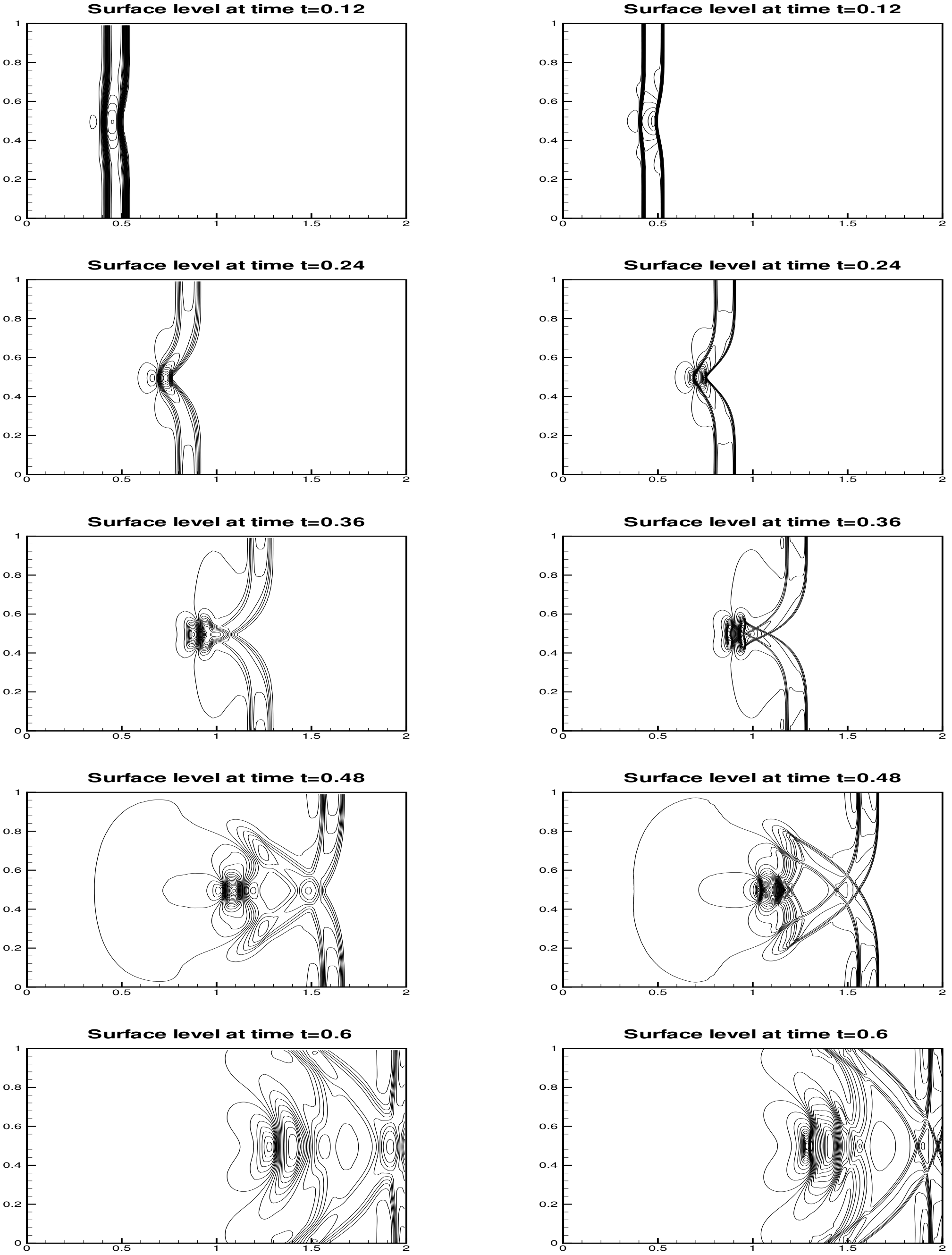


Figure 5.19: RKDG scheme: The contours of the surface level $h + b$ for the problem in Section 5.3.2. 30 uniformly spaced contour lines. From top to bottom: at time $t = 0.12$ from 0.99942 to 1.00656; at time $t = 0.24$ from 0.99318 to 1.01659; at time $t = 0.36$ from 0.98814 to 1.01161; at time $t = 0.48$ from 0.99023 to 1.00508; and at time $t = 0.6$ from 0.99514 to 1.00629. Left: results with a 200 × 100 uniform mesh. Right: results with a 600 × 300 uniform mesh.

$(t_1)^{\pm}_{i+\frac{1}{2}} = \rho^{\pm}_{i+\frac{1}{2}}$, which leads to

$$f(u^{\pm}_{i+\frac{1}{2}}) - \sum_j s_j(a(u,x)^{\pm}_{i+\frac{1}{2}})(t_j)^{\pm}_{i+\frac{1}{2}} = -(\rho u)^{\pm}_{i+\frac{1}{2}} + \frac{(\rho u)^{\pm}_{i+\frac{1}{2}}}{\rho^{\pm}_{i+\frac{1}{2}}}\rho^{\pm}_{i+\frac{1}{2}} = 0,$$

satisfying our requirement. For the RKDG scheme, we define

$$(t_1)_h(x) = \rho_h(x)$$

where $\rho_h(x)$ is the L^2 projection of $\rho(x)$ to the finite element space V_h. We can then easily verify the requirement

$$f(u_h) - \sum_j s_j(a_h(u_h, x))(t_j)_h = 0$$

for the steady state solution.

Next, we present the numerical result for a linear acoustic test [3]. The properties of the media are given by

$$c(x) = \sqrt{\frac{K(x)}{\rho(x)}} = 1 + 0.5\sin(10\pi x), \qquad Z(x) = \rho(x)c(x) = 1 + 0.25\cos(10\pi x).$$

The initial conditions are given by

$$\rho\,\varepsilon(x,0) = \begin{cases} \dfrac{-1.75 + 0.75\cos(10\pi x)}{c^2(x)}, & \text{if } 0.4 < x < 0.6 \\ \dfrac{-1}{c^2(x)}, & \text{otherwise} \end{cases}, \qquad u(x,0) = 0.$$

It is a test case where the impedance $Z(x)$ and hence the eigenvectors are both spatially varying. We perform the computation with 200 uniform cells, with the ending time $t = 0.4s$. An “exact” reference solution is computed with the same scheme over a 2000 uniform cells. The simulation results are shown in Figure 5.20. The numerical resolution shows very good agreement with the “exact” reference

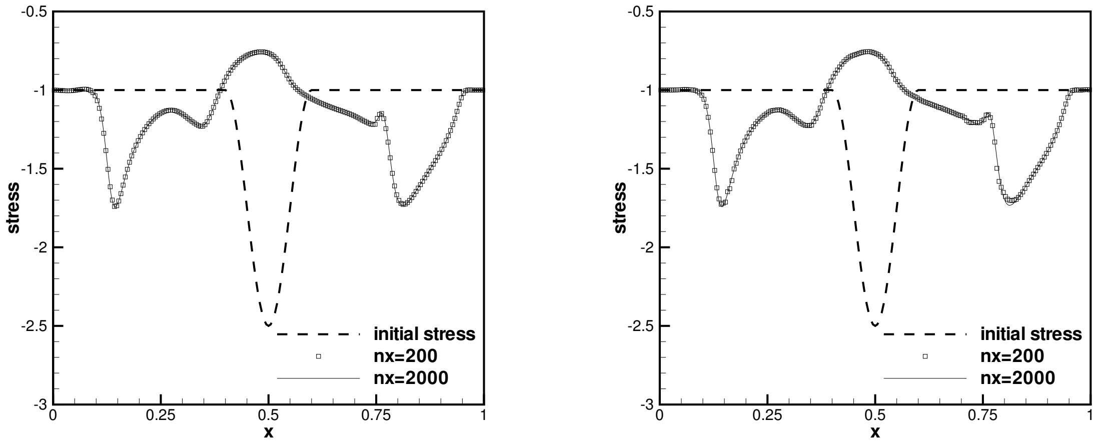


Figure 5.20: The numerical (symbols) and the "exact" reference (solid line) stress $\sigma(x)$ at time $t = 0.4s$. Left: FV schemes; right: RKDG schemes.

solution.

5.3.4 Chemosensitive movement

Originated from biology, chemosensitive movement [21, 15] is a process by which cells change their direction reacting to the presence of a chemical substance, approaching chemically favorable environments and avoiding unfavorable ones. Hyperbolic models for chemotaxis are recently introduced [21] and take the form

$$\begin{cases} n_t + (nu)_x = 0 \\ (nu)_t + (nu^2 + n)_x = n\chi'(c)\dfrac{\partial c}{\partial x} - \sigma nu \end{cases} \tag{5.35}$$

where the chemical concentration $c = c(x, t)$ is given by the parabolic equation

$$\frac{\partial c}{\partial t} - D_c \triangle c = n - c.$$

Here, $n(x, t)$ is the cell density, $nu(x, t)$ is the population flux and σ is the friction coefficient.

We would like to preserve the steady state solution to (5.35) with a zero population flux, which satisfies

$$n\chi'(c)c_x - n_x = 0, \qquad nu = 0. \tag{5.36}$$

where $c = c(x)$ does not depend on t in steady state. The first equality above does not seem to be of the form (5.17). However, (5.36) is equivalent to

$$a_1 \equiv \frac{n}{e^{\chi(c)}} = constant, \qquad a_2 \equiv nu = 0,$$

which is clearly in the form of (5.17). The first component of the source term is 0. A decomposition of the second component of the source term in the form of (5.3) is

$$n\chi'(c)\frac{\partial c}{\partial x} - \sigma nu = \frac{n}{e^{\chi(c)}}\frac{d}{dx}e^{\chi(c)} - \sigma nu$$

i.e. $s_1 = s_1(a_1) = \frac{n}{e^{\chi(c)}}$, $s_2 = s_2(a_2) = \sigma nu$, $t_1(x) = e^{\chi(c(x))}$, and $t_2(x) = x$.

For the finite volume scheme, we apply the WENO reconstruction to the function $(e^{\chi(c(x))}, 0)^T$, with coefficients computed from $(n, nu)^T$, to obtain $(e^{\chi(c(x))})^{\pm}_{i+\frac{1}{2}}$. We define

$$(t_1)^{\pm}_{i+\frac{1}{2}} = (e^{\chi(c(x))})^{\pm}_{i+\frac{1}{2}}, \qquad (t_2)^{\pm}_{i+\frac{1}{2}} = x_{i+\frac{1}{2}}.$$

In the case of steady state,

$$f(u^{\pm}_{i+\frac{1}{2}}) - \sum_j s_j(a(u,x)^{\pm}_{i+\frac{1}{2}})(t_j)^{\pm}_{i+\frac{1}{2}} = n^{\pm}_{i+\frac{1}{2}} - \frac{n^{\pm}_{i+\frac{1}{2}}}{(e^{\chi(c(x))})^{\pm}_{i+\frac{1}{2}}}(e^{\chi(c(x))})^{\pm}_{i+\frac{1}{2}} = 0,$$

which satisfies our requirement. For the RKDG scheme, we define

$$(t_1)_h(x) = (e^{\chi(c(x))})_h, \qquad (t_2)_h(x) = x$$

where $(e^{\chi(c(x))})_h$ is the L^2 projection of $e^{\chi(c(x))}$ to the finite element space V_h. A similar manipulation as in the finite volume case leads to

$$f(u_h) - \sum_j s_j(a_h(u_h, x))(t_j)_h = 0.$$

Our technique can also be applied to the two dimensional case of this application.

We give an numerical example here to test the high order accuracy for smooth solutions for our schemes. The initial conditions are taken as

$$n(x,0) = 1 + 0.2\ \cos(\pi x), \qquad u(x,0) = 0, \qquad x \in [-1,1]$$

with

$$c(x) = e^{-16x^2}, \qquad \chi(c) = \log(1+c), \qquad \sigma = 0$$

with a periodic boundary condition. Since the exact solution is not known explicitly for this problem, we use the same fifth order WENO scheme with $N = 5120$ cells to compute a reference solution and treat it as the exact solution when computing the numerical errors for the cell averages. Final time $t = 0.5s$ is used to avoid the development of shocks. The TVB constant M in the limiter for the RKDG scheme is taken as 13 for this example. Table 5.7 contains the L^1 errors and numerical orders of accuracy. We can clearly see that expected order accuracy is achieved for this example.

5.3.5 A model in fluid mechanics with spherical symmetry

A classical singularity arising in fluid mechanics in case of spherical symmetry leads to the following model equation

$$u_t + \left(\frac{u^2}{2}\right)_x = \frac{1}{x}u^2, \tag{5.37}$$

Table 5.7: L^1 errors and numerical orders of accuracy for the example in Section 5.3.4.

No. of cells	FV schemes				
	CFL	$\rho\epsilon$		ρu	
		L^1 error	order	L^1 error	order
20	0.6	9.70E-03		7.41E-03	
40	0.6	1.03E-03	3.24	8.85E-04	3.07
80	0.5	1.07E-04	3.26	8.80E-05	3.33
160	0.4	5.63E-06	4.25	5.63E-06	3.97
320	0.3	2.21E-07	4.67	1.89E-07	4.89
640	0.1	7.18E-09	4.94	6.07E-08	4.96
No. of cells	RKDG schemes				
		$\rho\epsilon$		ρu	
		L^1 error	order	L^1 error	order
20		1.27E-04		1.46E-04	
40		1.75E-05	2.85	2.07E-05	2.82
80		1.32E-06	3.73	1.89E-06	3.46
160		1.21E-07	3.45	1.97E-07	3.26
320		1.29E-08	3.23	2.27E-08	3.12
640		1.57E-09	3.03	2.76E-09	3.04

which has been considered in [6]. Notice that the source term is a nonlinear function of u. The steady state for this problem is given by

$$\frac{du}{dx} = \frac{u}{x} \qquad \Rightarrow \qquad a(u,x) \equiv \frac{u}{x} = constant \tag{5.38}$$

which is of the form (5.2) with $p(x) = 0$ and $q(x) = x$. The source term can be rewritten as

$$\frac{u^2}{x} = \left(\frac{u}{x}\right)^2 x = \left(\frac{u}{x}\right)^2 \left(\frac{x^2}{2}\right)_x$$

which is in the form of (5.3) with $s_1(a) = a^2 = \left(\frac{u}{x}\right)^2$ and $t_1(x) = \frac{x^2}{2}$. Note that here s_1 is a nonlinear function of a.

For finite volume schemes, we apply the WENO reconstruction to the function

$q(x) = x$, with coefficients computed from u, to obtain $q^{\pm}_{i+\frac{1}{2}}$. Since x is a polynomial with degree 1, the reconstructed $q^{\pm}_{i+\frac{1}{2}}$ should be exactly $x_{i+\frac{1}{2}}$, no matter how we compute the WENO coefficients. Hence, we can use $x_{i+\frac{1}{2}}$ directly, without applying WENO reconstruction on it. We then define $(t_1)^{\pm}_{i+\frac{1}{2}} = \frac{x^2_{i+\frac{1}{2}}}{2}$, which leads to

$$f(u^{\pm}_{i+\frac{1}{2}}) - \sum_j s_j(a(u,x)^{\pm}_{i+\frac{1}{2}})(t_j)^{\pm}_{i+\frac{1}{2}}$$

$$= \frac{(u^{\pm}_{i+\frac{1}{2}})^2}{2} - \left(\frac{u^{\pm}_{i+\frac{1}{2}}}{x^{\pm}_{i+\frac{1}{2}}}\right)^2 \frac{x^2_{i+\frac{1}{2}}}{2} = \frac{(u^{\pm}_{i+\frac{1}{2}})^2}{2} - \frac{(u^{\pm}_{i+\frac{1}{2}})^2}{2} = 0,$$

satisfying our requirement. For the RKDG scheme, we define

$$(t_1)_h(x) = \frac{x^2}{2}$$

and we can then easily verify the requirement

$$f(u_h) - \sum_j s_j(a_h(u_h,x))(t_j)_h = 0$$

for the steady state solution.

Next, we present a numerical result to demonstrate the well balanced property. The initial and boundary conditions are given by

$$u(x,0) = 0, \qquad x \in [-5,5] \tag{5.39}$$

$$u(x=-5,t) = 10, \qquad u(x=5,t) = -10. \tag{5.40}$$

The choice of these information allows us to compute the steady state, which is $u = -2x$. Numerical computations are performed by the well-balanced version of finite volume WENO schemes and RKDG methods. To see the benefit of well balanced schemes, we also use a non well balanced finite volume WENO schemes and RKDG methods, and compare the results. We use 100 uniform cells here. The

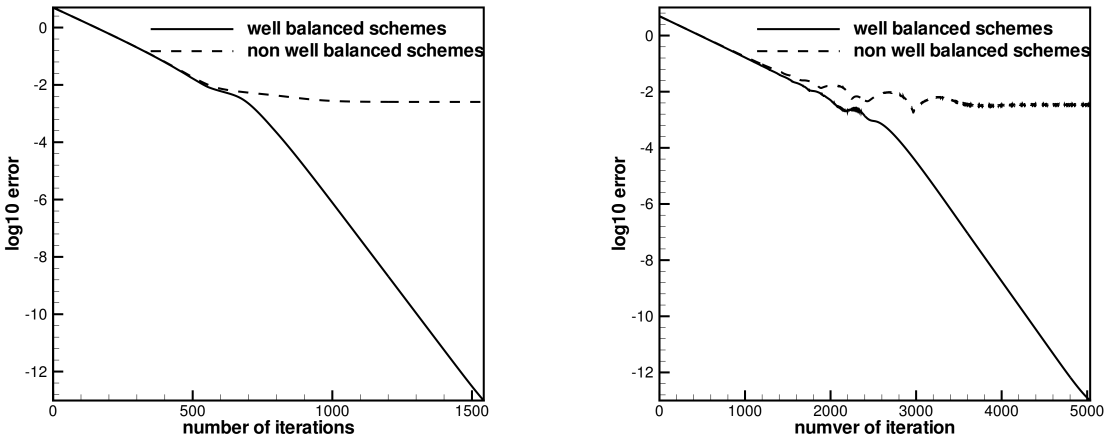


Figure 5.21: Comparison of the convergence history in L^1 error. Left: FV WENO schemes; right: RKDG schemes.

comparison of the convergence history, measured by the L^1 norm of the difference with the steady state, is given in Figure 5.21. The advantage of the well balanced schemes can be easily observed. Also, we compute the L^1 and L^∞ errors at time $t = 10$, with single precision and double precision. The results are shown in Table 5.8. We can clearly see that the errors are at the level of round-off errors for different precisions, verifying the well-balanced property.

Table 5.8: L^1 and L^∞ errors for different precisions for the steady state (5.38).

	FV		DG	
precision	L^1 error	L^∞ error	L^1 error	L^∞ error
single	6.06E-06	2.24E-05	2.63E-05	9.87E-05
double	1.60E-14	7.42E-14	3.25E-14	2.16E-13

5.3.6 Other applications

There are many other application problems which admit steady states that can be approximated by our well balanced schemes. These include the nozzle flow problem, a two phase flow model and a typical example with a stiff source term. We refer to [47] for more details of the first two models. The model with a stiff source term takes the form:

$$u_t + u_x = -\frac{1}{\epsilon} u(u-1). \tag{5.41}$$

We can easily check that our well balanced schemes can be applied to these models. Due to page limitation, we do not include computational results for these models here.

Chapter 6

A New Approach of High Order Well-balanced Finite Volume WENO Schemes and RKDG Methods for a Class of Hyperbolic Systems with Source Terms

In this chapter, we design a new approach of high order well balanced finite volume WENO schemes and RKDG finite element schemes. The general setup to obtain well balanced property is completely different from the one used in Chapter Five. In Section 6.1, we develop genuine high order well-balanced RKDG schemes for the shallow water equations. The well-balanced generalization of finite volume WENO schemes is presented in Section 6.2. Section 6.3 contains extensive numerical simulation results to demonstrate the behavior of our well balanced schemes for one and two dimensional shallow water equations, verifying high order accuracy, the exact C-property, and good resolution for smooth and discontinuous solutions. Application of these ideas to other balance laws, together with some selective numerical tests, are presented in Section 6.4.

6.1 Construction of well balanced RKDG schemes for shallow water equations

The traditional high order RKDG method has been presented in Section 2.3. In this section, we claim that, for one-dimensional and two-dimensional shallow water equations, this method is indeed a well balanced scheme for still water, based on a suitable choice of the initial value or the flux. This choice will not affect the property of the scheme, such as high order accuracy in smooth region and non-oscillatory shock resolution, and it increases the computational cost only slightly.

For the shallow water equations (1.3), we are interested in preserving the still water stationary solution (1.5). For this still water, the first equation $(hu)_x = 0$ is satisfied exactly for any consistent scheme since $hu = 0$. Let us concentrate on the second equation, which can be denoted by

$$(hu)_t + f(U)_x = g(h, b)$$

where $U = (h, hu)^T$ with the superscript T denoting the transpose.

As described in Section 2.3, in a RKDG method U is approximated by the piecewise polynomial U_h, which belongs to V_h defined in (2.21). We project the bottom function b into the same space V_h, to obtain an approximation b_h. This implies that $h_h + b_h = constant$ if $h + b = constant$.

Following the idea first introduced by Audusse et al. [2], and later used in the recent paper by Noelle et al. [30], our numerical scheme has the form:

$$\int_{I_j} \partial_t (hu)_h v_h dx - \int_{I_j} f(U_h) \partial_x v_h dx + \hat{f}^l_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}^r_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) = \int_{I_j} g(h_h, b_h) v_h dx. \tag{6.1}$$

Comparing with the standard RKDG scheme (2.22), we can see that the single valued fluxes $\hat{f}_{j+\frac{1}{2}}$ and $\hat{f}_{j-\frac{1}{2}}$ have been replaced by the left flux $\hat{f}^l_{j+\frac{1}{2}}$ and the right flux $\hat{f}^r_{j-\frac{1}{2}}$, respectively. We can rewrite the above scheme as:

$$\int_{I_j} \partial_t(hu)_h v_h dx - \int_{I_j} f(U_h)\partial_x v_h dx + \hat{f}_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) = \tag{6.2}$$
$$\int_{I_j} g(h_h, b_h) v_h dx + (\hat{f}_{j+\frac{1}{2}} - \hat{f}^l_{j+\frac{1}{2}}) v_h(x^-_{j+\frac{1}{2}}) - (\hat{f}_{j-\frac{1}{2}} - \hat{f}^r_{j-\frac{1}{2}}) v_h(x^+_{j-\frac{1}{2}}),$$

where $\hat{f}_{j+\frac{1}{2}} = F(U_h(x^-_{j+\frac{1}{2}}, t), U_h(x^+_{j+\frac{1}{2}}, t))$. The left side of (6.2) is the traditional RKDG scheme, and the right side is our approximation to the source term. The design of the left flux $\hat{f}^l_{j+\frac{1}{2}}$ and the right flux $\hat{f}^r_{j-\frac{1}{2}}$ will be explained later, however we point out here that $\hat{f}_{j+\frac{1}{2}} - \hat{f}^l_{j+\frac{1}{2}}$ and $\hat{f}_{j-\frac{1}{2}} - \hat{f}^r_{j-\frac{1}{2}}$ are high order correction terms at the level of $O(\triangle x^{k+1})$ and stay bounded when the numerical solution itself is bounded as dx is refined. Therefore, the scheme (6.1) is a $(k+1)$-th order conservative scheme and will converge to the weak solution.

In order to obtain the well balanced property, we need the residue

$$R = -\int_{I_j} f(U_h)\partial_x v_h dx + \hat{f}^l_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}^r_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) - \int_{I_j} g(h_h, b_h) v_h dx \tag{6.3}$$

to be zero if the still water stationary state (1.5) is reached. The following three conditions, *which only need to be valid for the still water stationary state*, are sufficient to guarantee this zero residue property.

- All the integrals in formula (6.3) should be calculated exactly for the still water. This can be easily achieved by using suitable Gauss-quadrature rules since h_h, b_h and v_h are polynomials in each cell I_j, hence f, g are both polynomials. Note that $(hu)_h = 0$ for the still water.

- We assume that

$$\hat{f}^l_{j+\frac{1}{2}} = f(U_h(x^-_{j+\frac{1}{2}}, t)), \qquad \hat{f}^r_{j-\frac{1}{2}} = f(U_h(x^+_{j-\frac{1}{2}}, t)) \tag{6.4}$$

for the still water. Note that this condition is not obvious. Later we will comment on how to make it possible for the RKDG method.

- We assume that U_h, which is the numerical approximation of U, is a steady state solution of the equation $(hu)_t + f(U)_x = g(h, b_h)$, where b_h has substituted b. This is true since $h_h + b_h = constant$ and $(hu)_h = 0$, which imply $\left(\frac{1}{2}gh_h^2\right)_x = -gh_h(b_h)_x$, or
$$\partial_x f(U_h) = g(h_h, b_h). \tag{6.5}$$

Proposition 6.1.1 *RKDG schemes which satisfy the above three conditions for the shallow water equations are exact for the still water stationary state (1.5).*

Proof. If these three conditions are satisfied, the residue R in (6.3) for still water reduces to

$$\begin{aligned} R &= -\int_{I_j} f(U_h)\partial_x v_h dx + \hat{f}^l_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) - \hat{f}^r_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) - \int_{I_j} g(h_h, b_h) v_h dx \\ &= -\int_{I_j} f(U_h)\partial_x v_h dx + f(U_h(x^-_{j+\frac{1}{2}}, t)) v_h(x^-_{j+\frac{1}{2}}) \\ &\quad -f(U_h(x^+_{j-\frac{1}{2}}, t)) v_h(x^+_{j-\frac{1}{2}}) - \int_{I_j} g(h_h, b_h) v_h dx \\ &= \int_{I_j} \partial_x f(U_h) v_h dx - \int_{I_j} g(h_h, b_h) v_h dx \\ &= \int_{I_j} (\partial_x f(U_h) - g(h_h, b_h)) v_h dx = 0 \end{aligned}$$

where the second equality is due to (6.4), the third equality follows from a simple integration by parts, and the last equality follows from (6.5). □

Remark 6.1.2 *For discontinuous solutions, the limiter on the function U_h is usually performed after each Runge-Kutta stage. This limiter procedure might destroy the preservation of the still water steady state $h + b = constant$. Therefore, following the idea presented in [2, 51], we apply the limiter procedure on the function $(h_h +$*

$b_h, (hu)_h)^T$ instead. The modified RKDG solution is then defined by $h_h^{mod} \equiv (h+b)_h^{mod} - b_h$. Since $\overline{h_h}^{mod} = \overline{(h+b)_h}^{mod} - \overline{b_h} = \overline{(h+b)_h} - \overline{b_h} = \overline{h_h}$, we observe that this procedure will not destroy the conservativity of h_h, which should be maintained during the limiter process.

For shallow water equations, the first and third conditions are obviously true, hence the only one remaining for us to check is the second condition. In order to fulfill it, we have two choices.

Choice A: Define the initial value and the approximation b_h by continuous piecewise polynomials. We would then have $b_h(x^-_{j+\frac{1}{2}}, t) = b_h(x^+_{j+\frac{1}{2}}, t)$. If the steady state $h_h + b_h = constant$ is reached, we will have a continuous h_h, i.e. $h_h(x^-_{j+\frac{1}{2}}, t) = h_h(x^+_{j+\frac{1}{2}}, t)$, which makes

$$\hat{f}_{j+\frac{1}{2}} = F(U_h(x^-_{j+\frac{1}{2}}, t), U_h(x^+_{j+\frac{1}{2}}, t)) = f(U_h(x^-_{j+\frac{1}{2}}, t)) = f(U_h(x^+_{j+\frac{1}{2}}, t))$$

We can therefore simply define the left and right fluxes as

$$\hat{f}^l_{j+\frac{1}{2}} = \hat{f}_{j+\frac{1}{2}}, \qquad \hat{f}^r_{j-\frac{1}{2}} = \hat{f}_{j-\frac{1}{2}},$$

which will fulfill the second condition. This make our scheme (6.1) to be identical to the traditional RKDG scheme without any modification.

In order to define the continuous piecewise polynomial approximations to the initial value and b, we can use the idea of essentially non-oscillatory (ENO) procedure [19]. Based on the values $u_{j+\frac{1}{2}}$, we can choose suitable stencils for each individual cell I_j by an ENO procedure, and then obtain a polynomial on I_j through an interpolation. This polynomial equals to $u_{j-\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$ at the two cell boundaries. Hence the global piecewise polynomial is continuous.

Choice B: Here we follow the idea of Audusse et al. [2]. After computing boundary values $U^{\pm}_{h,j+\frac{1}{2}}$, we set

$$h^{*,+}_{h,j+\frac{1}{2}} = \max\left(0, h^{+}_{h,j+\frac{1}{2}} + b^{+}_{h,j+\frac{1}{2}} - \max(b^{+}_{h,j+\frac{1}{2}}, b^{-}_{h,j+\frac{1}{2}})\right) \tag{6.6}$$

$$h^{*,-}_{h,j+\frac{1}{2}} = \max\left(0, h^{-}_{h,j+\frac{1}{2}} + b^{-}_{h,j+\frac{1}{2}} - \max(b^{+}_{h,j+\frac{1}{2}}, b^{-}_{h,j+\frac{1}{2}})\right) \tag{6.7}$$

and redefine the left and right values of U as:

$$U^{*,\pm}_{h,j+\frac{1}{2}} = \begin{pmatrix} h^{*,\pm}_{h,j+\frac{1}{2}} \\ (hu)^{\pm}_{h,j+\frac{1}{2}} \end{pmatrix} \tag{6.8}$$

Then the left and right fluxes $\hat{f}^{l}_{j+\frac{1}{2}}$ and $\hat{f}^{r}_{j-\frac{1}{2}}$ are given by:

$$\hat{f}^{l}_{j+\frac{1}{2}} = F(U^{*,-}_{h,j+\frac{1}{2}}, U^{*,+}_{h,j+\frac{1}{2}}) + \begin{pmatrix} 0 \\ \frac{g}{2}(h^{-}_{h,j+\frac{1}{2}})^2 - \frac{g}{2}(h^{*,-}_{h,j+\frac{1}{2}})^2 \end{pmatrix} \tag{6.9}$$

$$\hat{f}^{r}_{j-\frac{1}{2}} = F(U^{*,-}_{h,j-\frac{1}{2}}, U^{*,+}_{h,j-\frac{1}{2}}) + \begin{pmatrix} 0 \\ \frac{g}{2}(h^{+}_{h,j-\frac{1}{2}})^2 - \frac{g}{2}(h^{*,+}_{h,j-\frac{1}{2}})^2 \end{pmatrix} \tag{6.10}$$

Here F is a monotone flux as mentioned in Section 2.2. It is easy to check that $\hat{f}_{j+\frac{1}{2}} - \hat{f}^{l}_{j+\frac{1}{2}}$ and $\hat{f}_{j-\frac{1}{2}} - \hat{f}^{r}_{j-\frac{1}{2}}$ are indeed at the level of $O(\triangle x^{k+1})$ for general solutions, hence the original $(k+1)$-th order of accuracy is maintained. Under the still water stationary state, $h_h + b_h = constant$, hence it is easy to see $U^{*,-}_{h,j+\frac{1}{2}} = U^{*,+}_{h,j+\frac{1}{2}}$. The left flux then returns to:

$$\begin{aligned} \hat{f}^{l}_{j+\frac{1}{2}} &= \begin{pmatrix} 0 \\ \frac{g}{2}(h^{*,-}_{h,j+\frac{1}{2}})^2 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{g}{2}(h^{-}_{h,j+\frac{1}{2}})^2 - \frac{g}{2}(h^{*,-}_{h,j+\frac{1}{2}})^2 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \frac{g}{2}(h^{-}_{h,j+\frac{1}{2}})^2 \end{pmatrix} = f(U^{-}_{h,j+\frac{1}{2}}) \end{aligned}$$

Similarly,

$$\hat{f}^r_{j+\frac{1}{2}} = f(U^+_{h,j+\frac{1}{2}}). \tag{6.11}$$

Remark 6.1.3 *Clearly Choice A provides a simpler scheme with smaller computational cost, hence it would be preferred if it provides comparable numerical results to that of Choice B. Unfortunately, although it works well for small perturbation solutions from still water for a smooth bottom, the numerical resolution for a discontinuous bottom is not ideal. On the other hand, Choice B provides good numerical results for all the test cases we have experimented. In Section 6.3, we will report only the results obtained by Choice B to save space.*

We now consider the extension of the well-balanced high order RKDG schemes to 2-D shallow water equations

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0 \\ (hu)_t + \left(hu^2 + \dfrac{1}{2}gh^2\right)_x + (huv)_y = -ghb_x \\ (hv)_t + (huv)_x + \left(hv^2 + \dfrac{1}{2}gh^2\right)_y = -ghb_y \end{cases} \tag{6.12}$$

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. The still water stationary solution we are interested to preserve is

$$h + b = constant, \qquad hu = 0, \qquad hv = 0. \tag{6.13}$$

It is straightforward to extend our well balanced RKDG schemes to this two-dimensional problem. Also, the scheme can be applied on any triangulation.

6.2 Construction of well balanced finite volume WENO schemes for shallow water equations

In this section, we generalize the idea used in Section 6.1 to design a well balanced finite volume WENO schemes for the shallow water equations. The basic idea is the same as that for the RKDG methods. The only extra step is due to the fact that we only have the reconstructed pointwise values $U^{\pm}_{j+\frac{1}{2}}$ and would need to first define an approximation function U_h. We can then follow the procedure as before.

As before, we denote $U^{\pm}_{j+\frac{1}{2}}$ as the reconstructed left and right values at the interface $x_{j+\frac{1}{2}}$. The still water is given by $\bar{h}_j + \bar{b}_j = constant$ where as before the overbar denotes the cell average. We would like to have the reconstructed values to satisfy $h_{j+\frac{1}{2}} + b_{j+\frac{1}{2}} = constant$ as well. This can be achieved by the approach that we used in [48]. Basically, the WENO reconstruction can be eventually written out as

$$U^{+}_{j+\frac{1}{2}} = \sum_{k=-r+1}^{r} w_k \bar{U}_{j+k}, \qquad U^{-}_{j+\frac{1}{2}} = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{U}_{j+k}. \tag{6.14}$$

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 $\bar{u}$ and satisfy $\sum_{k=-r+1}^{r} w_k = \sum_{k=-r}^{r-1} \tilde{w}_k = 1$. We then use the same coefficients w_k and $\tilde{w}_k$ computed from above on $B = (b, 0)^T$ to obtain

$$B^{+}_{j+\frac{1}{2}} = \sum_{k=-r+1}^{r} w_k \bar{B}_{j+k}, \qquad B^{-}_{j+\frac{1}{2}} = \sum_{k=-r}^{r-1} \tilde{w}_k \bar{B}_{j+k}. \tag{6.15}$$

Hence,

$$U^{+}_{j+\frac{1}{2}} + B^{+}_{j+\frac{1}{2}} = \sum_{k=-r+1}^{r} w_k (\bar{U}_{j+k} + \bar{B}_{j+k}), \qquad U^{-}_{j+\frac{1}{2}} + B^{-}_{j+\frac{1}{2}} = \sum_{k=-r}^{r-1} \tilde{w}_k (\bar{U}_{j+k} + \bar{B}_{j+k}),$$

from which we know that the reconstructed values satisfy $h^{\pm}_{j+\frac{1}{2}} + b^{\pm}_{j+\frac{1}{2}} = constant$

for still water.

For the well balanced property, we only need to consider the second equation of (1.3). Our well balanced finite volume WENO scheme is given by

$$\triangle x_j \frac{d}{dt}\overline{hu}_j(t) = -\left(\hat{f}^l_{j+\frac{1}{2}} - \hat{f}^r_{j-\frac{1}{2}}\right) + \int_{I_j} g(h,b)dx \tag{6.16}$$

where $\hat{f}^l_{j+\frac{1}{2}}$ and $\hat{f}^r_{j-\frac{1}{2}}$ are the left and right fluxes as defined in Section 6.1. The residue R is denoted by the right side of the equation (6.16).

Here we also give three conditions which need to be valid for the still water stationary state (1.5):

- We use interpolation to obtain a high order polynomial h_h on the cell I_j, based on the boundary values $h^+_{j-\frac{1}{2}}$, $h^-_{j+\frac{1}{2}}$ and several other neighboring boundary values. For example, we can use $h^-_{j+\frac{3}{2}}$, $h^-_{j+\frac{1}{2}}$, $h^+_{j-\frac{1}{2}}$ and $h^+_{j-\frac{3}{2}}$ to interpolate a third degree polynomial. Similarly, we can use the same interpolation on b to obtain a polynomial b_h, and then use them to compute $\int_{I_j} g(h_h, b_h)dx$ exactly by using a suitable Gauss quadrature. In order to obtain $(2k)$-th order accuracy for the approximation of the source term, h_h and b_h need to approximate h and b with $(k+1)$-th order accuracy. We observe from the definition of h_h and b_h that the interpolated polynomials satisfy the following properties:

$$h_h(x_{j+\frac{1}{2}}) = h^-_{j+\frac{1}{2}}, \quad h_h(x_{j-\frac{1}{2}}) = h^+_{j-\frac{1}{2}}, \quad b_h(x_{j+\frac{1}{2}}) = b^-_{j+\frac{1}{2}}, \quad b_h(x_{j-\frac{1}{2}}) = b^+_{j-\frac{1}{2}},$$

$$h_h + b_h = constant \qquad \text{if} \qquad h^\pm_{j+\frac{1}{2}} + b^\pm_{j+\frac{1}{2}} = constant.$$

- We assume that the left and right fluxes $\hat{f}^l_{j+\frac{1}{2}}$ and $\hat{f}^r_{j-\frac{1}{2}}$ satisfy (6.4).

- We assume that interpolated polynomial $U_h(x,t)$ above, is a steady state solution of the equation $(hu)_t + f(U)_x = g(h, b_h)$, where b_h has substituted b. As before, this is true since $h_h + b_h = constant$ and $(hu)_h = 0$.

Proposition 6.2.1 *Finite volume WENO schemes which satisfy the above three conditions for shallow water equations are well balanced for still water stationary state (1.5).*

Proof. If these three conditions are satisfied, the residue R for still water reduces to

$$\begin{aligned} R &= -\hat{f}^l_{j+\frac{1}{2}} + \hat{f}^r_{j-\frac{1}{2}} + \int_{I_j} g(h_h, b_h)dx \\ &= -f(U_h(x^-_{j+\frac{1}{2}}, t)) + f(U_h(x^+_{j-\frac{1}{2}}, t)) + \int_{I_j} g(h_h, b_h)dx \\ &= -\int_{I_j} \partial_x f(U_h)dx + \int_{I_j} g(h_h, b_h)dx \\ &= -\int_{I_j} (\partial_x f(U_h) - g(h_h, b_h))dx = 0 \end{aligned}$$

where the second equality is due to (6.4), and the last equality follows from (6.5). □

Note that the first and third conditions are obviously true for the still water of the shallow water equations. As to the second one, we follow Choice B in Section 6.1, i.e. (6.8), (6.9) and (6.10).

Remark 6.2.2 *During the WENO reconstruction, we reconstruct k polynomials on the cell I_j, based on different stencils, and then define the boundary values $u^-_{j+\frac{1}{2}}$ and $u^+_{j-\frac{1}{2}}$ as convex combinations of those polynomials. We emphasize that such convex combination, when viewed as a function, is* not *a polynomial on I_j, due to the nonlinear nature of the weights. Therefore, the interpolated polynomial u_h is* not *that convex combination and must be recomputed. However, if we use ENO instead of WENO schemes, we can directly take u_h as the ENO reconstructed polynomial on I_j, thereby saving the computational cost to obtain it again by interpolation.*

Remark 6.2.3 *Audusse et al. [2] introduced a second order well balanced finite volume scheme, and recently, Noelle et al. [30] generalized it to higher order accuracy.*

The idea proposed here is a generalization of these schemes, by allowing more freedom in defining the polynomials h_h and b_h to save computational cost. If we interpolate h_h and b_h based on the two boundary values plus the center value of the cell I_j (which must be reconstructed), this will give us the fourth order discretization of the source term as introduced in [30].

Now let us consider the 2D shallow water equations (6.12) with the still water stationary solution (6.13) to be balanced. It is straightforward to extend our well balanced WENO schemes to this two-dimensional problem, at least for rectangular meshes. Let us look at the second equation in (6.12) for instance. As we mentioned in Section 2.2, the numerical scheme is given by:

$$\frac{d}{dt}\bar{u}_{ij}(t) = -\frac{1}{\triangle x_i}\left((\hat{f}_1)^l_{i+\frac{1}{2},j} - (\hat{f}_1)^r_{i-\frac{1}{2},j}\right) - \frac{1}{\triangle y_j}\left((\hat{f}_2)^l_{i,j+\frac{1}{2}} - (\hat{f}_2)^r_{i,j-\frac{1}{2}}\right) + g_{i,j}, \quad (6.17)$$

with

$$(\hat{f}_1)^{l,r}_{i+\frac{1}{2},j} = \sum_\alpha w_\alpha(\hat{f}_1(\cdot, y_j + \beta_\alpha \triangle y_j))^{l,r}_{i+\frac{1}{2}} \quad (6.18)$$

and similarly for $(\hat{f}_2)^{l,r}_{i,j+\frac{1}{2}}$, and

$$\begin{aligned} g_{i,j} &= -\frac{1}{\triangle x_i}\sum_\alpha w_\alpha\left(\int_{I_i} g\,(hb_x)(x, y_j + \beta_\alpha \triangle y_j)dx\right) \quad (6.19)\\ &\approx -\frac{1}{\triangle x_i \triangle y_j}\int_{I_i}\int_{I_j} ghb_x dxdy \end{aligned}$$

where the first quadrature summation in the y direction must be accurate to the order of the scheme and the integration in the x direction must be computed exactly (by Gauss-quadrature with enough exactness). If the still water stationary solution

(6.13) is given, the right side of the numerical scheme (6.17) becomes:

$$\begin{aligned} & -\frac{1}{\triangle x_i}\left((\hat{f}_1)^l_{i+\frac{1}{2},j} - (\hat{f}_1)^r_{i-\frac{1}{2},j}\right) - \frac{1}{\triangle x_i}\sum_\alpha w_\alpha \left(\int_{I_i} g\,(hb_x)(x, y_j + \beta_\alpha \triangle y_j)dx\right) \\ = & -\frac{1}{\triangle x_i}\sum_\alpha w_\alpha \left((\hat{f}_1(\cdot, y_j + \beta_\alpha \triangle y_j))^l_{i+\frac{1}{2}} - (\hat{f}_1(\cdot, y_j + \beta_\alpha \triangle y_j))^r_{i-\frac{1}{2}}\right) \\ & -\frac{1}{\triangle x_i}\sum_\alpha w_\alpha \left(\int_{I_i} g\,(hb_x)(x, y_j + \beta_\alpha \triangle y_j)dx\right), \end{aligned}$$

We can balance $(\hat{f}_1(\cdot, y_j + \beta_\alpha \triangle y_j))^l_{i+\frac{1}{2}} - (\hat{f}_1(\cdot, y_j + \beta_\alpha \triangle y_j))^r_{i-\frac{1}{2}}$ with $\int_{I_i} g\,(hb_x)(x, y_j + \beta_\alpha \triangle y_j)dx$ for each fixed α, by the same technique used in the 1D case. This means that at each Gauss point in the y direction, we interpolate polynomials as functions of x, and use them to compute the source term. Well balanced property is thus obtained. Similarly, we can handle the third equation in (6.12) in the same fashion.

Remark 6.2.4 *Both the well-balanced RKDG and finite volume WENO schemes are developed here. The RKDG schemes involve less modification for the well balanced property to hold, and are more flexible for general geometry, adaptivity and parallel implementation. On the other hand, the RKDG schemes rely on limiters to control spurious oscillations for discontinuous solutions, which are less robust than the WENO reconstruction procedure in the capability of maintaining accuracy in smooth regions and controlling oscillations for strong discontinuities simultaneously. We refer to [52] for a comparison of these two types of schemes.*

6.3 Numerical results for the shallow water equations

In this section we provide numerical results to demonstrate the good properties of the well balanced finite volume WENO and finite element RKDG schemes when applied to the shallow water equations. Fifth order finite volume WENO scheme and third order finite element RKDG scheme are implemented as examples. In all numerical tests, time discretization is by the third order TVD Runge-Kutta method in [41]. For finite volume 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 [39, 11] 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.

6.3.1 Test for the exact C-property

The purpose of the first test problem is to verify that the schemes indeed maintain the exact C-property over a non-flat bottom. We choose two different functions for the bottom topography given by ($0 \leq x \leq 10$):

$$b(x) = 5\, e^{-\frac{2}{5}(x-5)^2}, \tag{6.20}$$

which is smooth, and

$$b(x) = \begin{cases} 4 & \text{if } 4 \leq x \leq 8 \\ 0 & \text{otherwise,} \end{cases} \tag{6.21}$$

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 exact C-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 6.1 and 6.2 for the two bottom functions (6.20) and (6.21) and different precisions. For the RKDG 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 exact C-property.

Table 6.1: L^1 and L^∞ errors for different precisions for the stationary solution with a smooth bottom (3.16).

		L^1 error		L^∞ error	
	precision	h	hu	h	hu
FV	single	3.00E-05	1.10E-04	4.39E-05	5.19E-04
	double	5.04E-14	2.99E-13	1.12E-13	1.26E-12
	quadruple	6.48E-33	3.45E-32	2.17E-32	1.54E-31
RKDG	single	8.41E-06	3.15E-05	3.72E-05	2.06E-04
	double	3.02E-15	3.59E-15	1.60E-14	7.22E-14
	quadruple	8.06E-31	2.92E-33	8.05E-29	1.07E-31

Table 6.2: L^1 and L^∞ errors for different precisions for the stationary solution with a nonsmooth bottom (6.21).

		L^1 error		L^∞ error	
	precision	h	hu	h	hu
FV	single	1.80E-05	1.40E-04	3.24E-05	2.41E-04
	double	4.41E-14	2.57E-13	1.05E-13	1.30E-12
	quadruple	4.27E-32	3.71E-31	1.07E-31	1.46E-30
RKDG	single	5.72E-07	1.22E-07	9.54E-07	3.41E-07
	double	1.40E-15	3.16E-16	3.55E-15	7.77E-15
	quadruple	8.06E-31	1.65E-34	8.06E-29	4.12E-33

We have also computed stationary solutions using initial conditions which are not the still water stationary solutions and letting time evolve into a still water stationary solution, obtaining similar results with the exact C-property, i.e. the errors are at the level of round-off errors for different precisions.

6.3.2 Testing the orders of accuracy

In this example we will test the high order accuracy of our schemes for a smooth solution. We have the following bottom function and initial conditions

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

with periodic boundary conditions, see [46]. 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). Table 6.3 contains the L^1 errors for the cell averages and numerical orders of accuracy for the finite volume and RKDG schemes, respectively. We can clearly see that fifth order accuracy is achieved for the WENO scheme, and third order accuracy is achieved for the RKDG scheme. 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 Table 6.3. For the RKDG method, the CFL number is fixed at 0.18.

6.3.3 A small perturbation of a steady-state water

The following quasi-stationary test case was proposed by LeVeque [28]. 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.

The bottom topography consists of one hump:

$$b(x) = \begin{cases} 0.25(\cos(10\pi(x-1.5))+1) & \text{if } 1.4 \le x \le 1.6 \\ 0 & \text{otherwise} \end{cases} \tag{6.22}$$

Table 6.3: L^1 errors and numerical orders of accuracy for the example in Section 6.3.2.

	FV schemes				
No. of	CFL	h		hu	
cells		L^1 error	order	L^1 error	
25	0.6	1.28E-02		1.16E-01	
50	0.6	2.25E-03	2.50	2.25E-02	2.37
100	0.4	3.26E-04	2.79	2.75E-03	3.03
200	0.3	2.33E-05	3.80	2.00E-04	3.79
400	0.2	9.54E-07	4.61	8.20E-06	4.60
800	0.1	2.99E-08	4.99	2.58E-07	4.99
	RKDG schemes				
No. of		h		hu	
cells		L^1 error	order	L^1 error	order
25		2.35E-03		2.12E-02	
50		1.14E-04	4.36	1.01E-03	4.39
100		1.24E-05	3.20	1.09E-04	3.21
200		1.02E-06	3.59	8.97E-06	3.60
400		1.12E-07	3.19	9.79E-07	3.19
800		1.30E-08	3.09	1.14E-07	3.08

The initial conditions are given with

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 1 - b(x) + \epsilon & \text{if } 1.1 \leq x \leq 1.2 \\ 1 - b(x) & \text{otherwise} \end{cases} \tag{6.23}$$

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 [28]. Both sets of initial conditions are shown in Figure 6.1. The solution at time t=0.2s for the big pulse $\epsilon = 0.2$, obtained on a 200 cell uniform grid with simple transmissive boundary conditions, and compared with

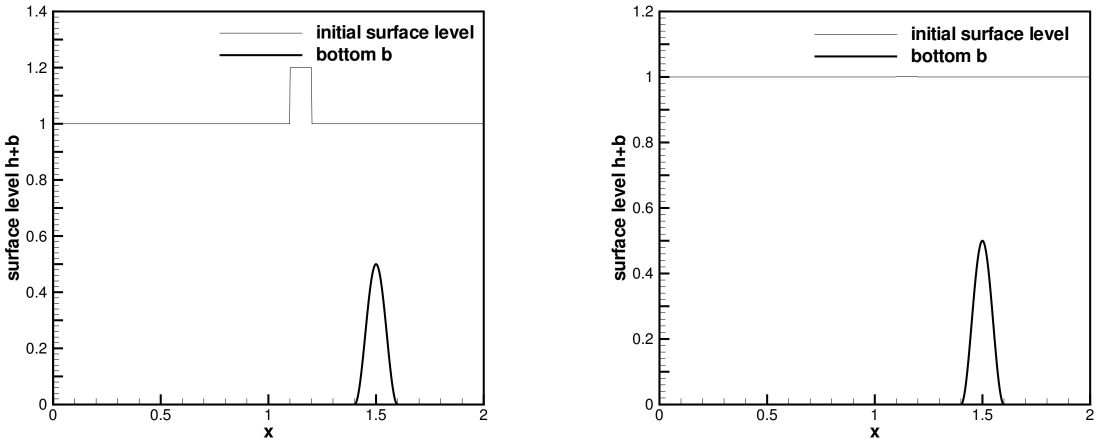


Figure 6.1: The initial surface level $h + b$ and the bottom b for a small perturbation of a steady-state water. Left: a big pulse ϵ=0.2; right: a small pulse ϵ=0.001.

a 3000 cell solution, is shown in Figure 6.2 for the FV scheme and in Figure 6.4 for the RKDG scheme. The results for the small pulse $\epsilon = 0.001$ are shown in Figures 6.3 and 6.5. At this time, the downstream-traveling water pulse has already passed the bump. We can clearly see that there are no spurious numerical oscillations.

6.3.4 The dam breaking problem over a rectangular bump

In this example we 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 [44].

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leq 1500/8 \\ 0 & \text{otherwise} \end{cases} \tag{6.24}$$

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

$$(hu)(x, 0) = 0 \quad \text{and} \quad h(x, 0) = \begin{cases} 20 - b(x) & \text{if } x \leq 750 \\ 15 - b(x) & \text{otherwise} \end{cases} \tag{6.25}$$

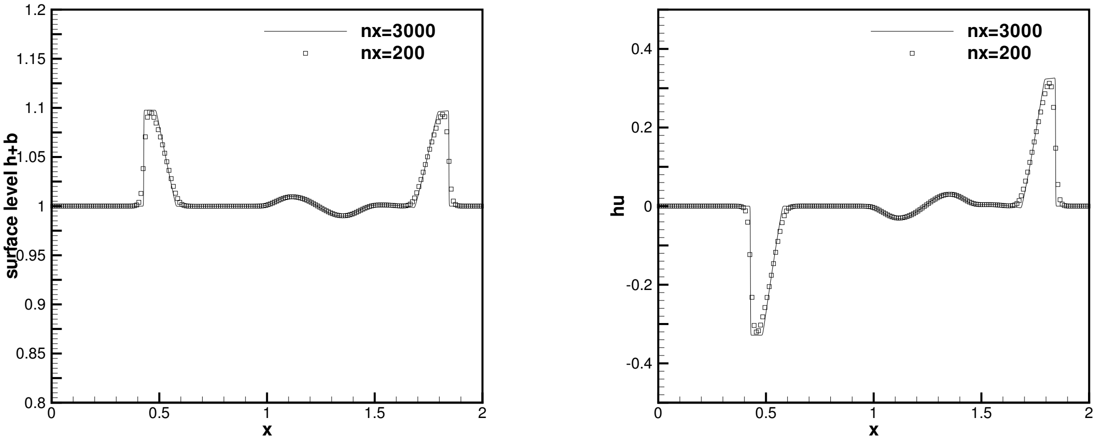


Figure 6.2: FV scheme: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

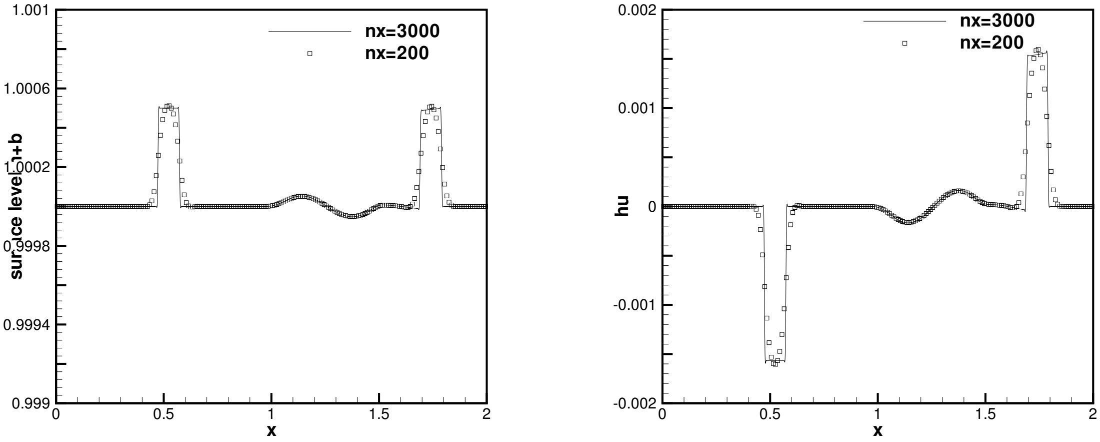


Figure 6.3: FV scheme: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

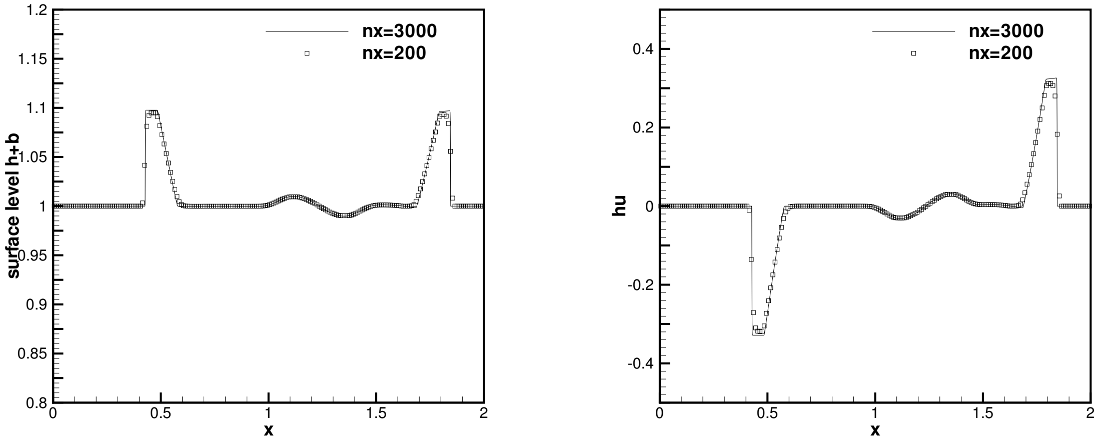


Figure 6.4: RKDG scheme: Small perturbation of a steady-state water with a big pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

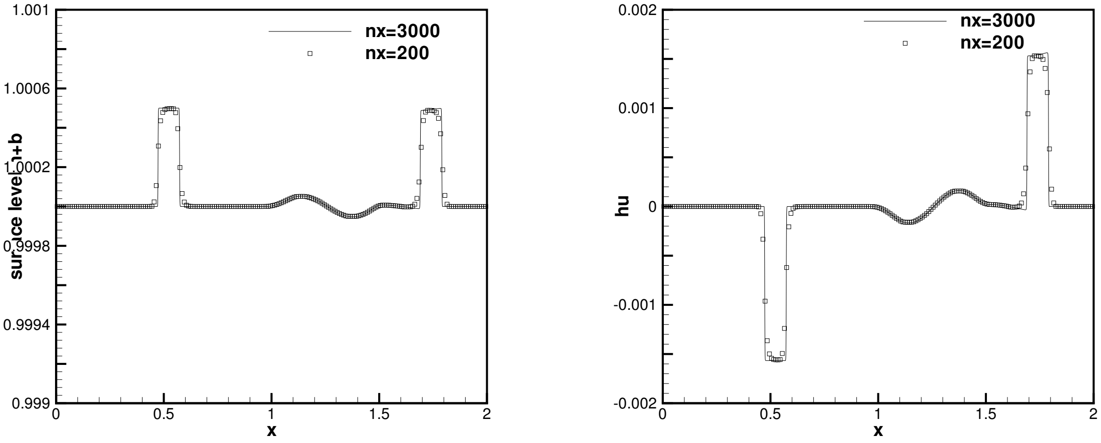


Figure 6.5: RKDG scheme: Small perturbation of a steady-state water with a small pulse. t=0.2s. Left: surface level $h + b$; right: the discharge hu.

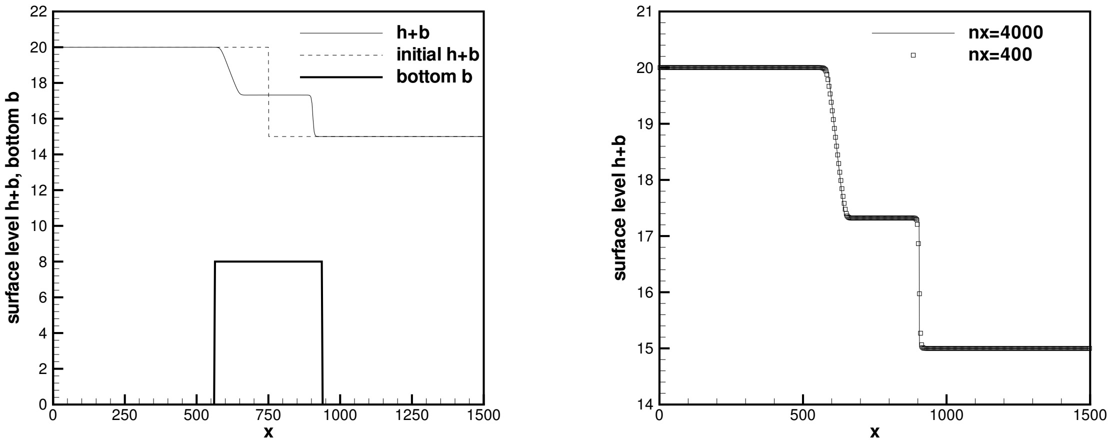


Figure 6.6: FV scheme: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

The numerical results obtained by the FV scheme with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Figures 6.6 and 6.7, with two different ending time t=15s and t=60s. Figures 6.8 and 6.9 demonstrate the numerical results by the RKDG scheme, with the same number of uniform cells. 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. Both schemes work well for this example, giving well resolved, non-oscillatory solutions using 400 cells which agree with the converged results using 4000 cells.

6.3.5 Steady flow over a hump

The purpose of this test case is to study the convergence in time towards steady flow over a bump. These are classical test problems for transcritical and subcritical flows, and they are widely used to test numerical schemes for shallow water equations. For example, they have been considered by the *working group on dam break modelling* [17], and have been used as a test case in, e.g. [43].

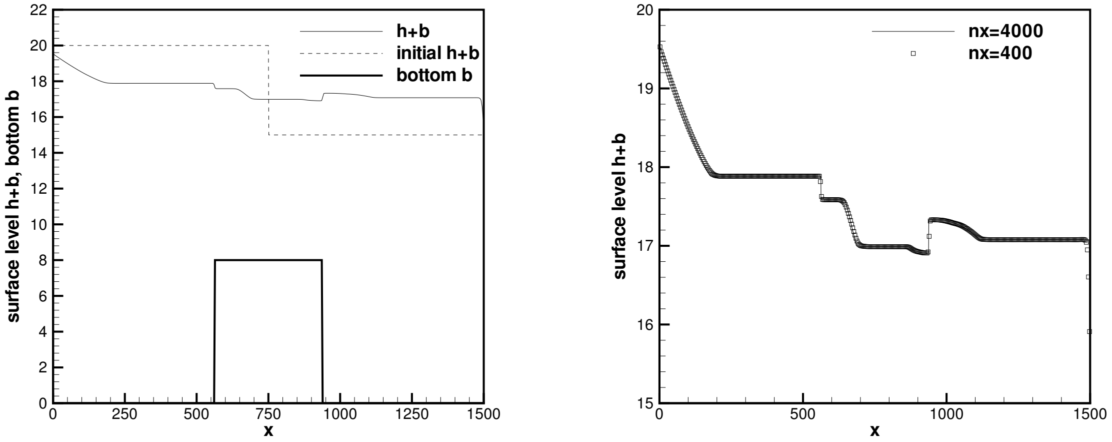


Figure 6.7: FV scheme: The surface level $h + b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

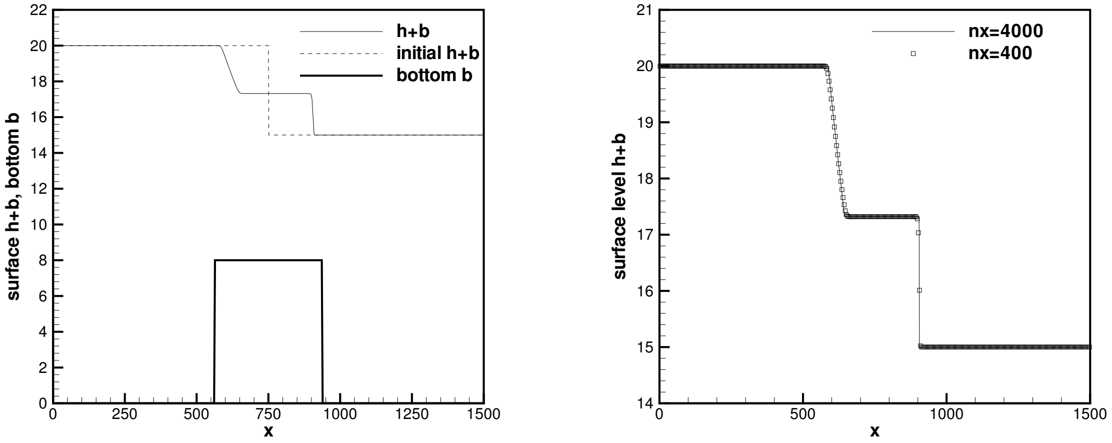


Figure 6.8: RKDG scheme: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

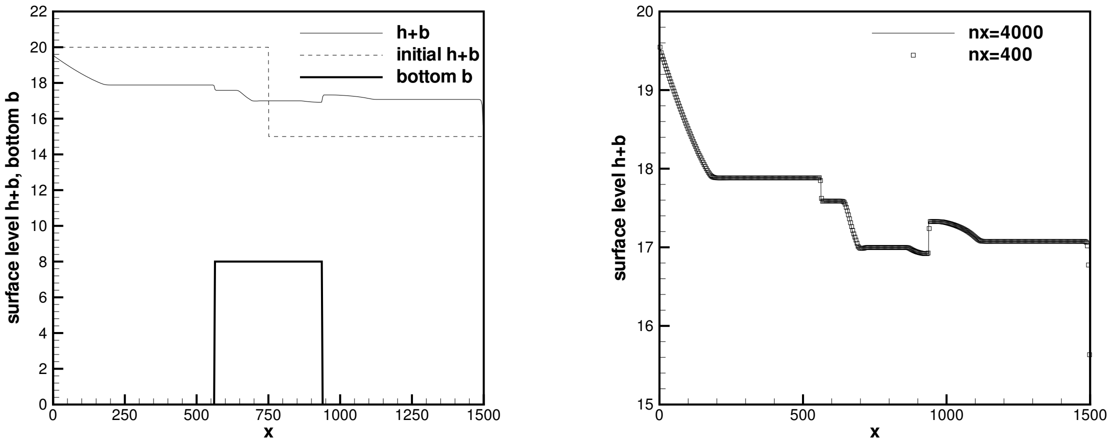


Figure 6.9: RKDG scheme: The surface level $h+b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

The bottom function is given by:

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12 \\ 0 & \text{otherwise} \end{cases} \tag{6.26}$$

for a channel of length $25m$. The initial conditions are taken as

$$h(x,0) = 0.5 - b(x) \quad \text{and} \quad u(x,0) = 0.$$

Depending on different boundary conditions, the flow can be subcritical or transcritical with or without a steady shock. The computational parameters common for all three cases are: uniform mesh size Δx = 0.125 m (200 cells), ending time t= 200 s. Analytical solutions for the various cases are given in Goutal and Maurel [17].

a): Transcritical flow without a shock.

- 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 subcrit-

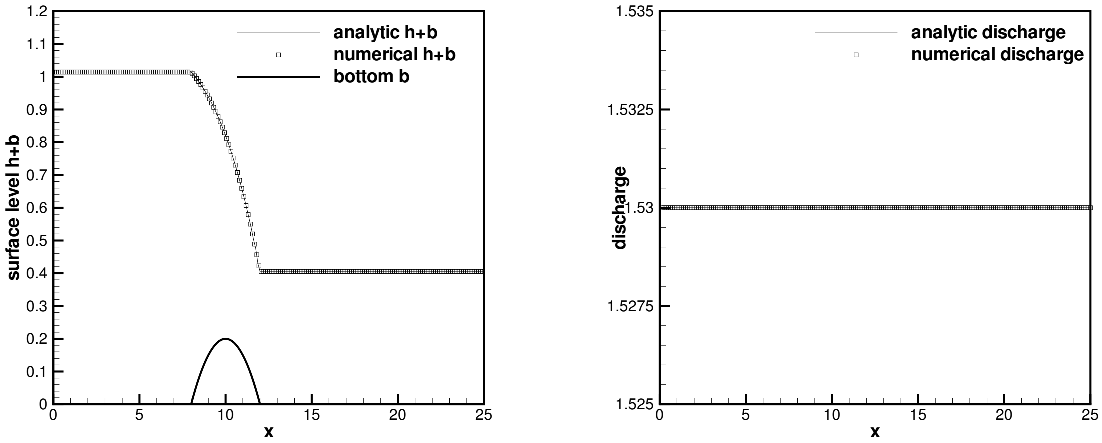


Figure 6.10: FV scheme: Steady transcritical flow over a bump without a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

ical.

The surface level $h + b$ and the discharge hu, as the numerical flux for the water height h in equation (1.3), are plotted in Figures 6.10 and 6.11, which show very good agreement with the analytical solution. The correct capturing of the discharge hu is usually more difficult than the surface level $h+b$, as noticed by many authors.

b): Transcritical flow with a shock.

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

In this case, the Froude number $Fr = u/\sqrt{gh}$ increases to a value larger than one above the bump, and then decreases to less than one. A stationary shock can appear on the surface. The surface level $h + b$ and the discharge hu, as the numerical flux for the water height h in equation (1.3), are plotted in Figure 6.12 and 6.13. In Figure 6.12 for the FV scheme, some minor oscillations can be observed near the jump. Here we also plot the numerical result with 400 uniform cells in Figure 6.14,

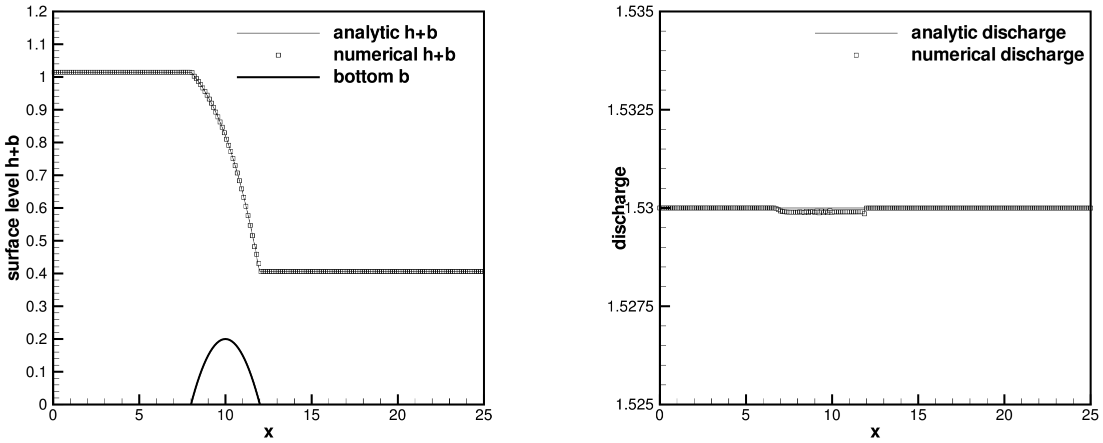


Figure 6.11: RKDG scheme: Steady transcritical flow over a bump without a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

where we can observe that the oscillation is completely removed. The reason for this phenomenon is still not known.

c): Subcritical flow.

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

- downstream: The water height h=2 m is imposed.

This is a subcritical flow. The surface level $h + b$ and the discharge hu, as the numerical flux for the water height h in equation (1.3), are plotted in Figure 6.15 and 6.16, which are in good agreement with the analytical solution.

6.3.6 Test for the exact C-property in two dimensions

This example is used to check that our schemes indeed maintain the exact C-property over a non-flat bottom for 2D shallow water equations. The two-dimensional hump

$$b(x, y) = 0.8e^{-50((x-0.5)^2+(y-0.5)^2)}, \qquad x, y \in [0, 1] \tag{6.27}$$

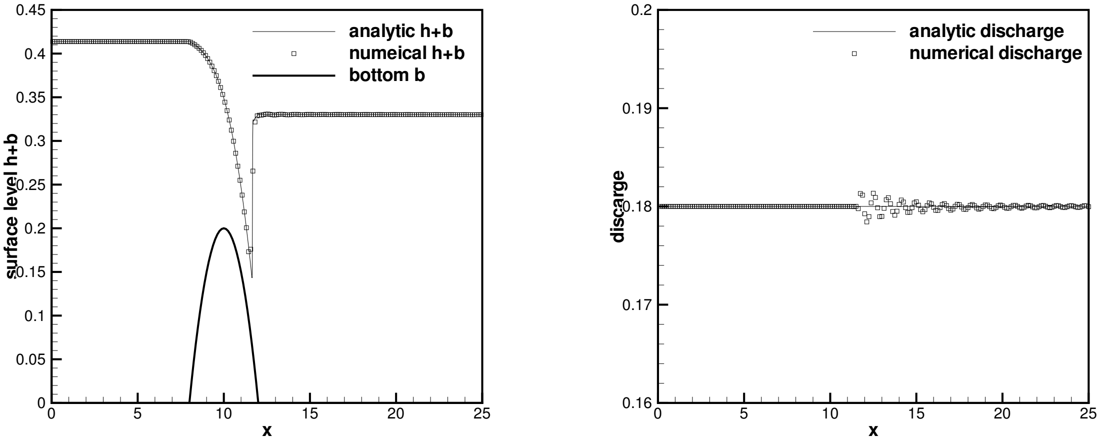


Figure 6.12: FV scheme: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

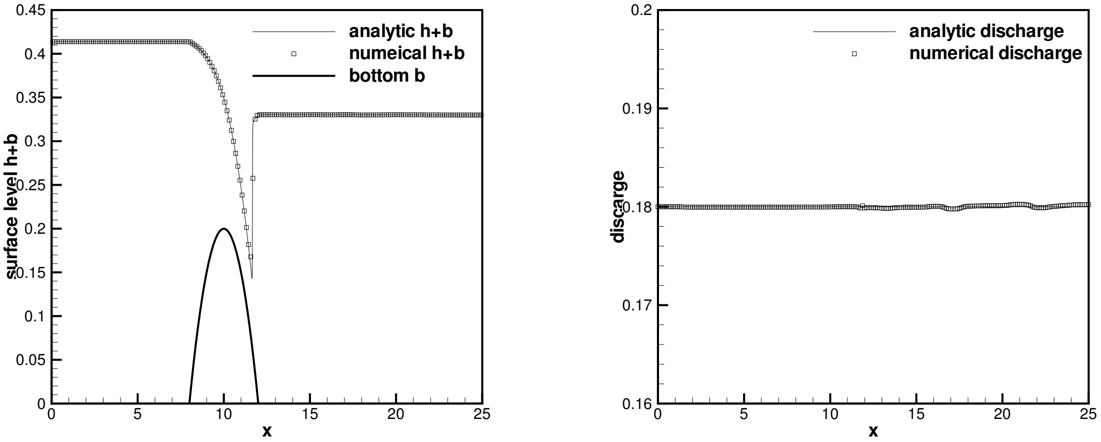


Figure 6.13: RKDG scheme: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

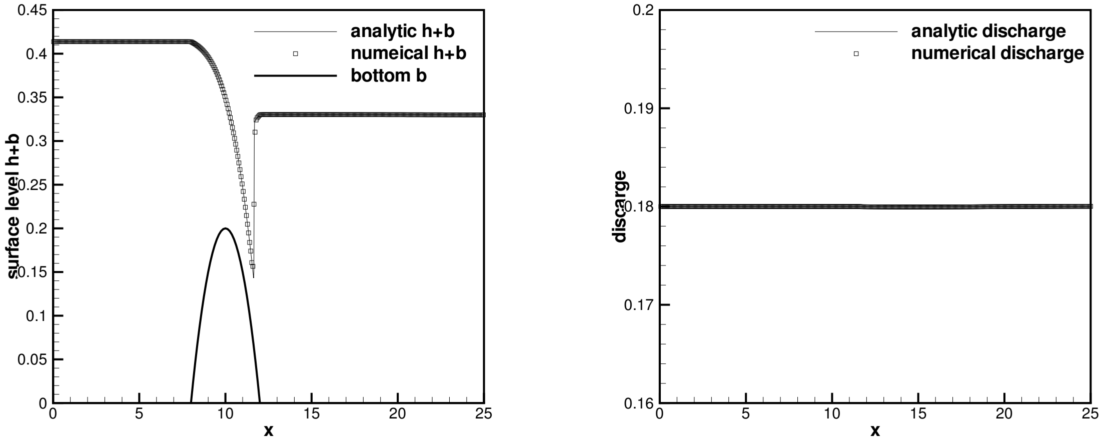


Figure 6.14: FV scheme with 400 uniform cells: Steady transcritical flow over a bump with a shock. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

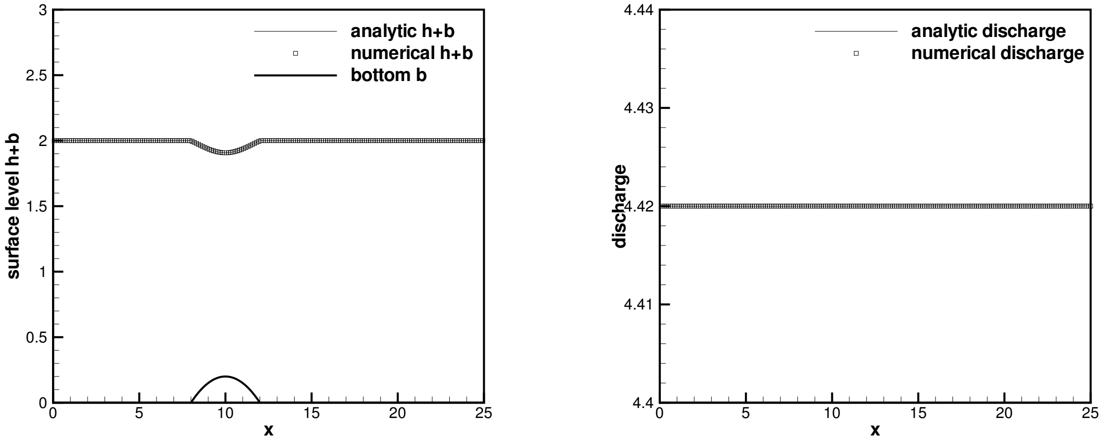


Figure 6.15: FV scheme: Steady subcritical flow over a bump. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

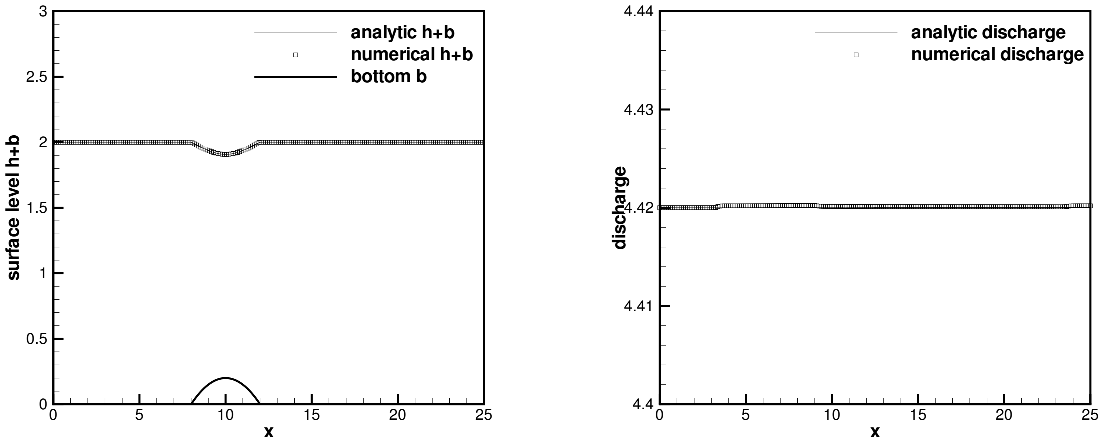


Figure 6.16: RKDG scheme: Steady subcritical flow over a bump. Left: the surface level $h + b$; right: the discharge hu as the numerical flux for the water height h.

is chosen to be the bottom. $h(x, y, 0) = 1 - b(x, y)$ is the initial depth of the water. Initial velocity is set to be zero. This surface should remain flat. The computation is performed to $t = 0.1$ using single, double and quadruple precisions with a 100×100 uniform mesh. Table 6.4 contains the L^1 errors for the water height h (which is not a constant function) and the discharges hu and hv for both schemes. We can clearly see that the L^1 errors are at the level of round-off errors for different precisions, verifying the exact C-property.

Table 6.4: L^1 errors for different precisions for the stationary solution in Section 6.3.6.

		L^1 error		
	precision	h	hu	hv
FV	single	8.77E-07	7.49E-07	6.93E-07
	double	1.49E-15	2.31E-15	2.30E-15
	quadruple	1.04E-33	9.87E-34	1.01E-33
RKDG	single	9.43E-08	4.84E-07	4.94E-07
	double	6.98E-17	2.31E-15	2.31E-15
	quadruple	6.14E-34	1.52E-33	1.53E-33

6.3.7 Testing the orders of accuracy

In this example we check the numerical orders of accuracy when the schemes are applied to the following two dimensional problem. The bottom topography and the initial data are given by:

$$b(x,y) = \sin(2\pi x) + \cos(2\pi y), \qquad h(x,y,0) = 10 + e^{\sin(2\pi x)}\cos(2\pi y),$$

$$(hu)(x,y,0) = \sin(\cos(2\pi x))\sin(2\pi y), \qquad (hv)(x,y,0) = \cos(2\pi x)\cos(\sin(2\pi y))$$

defined over a unit square, with periodic boundary conditions. The terminal time is taken as t=0.05 to avoid the appearance of shocks in the solution. Since the exact solution is also not known explicitly for this case, we use the same fifth order WENO scheme with an extremely refined mesh consisting of 1600×1600 cells to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. The TVB constant M in the limiter for the RKDG scheme is taken as 40 here. Tables 6.5 and 6.6 contain the L^1 errors and orders of accuracy for the cell averages. We can clearly see that, in this two dimensional test case, fifth order accuracy is achieved for the finite volume WENO scheme and close to third order accuracy is achieved for the RKDG scheme.

Table 6.5: FV scheme: L^1 errors and numerical orders of accuracy for the example in Section 6.3.7.

Number		h		hu		hv	
of cells	CFL	L^1 error	order	L^1 error	order	L^1 error	order
25×25	0.6	7.91E-03		2.12E-02		6.52E-02	
50×50	0.6	1.13E-03	2.81	2.01E-03	3.40	9.22E-03	2.82
100×100	0.6	8.89E-05	3.66	1.25E-04	4.00	7.19E-04	3.68
200×200	0.4	4.07E-06	4.45	5.19E-06	4.59	3.30E-05	4.45
400×400	0.3	1.42E-07	4.84	1.84E-07	4.82	1.15E-06	4.84
800×800	0.2	4.38E-09	5.02	5.99E-09	4.94	3.63E-08	4.99

Table 6.6: RKDG scheme: L^1 errors and numerical orders of accuracy for the example in Section 6.3.7.

Number of cells	h		hu		hv	
	L^1 error	order	L^1 error	order	L^1 error	order
25 × 25	2.45E-03		1.36E-02		2.05E-02	
50 × 50	5.73E-04	2.10	2.92E-03	2.22	4.75E-03	2.11
100 × 100	1.06E-04	2.43	5.31E-04	2.46	8.51E-04	2.48
200 × 200	1.71E-05	2.63	8.81E-05	2.60	1.39E-04	2.61
400 × 400	2.53E-06	2.75	1.32E-05	2.74	2.11E-05	2.72
800 × 800	3.52E-07	2.84	1.88E-06	2.81	3.01E-06	2.81

6.3.8 A small perturbation of a two dimensional steady-state water

This is a classical example to show the capability of the proposed scheme for the perturbation of the stationary state, given by LeVeque [28]. It is analogous to the test done previously in Section 6.3.3 in one dimension.

We solve the system in the rectangular domain $[0,2] \times [0,1]$. The bottom topography is an isolated elliptical shaped hump:

$$b(x,y) = 0.8\, e^{-5(x-0.9)^2-50(y-0.5)^2}. \tag{6.28}$$

The surface is initially given by:

$$\begin{aligned} h(x,y,0) &= \begin{cases} 1-b(x,y)+0.01 & \text{if } 0.05 \le x \le 0.15 \\ 1-b(x,y) & \text{otherwise} \end{cases} \\ hu(x,y,0) &= hv(x,y,0) = 0 \end{aligned} \tag{6.29}$$

So the surface is almost flat except for $0.05 \le x \le 0.15$, where h is perturbed upward by 0.01. Figures 6.17 and 6.18 display the right-going disturbance as it propagates past the hump, on two different uniform meshes with 200 × 100 cells and 600 × 300 cells for comparison. The surface level $h + b$ is presented at different times. The results indicate that both schemes can resolve the complex small features of the flow very well.

6.4 Other applications

In this section, we generalize high order well balanced schemes, designed in Sections 6.1 and 6.2, to other balance laws introduced in [47], including the elastic wave equation, the hyperbolic model for a chemosensitive movement, the nozzle flow, a model of fluid mechanics and a two phase flow model. Due to page limitation, only the elastic wave equation and chemosensitive movement model are investigated here, however our technique can also be applied to the other three cases. Some selective numerical tests are presented to show the good properties of our well balance schemes.

6.4.1 Elastic wave equation

We consider the propagation of compressional waves [3, 45] in an one-dimensional elastic rod with a given media density $\rho(x)$. The equations of motion in a Lagrangian frame are given by the balance laws:

$$\begin{cases} (\rho\varepsilon)_t + (-\rho u)_x = -u\dfrac{d\rho}{dx} \\ (\rho u)_t + (-\sigma)_x = 0, \end{cases} \tag{6.30}$$

where ε is the strain, u is the velocity and σ is a given stress-strain relationship $\sigma(\varepsilon, x)$. The equation of linear acoustics can be obtained from above if the stress-strain relationship is linear,

$$\sigma(\varepsilon, x) = K(x)\,\varepsilon$$

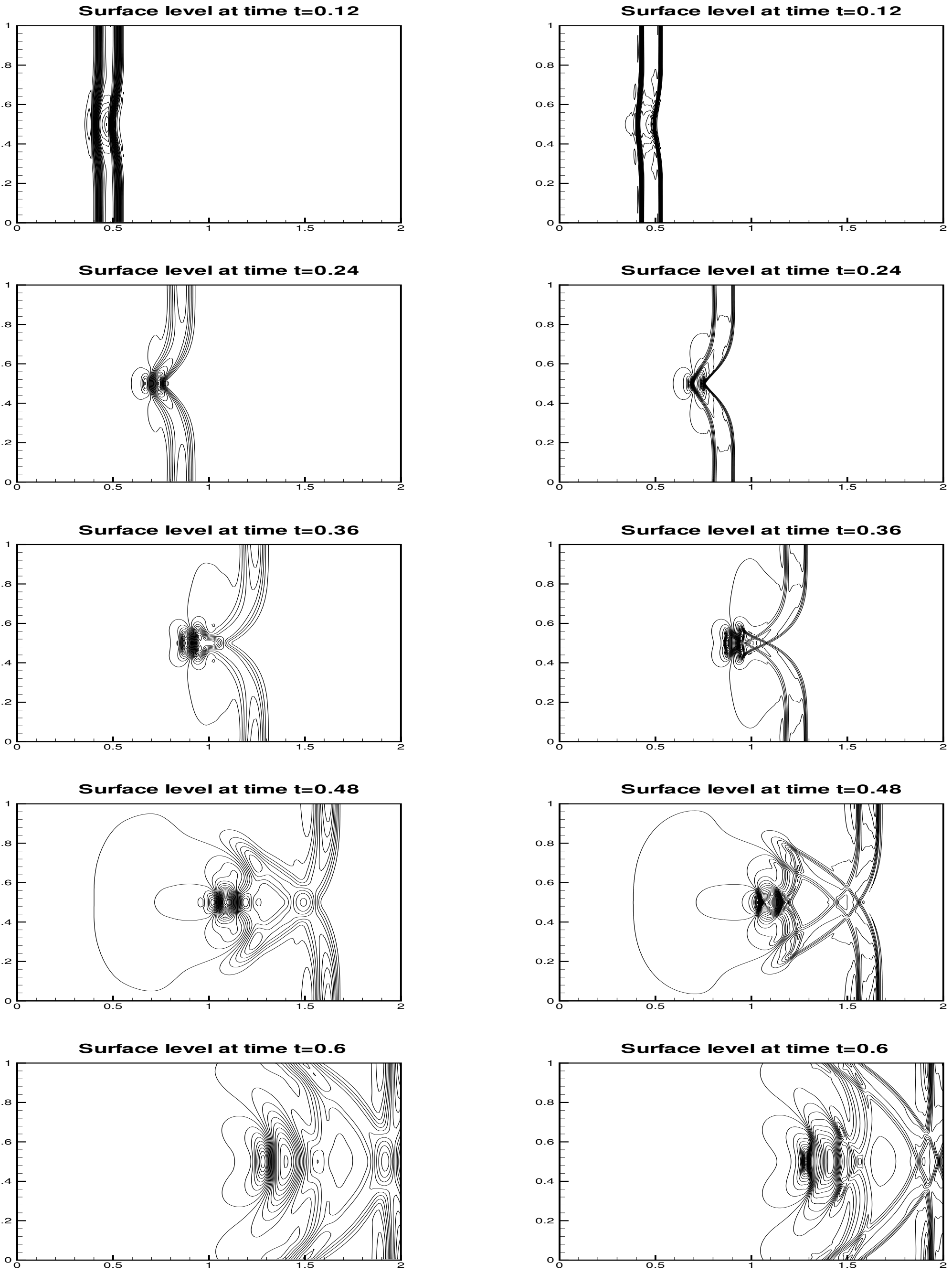
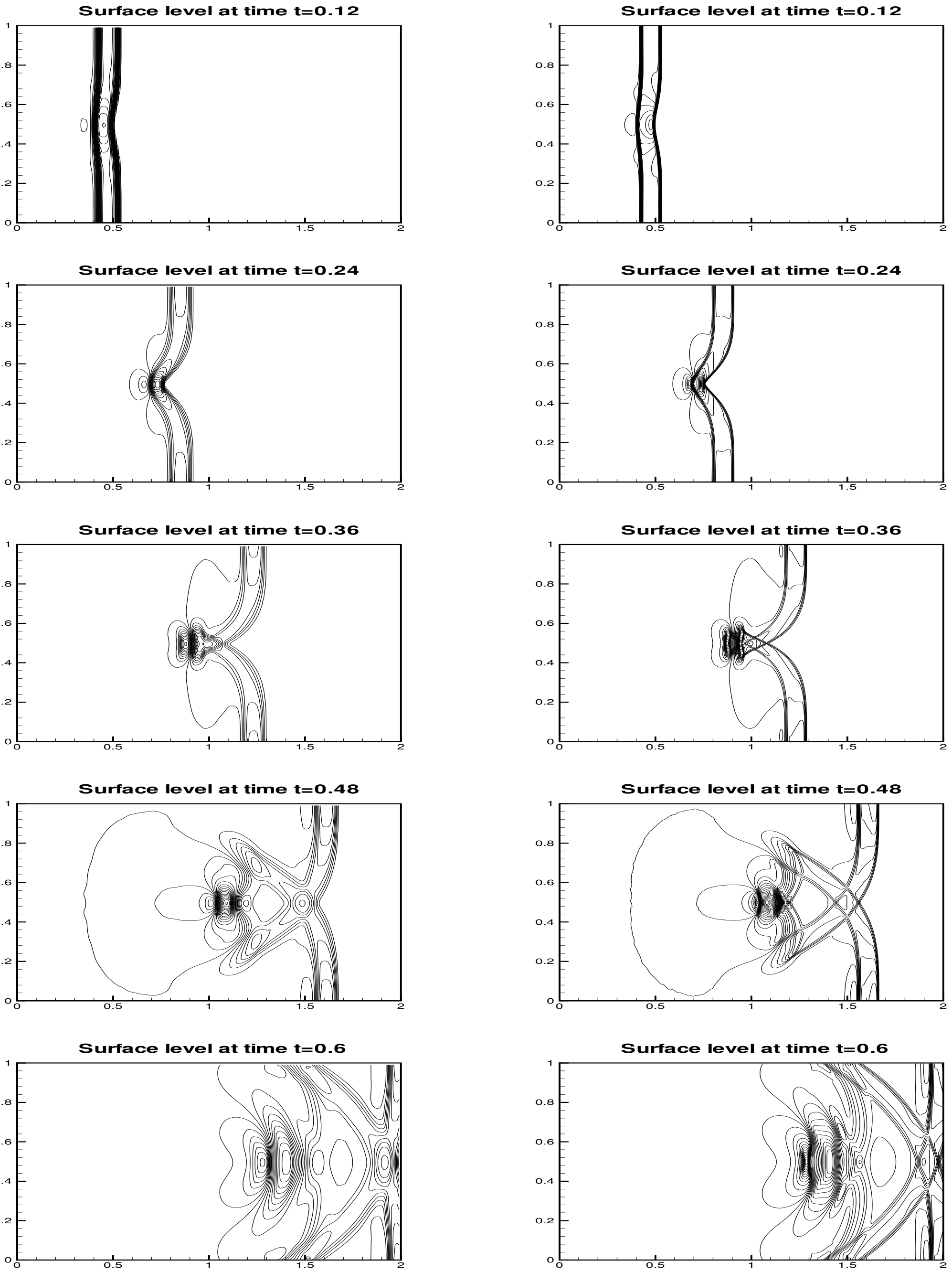


Figure 6.17: FV scheme: The contours of the surface level $h + b$ for the problem in Section 6.3.8. 30 uniformly spaced contour lines. From top to bottom: at time $t = 0.12$ from 0.99942 to 1.00656; at time $t = 0.24$ from 0.99318 to 1.01659; at time $t = 0.36$ from 0.98814 to 1.01161; at time $t = 0.48$ from 0.99023 to 1.00508; and at time $t = 0.6$ from 0.99514 to 1.00629. Left: results with a 200×100 uniform mesh. Right: results with a 600×300 uniform mesh.



Surface level at time t=0.12

Surface level at time t=0.12

Surface level at time t=0.24

Surface level at time t=0.24

Surface level at time t=0.36

Surface level at time t=0.36

Surface level at time t=0.48

Surface level at time t=0.48

Surface level at time t=0.6

Surface level at time t=0.6

Figure 6.18: RKDG scheme: The contours of the surface level $h + b$ for the problem in Section 6.3.8. 30 uniformly spaced contour lines. From top to bottom: at time $t = 0.12$ from 0.99942 to 1.00656; at time $t = 0.24$ from 0.99318 to 1.01659; at time $t = 0.36$ from 0.98814 to 1.01161; at time $t = 0.48$ from 0.99023 to 1.00508; and at time $t = 0.6$ from 0.99514 to 1.00629. Left: results with a 200×100 uniform mesh. Right: results with a 600×300 uniform mesh.

where $K(x)$ is the given bulk modulus of compressibility. The steady state we are interested to preserve for this problem is characterized by

$$a_1 \equiv \sigma(\varepsilon, x) = constant, \qquad a_2 \equiv u = constant.$$

Here we only show the well balanced property for the RKDG schemes. Similar idea can be used for the finite volume WENO schemes.

First, we project the initial value to obtain $U_h = ((\rho\varepsilon)_h, (\rho u)_h)^T$, and also apply the same procedure for ρ to obtain ρ_h. Then, we check the three conditions in Section 6.1 one by one. Only the first equation in (6.30) must be considered for the well balanced property.

1: If the steady state is reached, $u_h \equiv \frac{(\rho u)_h}{\rho_h}$ is constant and ρ_h is a polynomial, hence the integral of the source term can be calculated exactly.

2: We set

$$(\rho u)^{*,+}_{h,j+\frac{1}{2}} = \frac{(\rho u)^{+}_{h,j+\frac{1}{2}}}{\rho^{+}_{h,j+\frac{1}{2}}} \max(\rho^{+}_{h,j+\frac{1}{2}}, \rho^{-}_{h,j+\frac{1}{2}}) \tag{6.31}$$

$$(\rho u)^{*,-}_{h,j+\frac{1}{2}} = \frac{(\rho u)^{-}_{h,j+\frac{1}{2}}}{\rho^{-}_{h,j+\frac{1}{2}}} \max(\rho^{+}_{h,j+\frac{1}{2}}, \rho^{-}_{h,j+\frac{1}{2}}) \tag{6.32}$$

and redefine the left and right values of U as:

$$U^{*,\pm}_{h,j+\frac{1}{2}} = \begin{pmatrix} (\rho\varepsilon)^{\pm}_{h,j+\frac{1}{2}} \\ (\rho u)^{*,\pm}_{h,j+\frac{1}{2}} \end{pmatrix} \tag{6.33}$$

Then we define the left and right fluxes as:

$$\hat{f}^{l}_{j+\frac{1}{2}} = F(U^{*,-}_{h,j+\frac{1}{2}}, U^{*,+}_{h,j+\frac{1}{2}}) + \begin{pmatrix} -(\rho u)^{-}_{h,j+\frac{1}{2}} + (\rho u)^{*,-}_{h,j+\frac{1}{2}} \\ 0 \end{pmatrix} \tag{6.34}$$

$$\hat{f}^r_{j-\frac{1}{2}} = F(U^{*,-}_{h,j-\frac{1}{2}}, U^{*,+}_{h,j-\frac{1}{2}}) + \begin{pmatrix} -(\rho u)^+_{h,j-\frac{1}{2}} + (\rho u)^{*,+}_{h,j-\frac{1}{2}} \\ 0 \end{pmatrix}. \qquad (6.35)$$

The max in (6.31)-(6.32) was chosen in [2] to guarantee positive water height and was referred to as "hydrostatic reconstruction" there. Here it does not have a clear physical meaning and could be replaced by minimum or average as well.

3: $(\rho u)_h$, satisfying $u_h = constant$, is also a steady state solution of :

$$(-\rho u)_x = -u\frac{d\rho_h}{dx}.$$

With these three conditions, we can repeat the proof of Proposition 6.1.1 to show that our schemes are indeed well balanced and high order accurate.

Remark 6.4.1 *When performing the limiting on the function $(\rho u)_h$ after each Runge-Kutta stage to control spurious oscillations, we keep in mind that our purpose is to maintain the steady state solution $(\rho u)_h$ which satisfies $u_h = constant$. Here we follow the idea used in [48], and first check whether any limiting is needed based on the function u_h in each Runge-Kutta stage. If the answer is yes, then the actual limiter is implemented on $(\rho u)_h$.*

Next, we present the numerical result for a linear acoustic test [3]. The properties of the media are given by

$$c(x) = \sqrt{\frac{K(x)}{\rho(x)}} = 1 + 0.5\sin(10\pi x), \qquad Z(x) = \rho(x)c(x) = 1 + 0.25\cos(10\pi x).$$

The initial conditions are given by

$$\rho\varepsilon(x,0) = \begin{cases} \dfrac{-1.75 + 0.75\cos(10\pi x)}{c^2(x)}, & \text{if } 0.4 < x < 0.6 \\ \dfrac{-1}{c^2(x)}, & \text{otherwise} \end{cases}, \qquad u(x,0) = 0.$$

It is a test case where the impedance $Z(x)$ and hence the eigenvectors are both

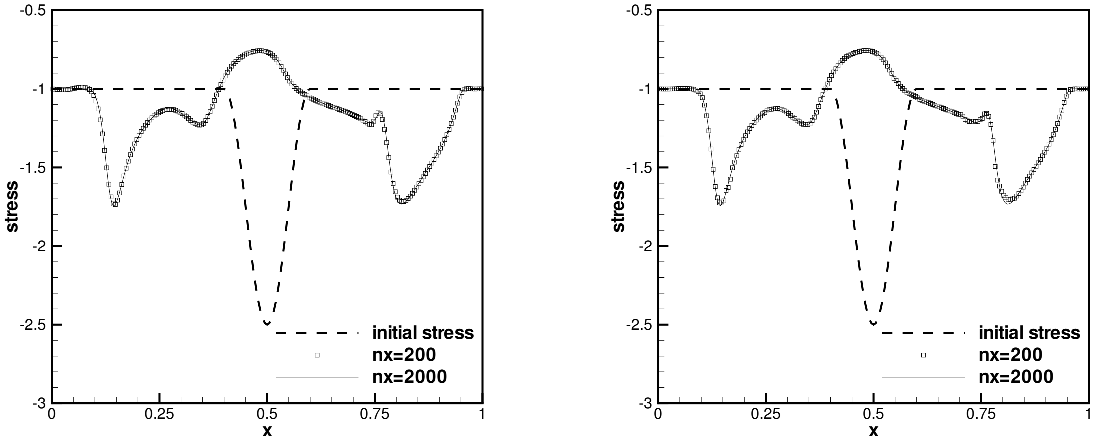


Figure 6.19: The numerical (symbols) and the "exact" reference (solid line) stress $\sigma(x)$ at time $t = 0.4s$. Left: FV schemes; right: DG schemes.

spatially varying. We perform the computation with 200 uniform cells, with the ending time $t = 0.4s$. An "exact" reference solution is computed with the same scheme over a 2000 grid point uniform cells. The simulation results are shown in Figure 6.19. The numerical resolution shows very good agreement with the "exact" reference solution.

6.4.2 Chemosensitive movement

Originated from biology, chemosensitive movement [21, 15] is a process by which cells change their direction reacting to the presence of a chemical substance, approaching chemically favorable environments and avoiding unfavorable ones. Hyperbolic models for chemotaxis are recently introduced [21] and take the form

$$\left\{ \begin{array}{l} n_t + (nu)_x = 0 \\ (nu)_t + (nu^2 + n)_x = n\chi'(c)\dfrac{\partial c}{\partial x} - \sigma nu \end{array} \right. \tag{6.36}$$

where the chemical concentration $c = c(x,t)$ is given by the parabolic equation

$$\frac{\partial c}{\partial t} - D_c \triangle c = n - c.$$

Here, $n(x,t)$ is the cell density, $nu(x,t)$ is the population flux and σ is the friction coefficient.

We would like to preserve the steady state solution to (6.36) with a zero population flux, which satisfies

$$\frac{n}{e^{\chi(c)}} = constant, \qquad nu = 0. \tag{6.37}$$

where $c = c(x)$ does not depend on t in steady state.

Here we only show the well balanced property for the RKDG schemes. Similar idea can be used for the finite volume WENO schemes.

First, we project the initial value to obtain $U_h = (n_h, (nu)_h)^T$, and also project $e^{\chi(c)}$ to obtain $(e^{\chi(c)})_h$. Then, we check the three conditions in Section 6.1 one by one. Only the second equation in (6.36) is relevant for the well balanced property.

1: The source term can be written as: $\frac{n}{e^{\chi(c)}} \frac{d}{dx} e^{\chi(c)} - \sigma nu$. If the steady state is reached, $\frac{n}{e^{\chi(c)}}$ is constant and $(e^{\chi(c)})_h$ is a polynomial, hence the integral of the source term can be calculated exactly.

2: We set

$$n^{*,+}_{h,j+\frac{1}{2}} = \frac{n^{+}_{h,j+\frac{1}{2}}}{(e^{\chi(c)})^{+}_{h,j+\frac{1}{2}}} \max((e^{\chi(c)})^{+}_{h,j+\frac{1}{2}}, (e^{\chi(c)})^{-}_{h,j+\frac{1}{2}}) \tag{6.38}$$

$$n^{*,-}_{h,j+\frac{1}{2}} = \frac{n^{-}_{h,j+\frac{1}{2}}}{(e^{\chi(c)})^{-}_{h,j+\frac{1}{2}}} \max((e^{\chi(c)})^{+}_{h,j+\frac{1}{2}}, (e^{\chi(c)})^{-}_{h,j+\frac{1}{2}}) \tag{6.39}$$

and redefine the left and right values of U as:

$$U^{*,\pm}_{h,j+\frac{1}{2}} = \left(\begin{array}{c} n^{*,\pm}_{h,j+\frac{1}{2}} \\ (nu)^{\pm}_{h,j+\frac{1}{2}} \end{array} \right) \tag{6.40}$$

Then we define the left and right fluxes as:

$$\hat{f}^{l}_{j+\frac{1}{2}} = F(U^{*,-}_{h,j+\frac{1}{2}}, U^{*,+}_{h,j+\frac{1}{2}}) + \left(\begin{array}{c} 0 \\ n^{-}_{h,j+\frac{1}{2}} - n^{*,-}_{h,j+\frac{1}{2}} \end{array} \right) \tag{6.41}$$

$$\hat{f}^{r}_{j-\frac{1}{2}} = F(U^{*,-}_{h,j-\frac{1}{2}}, U^{*,+}_{h,j-\frac{1}{2}}) + \left(\begin{array}{c} 0 \\ n^{+}_{h,j-\frac{1}{2}} - n^{*,+}_{h,j-\frac{1}{2}} \end{array} \right) \tag{6.42}$$

3: We note that $(nu)_h = 0$ and n_h, satisfying $\frac{n_h}{(e^{\chi(c)}_h} = constant$, is the steady state solution of :

$$(nu^2 + n)_x = \frac{n}{(e^{\chi(c)})_h} \frac{d}{dx}(e^{\chi(c)})_h - \sigma nu.$$

With these three conditions, we can repeat the proof of Proposition 6.1.1 to show that our new schemes are indeed well balanced and high order accurate.

The limiter procedure is performed similarly as in Section 6.4.1. We refer to [48] for more details.

The following example is to test the fifth order accuracy for smooth solutions, for which we take the initial conditions as

$$n(x,0) = 1 + 0.2\,\cos(\pi x), \qquad u(x,0) = 0, \qquad x \in [-1,1]$$

with

$$c(x) = e^{-16x^2}, \qquad \chi(c) = \log(1+c), \qquad \sigma = 0$$

with a periodic boundary condition. Since the exact solution is not known explicitly for this problem, we use the same fifth order WENO scheme with $N = 5120$ points to compute a reference solution and treat it as the exact solution when computing the numerical errors. Final time $t = 1.0s$ is used to avoid the development of shocks. The constant M is taken as 13 and the CFL number is 0.18 in the RKDG code. Table 6.7 contains the L^1 errors and numerical orders of accuracy. We can clearly see that the expected order accuracy is achieved for this example.

Table 6.7: L^1 errors and numerical orders of accuracy for the example in Section 6.4.2.

No. of points	FV schemes				
	CFL	$\rho\epsilon$		ρu	
		L^1 error	order	L^1 error	order
20	0.6	1.10E-002		8.76E-003	
40	0.6	1.20E-003	3.20	1.02E-003	3.10
80	0.5	1.19E-004	3.34	9.81E-005	3.38
160	0.4	6.27E-006	4.25	5.32E-006	4.20
320	0.3	2.48E-007	4.66	2.12E-007	4.65
640	0.1	8.09E-009	4.95	6.85E-008	4.96
No. of points	DG schemes				
		$\rho\epsilon$		ρu	
		L^1 error	order	L^1 error	order
20		1.13E-004		1.13E-004	
40		1.56E-005	2.86	1.42E-005	2.99
80		1.06E-006	3.88	9.58E-007	3.89
160		8.91E-008	3.57	8.09E-008	3.56
320		8.93E-009	3.32	8.10E-009	3.32
640		1.06E-009	3.07	9.59E-0010	3.08

Chapter 7

High Order Finite Volume Well-balanced WENO Schemes for the Moving Steady State of the Shallow Water Equations

In this chapter, we are interested in exactly preserving the moving steady state solution

$$hu = constant, \qquad \frac{1}{2}u^2 + g(h+b) = constant \tag{7.1}$$

of the shallow water equations (1.3), which has many applications in the world. This steady state problem does not belong in the class of balance laws introduced in Chapter Four, nor can it be balanced by the method used in Chapter Six. A new well balanced finite volume WENO scheme is presented here. Only one dimensional problem is discussed in this thesis. In Section 7.1, we introduce some notations to be used later. The algorithm to construct well balanced schemes is then introduced in Section 7.2. Section 7.3 contains extensive numerical simulation results to demonstrate the behavior of our well balanced WENO schemes, verifying high order accuracy, the well balanced property, and good resolution for smooth and discontinuous solutions.

7.1 Conservative and equilibrium variables

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

$$E := \frac{1}{2}u^2 + g(h + b) \tag{7.2}$$

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

$$h_t + m_x = 0 \tag{7.3}$$

$$u_t + E_x = 0. \tag{7.4}$$

Thus the steady states (4.5) are given by $m \equiv \mathit{constant}$, $E \equiv \mathit{constant}$. This motivates the introduction of the *equilibrium variables*

$$V := (m, E). \tag{7.5}$$

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

7.1.1 Variable transformations

Given conservative variables U and a bottom function b, the energy E (and hence the equilibrium variables $V = V(U)$) can be easily computed by (7.5). The difficulty lies in finding the inverse transform $U = U(V)$. For this, we introduce the Froude number

$$Fr := |u|/\sqrt{gh}, \tag{7.6}$$

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

$$\sigma := \mathrm{sign}(Fr - 1), \tag{7.7}$$

so

$$\sigma = \begin{cases} 1 & \text{supersonic flow} \\ 0 & \text{sonic flow} \\ -1 & \text{subsonic flow.} \end{cases} \tag{7.8}$$

Suppose now that $V = (m, E)$ and b are given. Under which conditions can we recover the conservative variable h from this information, and thus establish the desired transform $U = U(V)$? Let us denote the part of the energy depending on h by

$$\varphi(h) := \frac{m^2}{2h^2} + gh. \tag{7.9}$$

Here m is considered to be a fixed parameter. Our task is to find a unique solution h such that

$$\varphi(h) = E - gb. \tag{7.10}$$

If $m = 0$, then one can solve (7.10) as long as $E - gb > 0$. If $m \neq 0$, then $\varphi(h)$ is positive and convex. Its unique minimum is (h_0, φ_0) with

$$gh_0 = (g|m|)^{2/3}, \quad \varphi_0 = \frac{3}{2}(g|m|)^{2/3}. \tag{7.11}$$

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

$$E_0 = \varphi_0 + gb = \frac{3}{2}(g|m|)^{2/3} + gb. \tag{7.12}$$

If $E < E_0$, there is no solution to (7.10). If $E = E_0$, there is the unique solution $h = h_0$. If $E > E_0$, there are two solutions, one super- and the other one subsonic.

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

$$\hat{\varphi}(\hat{h}) = \frac{2}{3}\left(\frac{1}{2\hat{h}^2} + \hat{h}\right), \tag{7.13}$$

and the Froude number may be written as

$$Fr(\hat{h}) = \hat{h}^{-3/2}. \tag{7.14}$$

This shows that $\hat{h} = 1$, $\hat{h} > 1$ resp. $\hat{h} < 1$ correspond to sonic, sub- and supersonic states, see Figure 7.1. If we introduce $\hat{E} := (E - gb)/\varphi_0$, then (7.10) becomes

$$\hat{\varphi}(\hat{h}) = \hat{E}. \tag{7.15}$$

We summarize our results in the following Definition and Lemma.

Definition 7.1.1 *Let $m \in \mathbb{R}$ be given. A pair $(\hat{E}, \sigma) \in \mathbb{R} \times \{-1, 0, 1\}$ (resp. a triple $(E, b, \sigma) \in \mathbb{R}^2 \times \{-1, 0, 1\}$) is an admissible state if either*

$$\sigma = 0 \quad \text{and} \quad \hat{E} = 1 \quad (\text{resp. } E = E_0) \tag{7.16}$$

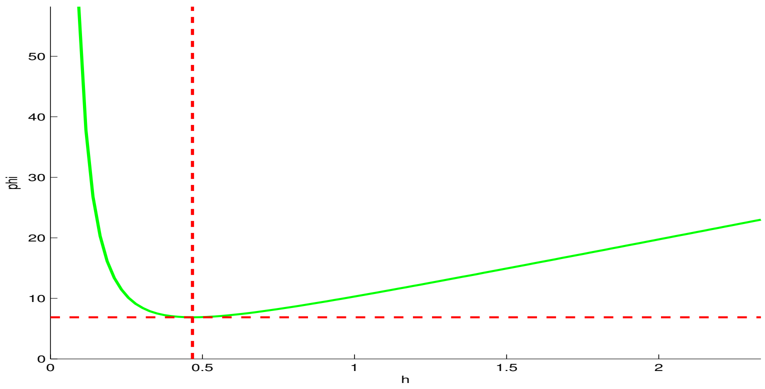


Figure 7.1: The normalized function $\hat{\varphi}(\hat{h})$. Supersonic ($\hat{h} < 1$), sonic ($\hat{h} = 1$) and subsonic ($\hat{h} > 1$) regions.

or

$$|\sigma| = 1 \quad \text{and} \quad \hat{E} > 1 \quad (\text{resp. } E > E_0). \tag{7.17}$$

Lemma 7.1.2 *Let m be given, and suppose that the pair $(\hat{E}, \sigma)$ is admissible. Then there exists a unique solution*

$$\hat{h} = \hat{h}(\hat{E}, \sigma) \tag{7.18}$$

such that

$$\begin{array}{lll} \hat{h} < 1 & \text{for } \sigma = 1 & \text{(supersonic flow)} \\ \hat{h} = 1 & \text{for } \sigma = 0 & \text{(sonic flow)} \\ \hat{h} > 1 & \text{for } \sigma = -1 & \text{(subsonic flow).} \end{array} \tag{7.19}$$

We call $\hat{h}(\hat{E}, \sigma)$ the admissible solution of (7.15).

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

Corollary 7.1.3 *Let m be given, and suppose that the triple (E, b, σ) is admissible.*

Then the unique admissible solution $h = h(m, E, b, \sigma)$ *of* (7.10) *is given by*

$$h(m, E, b, \sigma) = \frac{1}{g}(gm)^{2/3}\hat{h}(\hat{E}, \sigma). \tag{7.20}$$

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

7.2 High order well-balanced finite volume scheme

In this section, we design a high order finite volume WENO scheme for the shallow water equation (1.3), with the objective to maintain the general moving steady state solution (7.1). We start with the one dimensional case, and left the two dimensional problem for further investigation. The basic framework of the well balanced scheme follows the one introduced by Audusse et al. [2], and later used in the recent papers [30, 49]. However, the approximation of the flux and source terms requires more attention due to the complexity of the moving steady state. For simplicity, we denote the shallow water equations (1.3) by

$$U_t + f(U)_x = s(U, b), \tag{7.21}$$

where U represents (h, hu), $f(U)$ is the flux and $s(U, b)$ stands for the source term.

7.2.1 Framework of the discretization

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

We solve an integrated version of (7.21) over the interval I_i, Our conservative finite volume scheme takes the classical semidiscrete form

$$\frac{d}{dt}\overline{U}_i(t) = -\frac{1}{\Delta x_i}\left(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}\right) + \frac{1}{\Delta x_i}s_i =: \frac{1}{\Delta x_i}r_i. \tag{7.22}$$

where $\hat{f}_{i+\frac{1}{2}}$ is a consistent, Lipschitz continuous numerical flux for the homogeneous shallow water equations and s_i is a high order approximation to the integral of the source term $\int_{I_i} s(h(x,t), b(x))dx$. For later reference, we call the RHS of (7.22) the *residual* $r_i/\Delta x_i$. Thus a well-balanced scheme is one for which all residuals vanish at steady state.

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

Lemma 7.2.1 *The numerical scheme* (7.22) *is formally k-th order accurate if the following holds in smooth regions:*

i) $\hat{f}_{i+\frac{1}{2}} = f(U(x_{i+\frac{1}{2}}, t)) + O((\triangle x_i)^k)$, *where the O term is smooth.*

ii) $s_i = \int_{I_i} s(h,b)dx + O((\Delta x_i)^{k+1})$

The following Lax-Wendroff theorem proves that the numerical solution converges to a weak solution of this conservation law.

Lemma 7.2.2 *The numerical approximation computed with scheme* (7.22) *converges to a weak solution of* (7.21)*, if the following conditions are satisfied:*

i) the numerical flux $\hat{f}_{i+\frac{1}{2}}$ *is a consistent flux* (2.7)*;*

ii) the approximation to the source term $\frac{1}{\Delta x_i}s_i$ *stay bounded when the numerical solution itself is bounded, as the grid size* Δx *is refined.*

The proof of these lemmas are straightforward.

We choose a TVD Runge-Kutta discretization [41] in time. In order to complete the definition of the scheme, we need to introduce the spatial reconstruction, the source term discretization, and the numerical fluxes. This will be done in Sections 7.2.2 and 7.2.3.

7.2.2 Equilibrium-limited reconstructions in the cell interior

Assume the initial values $\bar{U}_i$ and $\bar{b}_i$ are given. We apply the WENO reconstruction procedure on $\bar{b}_i$ to obtain $b_i, b^{\pm}_{i+\frac{1}{2}}$, and the approximations of $b(x)$ at the relevant Gaussian points. If $b(x)$ is known at all points, this WENO reconstruction procedure is unnecessary.

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

Now we need to address one of the more subtle points of the well-balanced algorithm. Even if the initial data are in perfect equilibrium, say $V(x) \equiv \bar{V}$ for some constant equilibrium state $\bar{V}$, the WENO-reconstructed values $U_i, U^{\pm}_{i+\frac{1}{2}}$ and hence $V_i, V^{\pm}_{i+\frac{1}{2}}$ may not be in equilibrium any more. The problem is the total energy $E = \frac{1}{2}u^2 + g(h+b)$. First of all, the topography b may be a general function of x. Second, the velocity depends nonlinearly on height and momentum. For the lake at rest, the second problem disappears since $u = 0$. The first problem can be fixed by reconstructing not b, but $h+b$ and recovering $b_i, b^{\pm}_{i+\frac{1}{2}}$ as $(h+b)_i - h_i$, $(h+b)^{\pm}_{i+\frac{1}{2}} - h^{\pm}_{i+\frac{1}{2}}$, see [2, 30].

For moving equilibria, this is much less straightforward. Our solution proceeds as follows: first we define a local reference state $\bar{V}_i$ for each cell I_i as the solution of the implicit equation

$$\bar{U}_i = \frac{1}{\triangle x_i} \int_{I_i} U(\bar{V}_i, b(x), \sigma(\bar{U}_i))\, dx, \tag{7.23}$$

where the conservative cell average $\bar{U}_i$ is considered to be given. By definition, these reference values satisfy

Lemma 7.2.3 *Suppose the function $U(x)$ is in equilibrium, i.e. there is an equilibrium state $\bar{V}$ such that*

$$V(U(x), b(x)) \equiv \bar{V} \qquad \mathit{in}\ \mathcal{D}.$$

Then

$$\bar{V}_i = \bar{V} \quad \text{for all} \ \ i. \tag{7.24}$$

In actual implementation, we use a Gauss quadrature of sufficient accuracy to approximate the integral in (7.23). That is, the reference energy $\bar{E}_i$ is implicitly defined by the equation

$$\bar{h}_i = \frac{1}{\Delta x_i} \sum_\alpha \omega_\alpha h(\overline{hu}_i, \bar{E}_i, b_{i+\alpha}, \sigma(\bar{U}_i)). \tag{7.25}$$

A Newton iteration is then used to solve (7.25) with the initial guess of $\bar{E}_i$ being

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

The conclusion of Lemma 7.2.3 still holds for the reference value $\bar{E}_i$ defined in (7.25). The relevance of Lemma 7.2.3 (and its discrete analogue) is that it provides an indicator that we have reached equilibrium, since in this case all the values $\bar{V}_i$ coincide.

Next we show how to use the local reference values $\bar{V}_i$ to modify the WENO reconstructed values $V^{\pm}_{i+\frac{1}{2}}$ and V_i in such a way that they maintain any present global equilibrium state $\bar{V}$. For this we use the TVD type limiter function

$$\lim(w; \bar{w}_i, \bar{w}_{i\pm1}) := \bar{w}_i + m(w - \bar{w}_i, \bar{w}_{i+1} - \bar{w}_i, \bar{w}_i - \bar{w}_{i-1}), \tag{7.26}$$

where

$$m(a_1, a_2, a_3) = \begin{cases} s \min_{1\leq n\leq 3}|a_n| & \text{if } s = sign(a_1) = sign(a_2) = sign(a_3), \\ 0, & \text{otherwise.} \end{cases} \tag{7.27}$$

Of course, other limiters should be possible as well.

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

$$\breve{V}^{\pm}_{i+\frac{1}{2}} = \lim(V^{\pm}_{i+\frac{1}{2}}; \bar{V}_i, \bar{V}_{i\pm1}). \tag{7.28}$$

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

Lemma 7.2.4 *At steady state, where $V(x) \equiv \bar{V}$, the limited values* (7.28) *satisfy*

$$\breve{V}^{\pm}_{i+\frac{1}{2}} = \breve{V}_i = \bar{V}_i = \bar{V} \quad \text{for all } i. \tag{7.29}$$

Therefore we call (7.26)–(7.28) *the* equilibrium limiter.

Proof. If $V(x) \equiv \bar{V}$, then $\bar{V}_i = \bar{V}$ for all i due to (7.24). Therefore, the function m in (7.26)–(7.27) vanishes, and

$$\lim(V^{\pm}_{i+\frac{1}{2}}; \bar{V}_i, \bar{V}_{i\pm1}) = \bar{V}_i = \bar{V}. \tag{7.30}$$

□

We still have to modify our reconstructed values $\breve{V}^{\pm}_{i+\frac{1}{2}}$ and $\breve{V}_i$ once more due to the following observation:

Remark 7.2.5 *We may only have first order accuracy at smooth extrema if the above TVD limiter is used.*

To improve the accuracy, we have two choices.

1: We have the pointwise values $V^{\pm}_{i+\frac{1}{2}}$ which are high order accurate for general solutions, but do not equal to a constant for the moving steady state solution. On the other hand, we have the limited values $\check{V}^{\pm}_{i+\frac{1}{2}}$ which satisfy the well balanced requirement for steady states, but are only first order accurate at smooth extrema for general solutions. The idea is to use a convex combination of the two candidate values

$$\tilde{V}^{\pm}_{i+\frac{1}{2}} := V^{\pm}_{i+\frac{1}{2}} + \text{iflag}\,(\check{V}^{\pm}_{i+\frac{1}{2}} - V^{\pm}_{i+\frac{1}{2}}). \tag{7.31}$$

The goal is to weight the combined value $\tilde{V}^{\pm}_{i+\frac{1}{2}}$ toward $\check{V}^{\pm}_{i+\frac{1}{2}}$ for steady state soutions, and toward $V^{\pm}_{i+\frac{1}{2}}$ away from them. One possible definition of the indicator 'iflag' is based on a wide support. The idea we follow is: We find $2\sqrt{N}$ (N is the number of cells) consequent points on two sides of cell I_i, and sum $\min(\bar{V}_{j+1} - \bar{V}_j, \bar{V}_j - \bar{V}_{j-1})$ over any j falling into this category. If we denote this result as q, the indicator 'iflag' is then defined as $\exp(-qN)$, which will be one if it is near steady state, i.e. q is near zero. Numerical results to be presented later show that this works well.

2: Another way to improve the accuracy at the extreme points is to change the TVD limiter procedure m to a new limiter. Instead of using the information from a wide support, we are trying to approximate the high derivative. The details are given below.

We first define these three values:

$$d1 = \frac{1}{2}\|\bar{V}_{i+1} - \bar{V}_{i-1}\|, \qquad d2 = \|\bar{V}_{i+1} - 2\bar{V}_i + \bar{V}_{i-1}\|,$$

$$d = \|V^{-}_{i+\frac{1}{2}} - \bar{V}_i\| + 2\|V_i - \bar{V}_i\| + \|V^{+}_{i-\frac{1}{2}} - \bar{V}_i\|,$$

where $d1$, $d2$ are the approximations to absolute value of the first and second derivatives of V. If our V is a smooth function, by the Taylor's expansion, we can know that actually the following relation can be obtained:

$$\begin{aligned} d &= \|V^-_{i+\frac{1}{2}} - \bar{V}_i\| + 2\|V_i - \bar{V}_i\| + \|V^+_{i-\frac{1}{2}} - \bar{V}_i\| \\ &\approx \|\frac{V''}{12}\Delta x^2 + \frac{V'}{2}\Delta x\| + 2\|\frac{V''}{24}\Delta x^2\| + \|\frac{V''}{12}\Delta x^2 - \frac{V'}{2}\Delta x\| \\ &\approx \|\frac{d2}{12} + \frac{d1}{2}\| + 2\|\frac{d2}{24}\| + \|\frac{d2}{12} - \frac{d1}{2}\| \\ &\leq d1 + \frac{d2}{4} \end{aligned}$$

Consider the flag value $2\frac{d1+d2/4}{d}$. For smooth V, we know this flag value must be greater than 1 (even near the extreme). However, if the steady state solution is reached, $\bar{V}$ is constant and then $d1$ and $d2$ are both zero, hence this flag value equals to zero. If we define the limited value as

$$\tilde{V}^-_{i+\frac{1}{2}} := \bar{V}_i + \min(1, 2\frac{d1 + d2/4}{d})(\check{V}^-_{i+\frac{1}{2}} - \bar{V}_i) \tag{7.32}$$

this would give us a high order approximation, which also satisfies our well balanced requirement. $\tilde{V}^+_{i-\frac{1}{2}}$ and $\tilde{V}_i$ can be defined in a similar way.

To summarize, the *equilibrium balanced* variables $\tilde{V}^\pm_{i+\alpha}$ and $\tilde{V}^\pm_{i+\alpha}$ are given either directly

- by the TVD limited values $\check{V}^\pm_{i+\alpha}$ and $\check{V}^\pm_{i+\alpha}$ in (7.28), or
- by the weighted switch (7.31), or
- by the limiter type switch (7.32).

The corresponding conservative variables are given by

$$\tilde{U}^\pm_{i+\alpha} := U(\tilde{V}^\pm_{i+\alpha}, b^\pm_{i+\alpha}, \sigma^\pm_{i+\alpha}) \quad \text{for} \quad \alpha \in \{0, \frac{1}{2}\}. \tag{7.33}$$

As an immediate consequence of Lemma 7.2.4 we have

Corollary 7.2.6 *If $\bar{V}_{i\pm1} = \bar{V}_i$, then the equilibrium-limited values* (7.33) *satisfy*

$$V(\tilde{U}_i, b_i) = V(\tilde{U}^{\pm}_{i+\frac{1}{2}}, b^{\pm}_{i+\frac{1}{2}}) = \bar{V}_i. \tag{7.34}$$

The well balanced numerical test in Section 7.3 shows that the second method gives better results than the third, since the steady state cannot be preserved up to the roundoff error by using the latter method. The errors are at the level of 10^{-8} for double precision. But it can capture the small perturbation of the steady state well. Among all our numerical tests, we will use the second one, i.e. (7.31) together with (7.33).

7.2.3 A well-balanced quadrature rule for the source term

Suppose we have any points $x_L < x_R$ with corresponding values $(U, b)_L$ and $(U, b)_R$. Then we define the numerical residuum of the cell $[x_L, x_R]$ by

$$\begin{aligned} r^R_L &:= -Df^R_L + s^R_L \\ &:= -f(U_R) + f(U_L) \\ &\quad - g\left(\frac{h_R + h_L}{2} - \frac{m_L m_R (h_R - h_L)^2}{2m_L m_R(h_L + h_R) - 4g(h_L)^2(h_R)^2}\right)(b_R - b_L) \end{aligned} \tag{7.35}$$

Let us denote the central bracket on the RHS of (7.35) by $\tilde{h}^R_L$ and thus rewrite the source term as

$$s^R_L = -g\,\tilde{h}^R_L\,(b_R - b_L). \tag{7.36}$$

Note that this quadrature of the source term generalizes the well-balanced discretization for the lake at rest (i.e. $m_L = m_R = 0$), which is

$$-g\frac{h_L + h_R}{2}(b_R - b_L). \tag{7.37}$$

The quadrature (7.35) is motivated by analyzing the flux difference for moving equilibria. In this case we have $m_L = m_R = m$ and $E_L = E_R$. Let $\bar{h} := (h_L + h_R)/2$. Now the flux difference becomes

$$\begin{aligned} & f(U_R) - f(U_L) \\ &= \frac{1}{2} g(h_R)^2 + \frac{m^2}{h_R} - \frac{1}{2} g(h_L)^2 - \frac{m^2}{h_L} \\ &= \left(g\bar{h} - \frac{m^2}{h_L h_R} \right) (h_R - h_L) \end{aligned} \tag{7.38}$$

Now we use the relation

$$h_R - h_L = -(b_R - b_L) \frac{g h_L^2 h_R^2}{g h_L^2 h_R^2 - m^2 \bar{h}}, \tag{7.39}$$

which holds if $E_L = E_R$. Combining (7.38) and (7.39) yields

Lemma 7.2.7 *Suppose that $V(U_L, b_L) = V(U_R, b_R)$. Then*

$$f(U_R) - f(U_L) = -g\, \tilde{h}_L^R\, (b_R - b_L), \tag{7.40}$$

so the residuum r_L^R vanishes at steady state.

In the following two paragraphs we treat the interior residual and the residual in the boundary of each cell separately.

7.2.3.1 The cell boundary residual

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

To derive our scheme, we separate the boundary into two layers, see Figure 7.2. Take, for example the left boundary of cell i. We introduce points $x_C = x_{i-\frac{1}{2}} <$

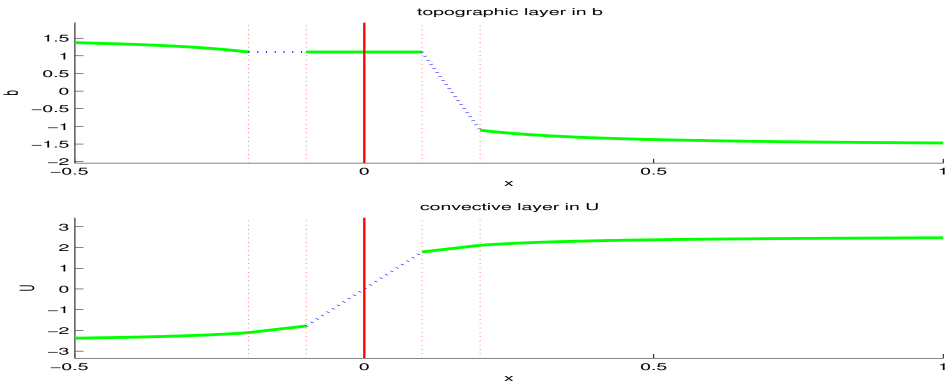


Figure 7.2: The boundary layer model. Top: discontinuous topography b. Bottom: shock-discontinuity in U. Transition layers are marked by dotted lines.

$x_B < x_A$ which are separated by an infinitesimal distance. Together with these we introduce the values

$$(U_A, b_A) := (\tilde{U}^+_{i-\frac{1}{2}}, b^+_{i-\frac{1}{2}}), \tag{7.41}$$

$$(U_B, b_B) := (\hat{U}^+_{i-\frac{1}{2}}, \hat{b}_{i-\frac{1}{2}}), \tag{7.42}$$

$$(U_C, b_C) := (\hat{U}_{i-\frac{1}{2}}, \hat{b}_{i-\frac{1}{2}}). \tag{7.43}$$

The values at point x_A are adjacent to the interior of the cell. The value $b^+_{i-\frac{1}{2}}$ is the WENO reconstructed bottom topography, and $\tilde{U}^+_{i-\frac{1}{2}}$ is the WENO reconstructed and equilibrium limited conservative variable (7.33). At the point x_B the topography from the left of cell i and the right of cell $i-1$ is merged,

$$\hat{b}_{i-\frac{1}{2}} = \max(b^-_{i-\frac{1}{2}}, b^+_{i-\frac{1}{2}}). \tag{7.44}$$

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

$$\hat{U}^+_{i-\frac{1}{2}} := U(\tilde{V}^+_{i-\frac{1}{2}}, \hat{b}_{i-\frac{1}{2}}, \sigma^+_{i-\frac{1}{2}}). \tag{7.45}$$

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

$$f(\hat{U}_{i-\frac{1}{2}}) = \hat{f}_{i-\frac{1}{2}} = F(\hat{U}^-_{i-\frac{1}{2}}, \hat{U}^+_{i-\frac{1}{2}}). \tag{7.46}$$

We can therefore distinguish two boundary layers within each cell. We call $[x_C, x_B]$ the convective and $[x_B, x_A]$ the topographic layer. In the convective layer, the topography is constant, so the source term s^B_C disappears and the residuum becomes a pure flux difference

$$r_i^{conv} := -f(U_B) + f(U_C) = -f(\hat{U}^+_{i-\frac{1}{2}}) + f(\hat{U}_{i-\frac{1}{2}}) \tag{7.47}$$

as for the homogeneous conservation law.

In the topographic layer the bottom b changes while the equilibrium variables $V = (m, E)$ remain constant. According to Lemma 7.2.7 the overall residuum vanishes,

$$r_i^{topo} = 0, \tag{7.48}$$

and by definition (7.35) of the residuum

$$s_i^{topo} = f(U_B) - f(U_A) = f(\hat{U}^+_{i-\frac{1}{2}}) - f(\tilde{U}^+_{i-\frac{1}{2}}). \tag{7.49}$$

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

7.2.3.2 The interior residual

Based on the reconstructed values $\hat{U}^\pm_{i+\frac{1}{2}}$ in (7.45) we now define the residual in the interior of the cell as

$$\bar{r}_i^{int} := r_L^R \tag{7.50}$$

with $(U_L, b_L) = (\hat{U}^+_{i-\frac{1}{2}}, b^+_{i-\frac{1}{2}})$ and $(U_R, b_R) = (\hat{U}^-_{i+\frac{1}{2}}, b^-_{i+\frac{1}{2}})$. This residual is so far only second order accurate. But we can directly adapt the extrapolation idea used in the paper of Noelle et al. [30], and obtain a high order discretization.

We first subdivide each cell into N subcells and apply the quadrature (7.36) to all subcells. Then we can have the following quadratures S_N:

$$S_N = \sum_{j=1}^{N} s_L^R(U^+_{j-1}, U^-_j, b^+_{j-1}, b^-_j) \tag{7.51}$$

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

$$\begin{aligned} S_N &= \sum_{j=1}^{N} s_L^R(U^+_{j-1}, U^-_j, b^+_{j-1}, b^-_j) \\ &= \sum_{j=1}^{N} \left(f(U^-_j) - f(U^+_{j-1})\right) \\ &= f(U^-_N) - f(U^+_0) = f(U^-_{i+\frac{1}{2}}) - f(U^+_{i-\frac{1}{2}}). \end{aligned}$$

This shows that S_N is also a second order well balanced approximation to the source term. Note that the quadrature S_1 (7.36) is second order accurate and symmetric, therefore, there exists an asymptotic expansion:

$$S_N = S + c_1 \left(\frac{\Delta x}{N}\right)^2 + c_2 \left(\frac{\Delta x}{N}\right)^4 + \cdots, \tag{7.52}$$

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

$$\frac{4S_2 - S_1}{3}. \tag{7.53}$$

Compared with the second order discretization (7.36), the fourth order well balanced scheme here needs one additional reconstructed point value at the cell center per cell, which is necessary for the computation of S_2.

7.2.4 Summary of the scheme

The fourth order well-balanced scheme is given by

$$\frac{d}{dt}\bar{U}_i := \frac{1}{\Delta x_i}\left(-F(\hat{U}^-_{i+\frac{1}{2}},\hat{U}^+_{i+\frac{1}{2}}) + F(\hat{U}^-_{i-\frac{1}{2}},\hat{U}^+_{i-\frac{1}{2}}) + s_i\right). \tag{7.54}$$

Here the function $F(\cdot,\cdot)$ is a conservative, Lipschitz continuous numerical flux consistent with the shallow water flux, i.e. $F(U,U) = f(U)$ for all U. The left and right values $\hat{U}^\pm_{i+\frac{1}{2}}$ at the cell interface are defined in (7.45).

The total source term s_i is given by

$$s_i := \frac{4S_2 - S_1}{3} + f(\hat{U}^+_{i-\frac{1}{2}}) - f(\tilde{U}^+_{i-\frac{1}{2}}) + f(\tilde{U}^-_{i+\frac{1}{2}}) - f(\hat{U}^-_{i+\frac{1}{2}}). \tag{7.55}$$

The extrapolated interior source term $(4S_2 - S_1)/3$ is defined by

$$S_1 := s_L^R(\tilde{U}^+_{i-\frac{1}{2}}, \tilde{U}^-_{i+\frac{1}{2}}, b^+_{i-\frac{1}{2}}, b^-_{i+\frac{1}{2}}) \tag{7.56}$$

$$S_2 := \left(s_L^R(\tilde{U}^+_{i-\frac{1}{2}}, \tilde{U}_i, b^+_{i-\frac{1}{2}}, b_i) + s_L^R(\tilde{U}_i, \tilde{U}^-_{i+\frac{1}{2}}, b_i, b^-_{i+\frac{1}{2}})\right) \tag{7.57}$$

and the well-balanced quadrature of the source term s_L^R is given by

$$s_L^R(U_L, U_R, b_L, b_R) := -g\left(\frac{h_R + h_L}{2} - \frac{m_L m_R (h_R - h_L)^2}{2 m_L m_R (h_L + h_R) - 4g(h_L)^2(h_R)^2}\right)(b_R - b_L) \tag{7.58}$$

The scheme is completed by a TVD Runge-Kutta discretisation [41] in time.

Algorithm 7.2.8 *An implementation of this algorithm may follow the following steps:*

1. *Compute the initial cell average of U and bottom b based on the initial data.*

Apply the WENO reconstruction on $\bar{b}_i$ to obtain point values of b (may be ignored if bottom b is prescribed as a function of x.

2. *At each time step, perform the usual WENO-LF or WENO-LLF approximation on the cell average $\bar{U}_i$, and obtain $U^{\pm}_{i+\frac{1}{2}}$, hence $V^{\pm}_{i+\frac{1}{2}}$. Compute U_i and V_i to obtain fourth order accuracy.*

3. *Compute the reference value $\bar{V}_i$ as the implicit solution of equation* (7.23).

4. *Apply the equilibrium limiter* (7.26) *on the cell averages $\bar{V}_i, \bar{V}_{i\pm 1}$, and on the pointvalues $V^{\pm}_{i+\frac{1}{2}}, V_i$, to get the limited values $\check{V}^{\pm}_{i+\frac{1}{2}}$ and $\check{V}_i$. Then apply the limiter* (7.31) *to obtain $\tilde{V}^{\pm}_{i+\frac{1}{2}}$ and $\tilde{V}_i$.*

5. *Compute the numerical fluxes on the RHS of* (7.54).

6. *Compute the high order discretization to the source term* (7.55)–(7.58).

7. *Apply a TVB Runge-Kutta scheme [41] to* (7.54) *to advance $\bar{U}_i(t)$ in time.*

Collecting the results of this section it is straightforward to prove the following

Theorem 7.2.9 *The WENO scheme* (7.54)–(7.58) *maintains the moving steady state solution (4.5) exactly and is high order accurate. The same holds for the fully discrete scheme.*

Proof. Suppose that the initial data are a moving steady state, $V(x) \equiv \bar{V}$. Then Lemma 7.2.3 implies that all reference values $\bar{V}_i$ coincide with $\bar{V}$. Corollary 7.2.6 implies that $V(\tilde{U}_i, b_i) = V(\tilde{U}^{\pm}_{i+\frac{1}{2}}, b^{\pm}_{i+\frac{1}{2}}) = \bar{V}_i$. Now Lemma 7.2.7 implies that the interior residual vanishes, $r_i^{int} = 0$. Since we know from (7.48) that there is no residual in the topographic layer, $r_i^{topo} = 0$, it remains to show that the residual in the convective layer, r_i^{conv} vanishes as well. For this we study not only the values $\hat{U}^{+}_{i-\frac{1}{2}}$ and $\hat{U}_{i-\frac{1}{2}}$, but also the corresponding value $\hat{U}^{-}_{i-\frac{1}{2}}$ from the neighboring cell I_{i-1}.

Since $\bar{V}_{i-1} = \bar{V}_i$, it follows that $\hat{U}^-_{i-\frac{1}{2}} = \hat{U}^+_{i-\frac{1}{2}}$ and hence $\hat{f}_{i-\frac{1}{2}} = f(\hat{U}^+_{i-\frac{1}{2}})$. Therefore

$$r_i^{conv} = f_{i-\frac{1}{2}} - f(\hat{U}^+_{i-\frac{1}{2}}) = 0 \tag{7.59}$$

and

$$r_i = r_i^{int} + r_i^{topo} + r_i^{conv} = 0, \tag{7.60}$$

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

We can easily check the two conditions of Lemma 7.2.1 are satisfied for our scheme. This proves the high order accuracy. □

Remark 7.2.10 *The steady state solution we are trying to maintain in the previous well balanced scheme is $m =$ constant and $E =$ constant globally. There also exists the cases when a shock appears in the steady state, hence m, E are both piecewise constants, and satisfy the Rankine-Hugoniot jump condition at the shock. In that case, we may meet problems near the shock position – the condition of the theorem (7.2.7) is not true any more.*

Here we only deal with the case when the shock position is exactly on the cell boundary and leave the general case for further investigation. Suppose we already know the shock position at the beginning, then we perform the one sided limiter procedure on the cells next to the shock. This will make sure that $\hat{V}^{\pm}_{i+\frac{1}{2}}$ equal to constant piecewisely. Since the Roe's flux is the only one which can capture the shock exactly, we use WENO-Roe instead of WENO-LF here. And the condition of the theorem (7.2.7) is still true since the shock is exact on the boundary and $\hat{V}^-_{i+\frac{1}{2}} = \hat{V}^+_{i-\frac{1}{2}}$ is true for any i. Then we can still have a well balanced scheme for this special case.

7.3 One dimensional numerical results

In this section we present numerical results of our fourth order finite volume WENO scheme satisfying the well balanced property for the one dimensional shallow water equations (1.3). In all the examples, time discretization is by the classical third order Runge-Kutta method, and the CFL number is taken as 0.6, except for the accuracy tests where smaller time step is taken to ensure that spatial errors dominate. The gravitation constant g is taken as $9.812 m/s^2$.

7.3.1 Well balanced test

The purpose of the first test problems is to verify the well balanced property of our algorithm towards the moving steady state solution. These steady state problems are classical test cases for transcritical and subcritical flows, and they are widely used to test numerical schemes for shallow water equations. For example, they have been used as a test case in, e.g. [43]. Here, our purpose is to maintain these steady state solutions exactly.

The bottom function is given by:

$$b(x) = \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 \le x \le 12 \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

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:

$$E = \frac{1.53^2}{2 \times 0.66^2} + 9.812 \times 0.66, \qquad m = 1.53, \tag{7.2}$$

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 7.1. In order to demonstrate that the steady state 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 7.1 for different precisions. We can clearly see that the L^1 and L^∞ errors are at the level of round-off errors for different precisions, verifying the well balanced property.

Table 7.1: L^1 and L^∞ errors for different precisions for the transcritical flow without a shock.

	L^1 error		L^∞ error	
precision	h	hu	h	hu
single	3.43E-05	5.61E-05	9.35E-04	6.56E-05
double	5.63E-16	1.51E-15	2.05E-15	6.66E-15
quadruple	5.38E-34	2.14E-33	1.73E-33	6.55E-33

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

$$E = \begin{cases} \dfrac{0.18^2}{2 \times 0.4137^2} + 9.812 \times 0.41372 & \text{if } x \leq 11.665511784112317 \\ \dfrac{0.18^2}{2 \times 0.33^2} + 9.812 \times 0.33 & \text{otherwise} \end{cases} \qquad m = 0.18, \tag{7.3}$$

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.

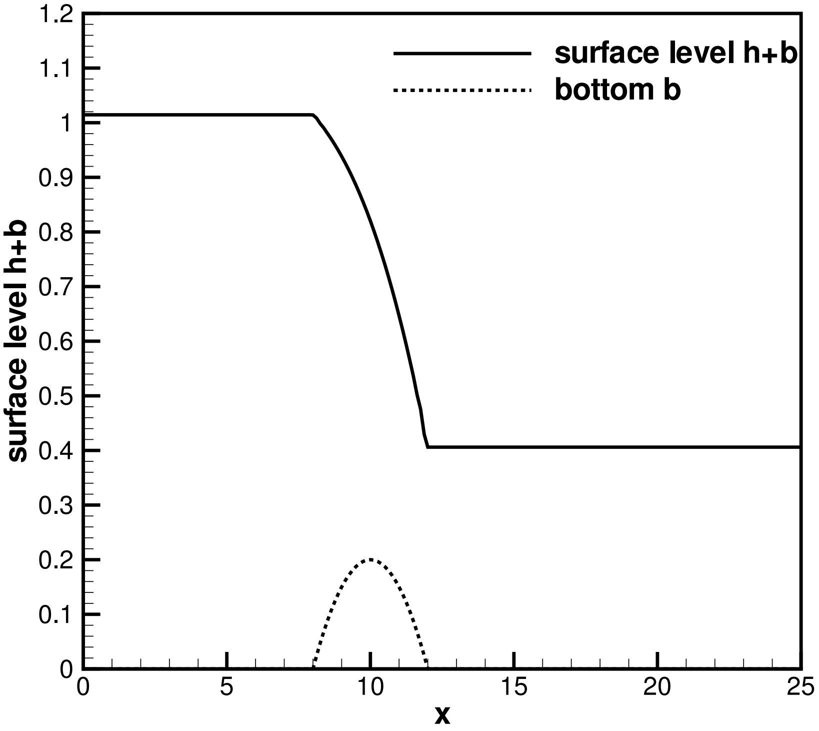


Figure 7.1: The surface level $h+b$ and the bottom b for the transcritical flow without a shock.

This steady state should be exactly preserved. As we mentioned in Section 7.2, 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. Also, we mentioned that the left and right approximated values of bottom at the shock must be exact, so that the Roe's flux can capture this shock exactly. Here we compute the solution until $t = 20$ using $N = 400$ uniform mesh points. The computed surface level $h + b$ and the bottom b are plotted in Figure 7.2. In order to demonstrate that the steady state 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 and the discharge hu in Tables 7.2 for different precisions. We can clearly see that the L^1 and L^∞ errors are at the level of round-off errors for different precisions, verifying the well balanced property.

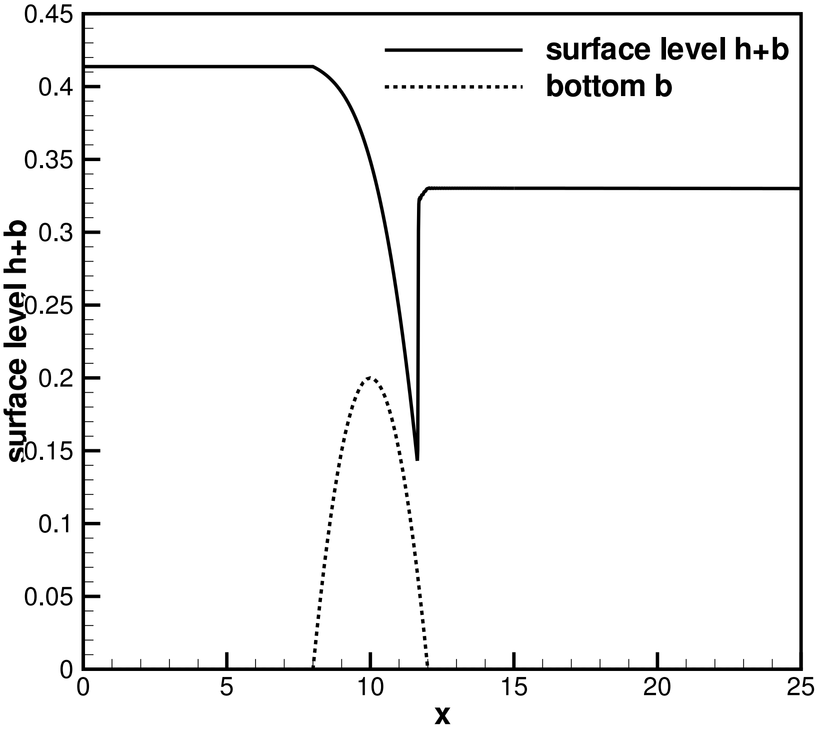


Figure 7.2: The surface level $h+b$ and the bottom b for the transcritical flow with a shock.

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

$$E = 22.06605, \qquad m = 4.42, \tag{7.4}$$

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. when the flow is subcritical.

Table 7.2: L^1 and L^∞ errors for different precisions for the transcritical flow with a shock.

precision	L^1 error		L^∞ error	
	h	hu	h	hu
single	7.20E-06	1.38E-06	2.03E-03	1.42E-05
double	4.26E-18	4.01E-18	2.22E-16	2.22E-16

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 7.3. In order to demonstrate that the steady state 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 and the discharge hu in Tables 7.3 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.

Table 7.3: L^1 and L^∞ errors for different precisions for the subcritical flow.

precision	L^1 error		L^∞ error	
	h	hu	h	hu
single	2.39E-05	8.51E-05	3.97E-05	1.86E-04
double	2.66E-16	3.03E-15	1.11E-15	9.77E-15
quadruple	1.34E-35	4.87E-34	5.78E-34	3.08E-33

7.3.2 Testing the orders of accuracy

In this example we will test the high order accuracy of our schemes for a smooth solution. Following the examples presented in [46], we have the bottom function and

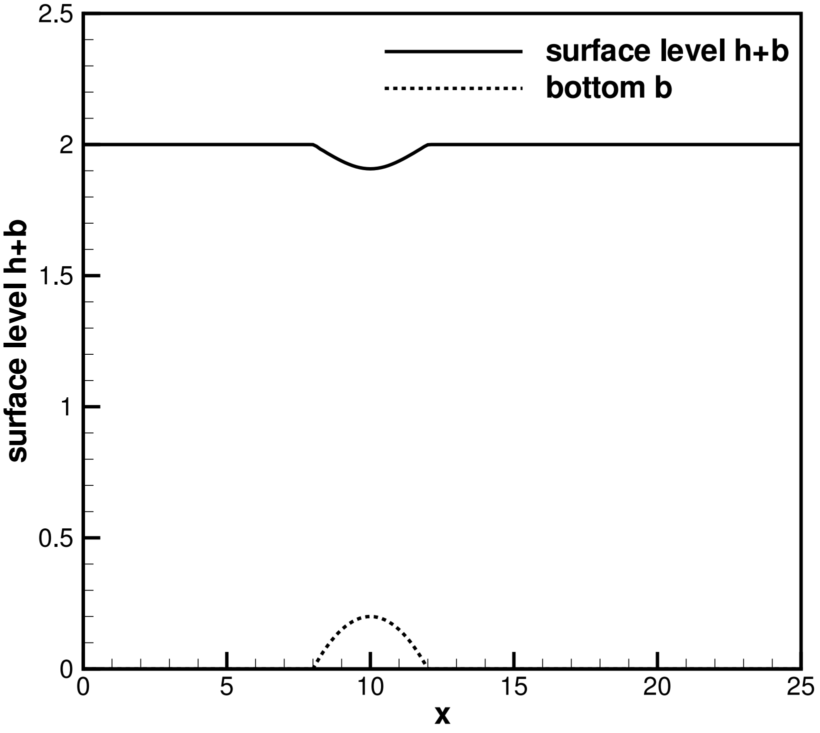


Figure 7.3: The surface level $h + b$ and the bottom b for the subcritical flow.

initial conditions

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

with periodic boundary conditions. Since the exact solution is not known explicitly for this case, we use the fifth order finite volume non well-balanced WENO scheme with $N = 12,800$ cells to compute a reference solution, and treat this reference solution as the exact solution in computing the numerical errors. We compute up to $t = 0.1$ when the solution is still smooth (shocks develop later in time for this problem). Table 7.4 contains the L^1 errors for the cell averages and numerical orders of accuracy for the finite volume schemes, respectively. Notice that the CFL number we have used decreases with the mesh size and is recorded in Table 7.4. We can easily observe the fifth-order accuracy for the WENO schemes. Note that the fifth-order WENO reconstruction has been used in space, but the source term is approximated by a fourth order accurate extrapolation. Hence the approximation of the source term in the algorithm contributes less to the overall error. This phenomena has been investigated in [30].

Table 7.4: L^1 errors and numerical orders of accuracy for the example in Section 7.3.2.

No. of cells	CFL	h		hu	
		L^1 error	order	L^1 error	order
25	0.6	8.60E-04		1.77E-02	
50	0.6	4.21E-05	4.35	1.26E-03	3.81
100	0.4	1.38E-06	4.93	5.32E-05	4.56
200	0.3	5.55E-08	4.64	2.42E-06	4.46
400	0.2	1.83E-09	4.93	8.42E-08	4.84
800	0.1	4.39E-11	5.37	1.76E-09	5.58

7.3.3 A small perturbation of a moving steady-state water

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

In the subsection 7.3.1, we present three steady state solutions and show that our numerical schemes do maintain them exactly. In this test case, we impose to them a small perturbation 0.01 on the height in the interval [5.75,6.25].

Theoretically, this disturbance should split into two waves, propagating left and right. 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 7.4 for the transcritical flow without a shock, in Figure 7.5 for the transcritical flow with a shock and in Figure 7.6 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.

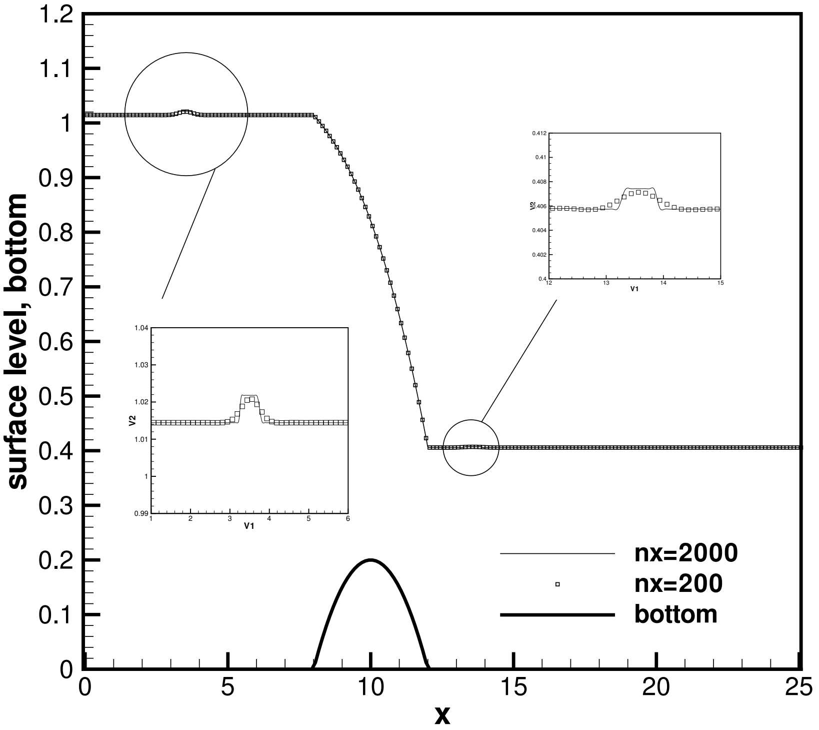


Figure 7.4: Small perturbation of the transcritical flow without a shock.

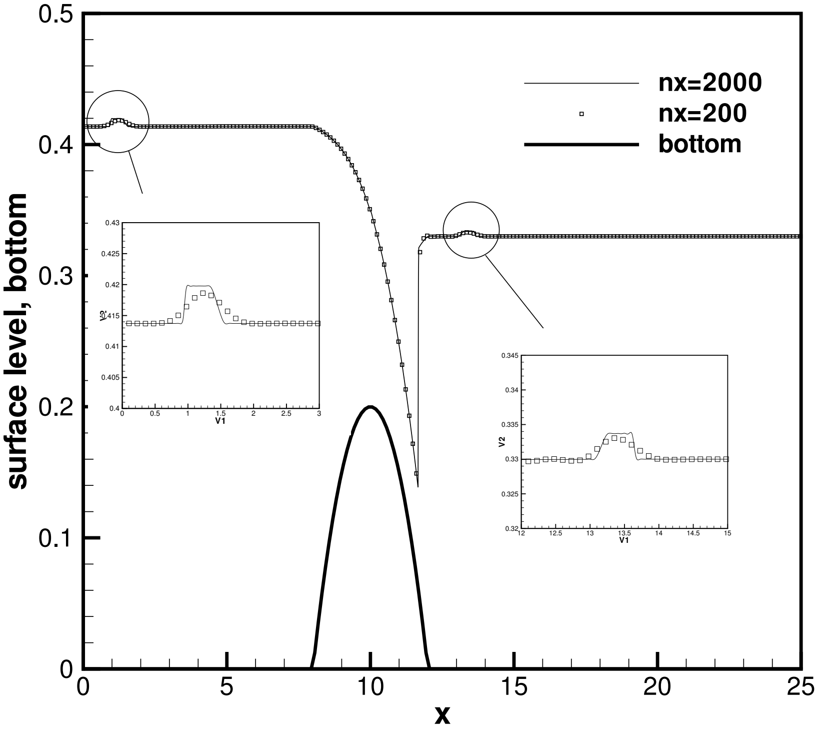


Figure 7.5: Small perturbation of the transcritical flow with a shock.

7.3.4 The dam breaking problem over a rectangular bump

In this traditional test case we simulate the dam breaking problem over a rectangular bump, which produces a rapidly varying flow over a discontinuous bottom topography. This example was used in [44, 46, 30].

The bottom topography takes the form:

$$b(x) = \begin{cases} 8 & \text{if } |x - 750| \leq 1500/8 \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

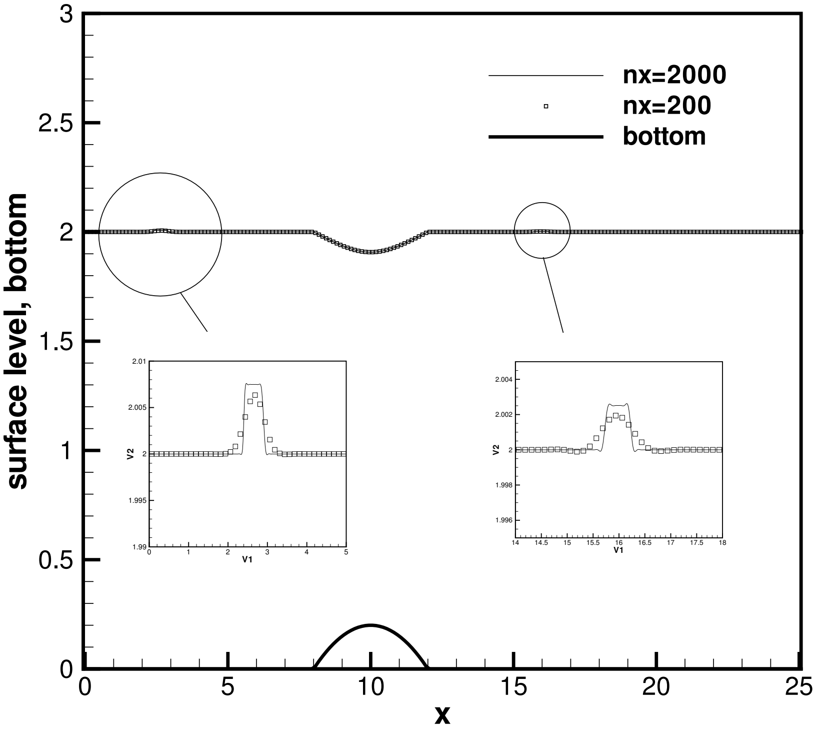


Figure 7.6: Small perturbation of the subcritical flow.

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

$$(hu)(x,0) = 0 \quad \text{and} \quad h(x,0) = \begin{cases} 20 - b(x) & \text{if } x \leq 750 \\ 15 - b(x) & \text{otherwise} \end{cases} \tag{7.6}$$

We use open boundary conditions on both sides. In the beginning, we observe the standard rarefaction and shock waves which form the solution of the Riemann problem of the homogeneous shallow water equations. The numerical results with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Figures 7.7 at ending time t=15s. At time T≈17, the waves reach the discontinuous edges of the bottom. After that, a part of the wave is transmitted, another part reflected, and a remaining part becomes a standing wave. Later on, this wave system keeps interacting. When the time T reaches 60, six waves appears in our solution. The numerical results with 400 uniform cells (and a comparison with the results using 4000 uniform cells) are shown in Figures 7.8 at ending time t=60s.

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

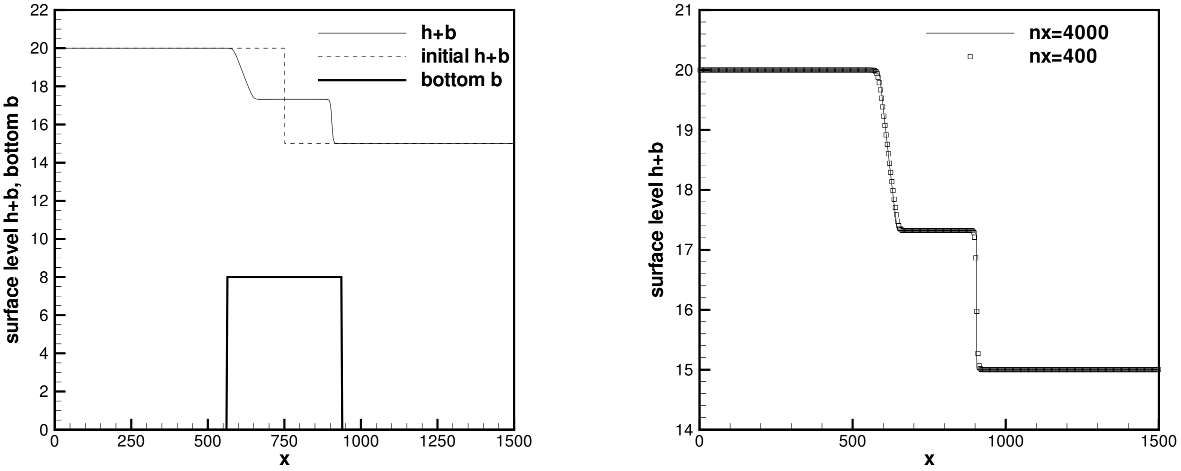


Figure 7.7: The surface level $h + b$ for the dam breaking problem at time t=15s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

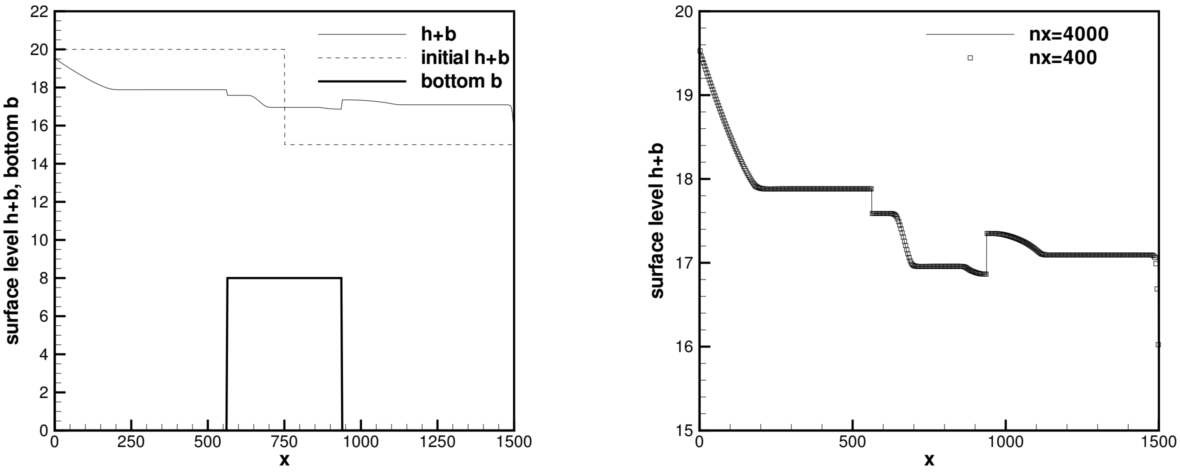


Figure 7.8: The surface level $h + b$ for the dam breaking problem at time t=60s. Left: the numerical solution using 400 grid cells, plotted with the initial condition and the bottom topography; Right: the numerical solution using 400 and 4000 grid cells.

Chapter 8

Conclusion

8.1 Summary

The main contribution of this thesis is the design of well balanced high order numerical schemes for a class of hyperbolic systems with separable source terms. This class of hyperbolic systems includes the shallow water equations, the elastic wave equation, the hyperbolic model for a chemosensitive movement, the nozzle flow, a two phase flow model, a model of fluid mechanics in case of spherical symmetry and other systems.

In Chapter Two, we give a short review of three common used high order numerical schemes, including finite difference WENO, finite volume WENO and finite element discontinuous Galerkin schemes.

In Chapter Three, Four and Five, we first design high order well balanced WENO schemes for the still water solution of the shallow water equations, and then generalize our idea to a general class of balance laws with separable source terms. Well balanced high order finite volume WENO schemes and finite element discontinuous Galerkin schemes are also designed for the same class of balance laws, which are more suitable for computations in complex geometry and / or for using adaptive meshes. The key ingredient in our design is a special decomposition of the source term before discretization, which allows us to design specific approximations such that the resulting WENO schemes satisfy the well balanced property, and at the same time maintain their original high order accuracy and essentially non-oscillatory property for general solutions.

We discuss a new approach of high order well balanced finite volume WENO schemes and RKDG finite element schemes in Chapter Six. Traditional RKDG methods with a special treatment of the flux are proven to be well balanced for certain steady state solutions, and can maintain their original high order accuracy and essentially non-oscillatory property for general solutions. Finite volume WENO schemes can be modified due to similar ideas to obtain those properties. Comparing with the well balanced schemes developed in Chapter Five, the well balanced RKDG schemes here are simpler and involve less modification to the original RKDG methods, while the well balanced WENO finite volume schemes here and that in Chapter Five are comparable in computational cost. Similar idea can be generalized to the finite difference WENO scheme, but it is more complicated compared with the scheme presented in Chapter Three. Hence we do not include it in this thesis.

In Chapter Seven, a high order well balanced finite volume WENO scheme has been designed for the moving steady state solution of the shallow water equations. Only one dimensional case is considered so far. This example can not be treated by the numerical schemes introduced in the previous chapters, and need different techniques to obtain the well balanced property. A special discritization of the source term and the flux terms is introduced there.

8.2 Ongoing and Future Work

In this thesis, we have designed high order well balanced numerical schemes for a class of hyperbolic systems with source terms. Our long term plan is to develop such schemes for more general systems, and explore more applications in the framework of well balanced schemes.

Stiff source term problems arising from chemical reaction, combustion and turbulent modeling have attracted more and more attention in these years. They have wide application in the real world. There are some connections between the well balanced schemes and those problems. Several stiff source term problems can already be solved using the framework mentioned in this thesis. More connections will be explored.

Designing well balanced schemes for multi-layer shallow water equations and high order numerical scheme for general nonconservative systems will also be investigated. The two-layer shallow water equations involve nonconservative terms, which also appear in many other models from physical world, such as two phase flows. The presence of nonconservative products makes it difficult to define weak solutions. Pares and Castro et al. [31, 7] have designed some numerical methods, which depend on the choice of the family of paths, and their high order version is not consistent with conservative schemes for the hyperbolic problems. Application of the idea of "well balanced property" there would be interesting.

Bibliography

[1] J.D. Anderson, *Computational Fluids Dynamics*, McGraw-Hill, New York, 1995.

[2] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein and B. Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM Journal on Scientific Computing, 25 (2004), pp.2050-2065.

[3] D.S. Bale, R.J. LeVeque, S. Mitran and J.A. Rossmanith, *A wave propagation method for conservation laws and balance laws with spatially varying flux functions*, SIAM Journal on Scientific Computing, 24 (2002), pp.955-978.

[4] D.S. Balsara and C.-W. Shu, *Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy*, Journal of Computational Physics, 160 (2000), pp.405-452.

[5] A. Bermudez and M.E. Vazquez, *Upwind methods for hyperbolic conservation laws with source terms*, Computers and Fluids, 23 (1994), pp.1049-1071.

[6] R. Botchorishvili, B. Perthame and A. Vasseur, *equilibrium schemes for scalar conservation laws with stiff sources*, Mathematics of Computation, 72(2003), pp.131-157. Also, an extended version containing more numerical examples is located at http://www.inria.fr/rrrt/rr-3891.html

[7] M.J. Castro, J.M. Gallardo and C. Pares, *High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative*

products. Applications to shallow water systems, Mathematics of Computation, to appear.

[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, pp.69-224.

[9] B. Cockburn, S. Hou and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Mathematics of Computation, 54 (1990), pp.545-581.

[10] 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*, Journal of Computational Physics, 84 (1989), pp.90-113.

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

[12] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, Journal of Computational Physics, 141 (1998), pp.199-224.

[13] B. Cockburn and C.-W. Shu, *Runge-Kutta Discontinuous Galerkin methods for convection-dominated problems*, Journal of Scientific Computing, 16 (2001), pp.173-261.

[14] N. Crnjaric-Zic, S. Vukovic and L. Sopta, *Balanced finite volume WENO and central WENO schemes for the shallow water and the open-channel flow equations*, Journal of Computational Physics, 200 (2004), pp.512-548.

[15] F. Filbet and C.-W. Shu, *Approximation of hyperbolic models for chemosensitive movement*, SIAM Journal on Scientific Computing, v27 (2005), pp.850-872.

[16] Ll. Gascón and J.M. Corberán, *Construction of second-order TVD schemes for nonhomogeneous hyperbolic conservation laws*, Journal of Computational Physics, 172 (2001), pp.261-297.

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

[18] J.M. Greenberg and A.Y. LeRoux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM Journal on Numerical Analysis, 33 (1996), pp.1-16.

[19] A. Harten, B. Engquist, S. Osher and S. Chakravathy, *Uniformly high order accurate essentially non-oscillatory schemes, III*, Journal of Computational Physics, 71 (1987), pp.231-303.

[20] A. Harten, P.D. Lax and B. Van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Review, 25 (1983), pp.35-61.

[21] T. Hillen, *Hyperbolic models for chemosensitive movement*, Mathematical Models and Methods in Applied Sciences, 12 (2002), pp.1007-1034.

[22] C. Hu and C.-W. Shu, *Weighted essentially non-oscillatory schemes on triangular meshes*, Journal of Computational Physics, 150 (1999), pp.97-127.

[23] M.E. Hubbard and P. Garcia-Navarro, *Flux difference splitting and the balancing of source terms and flux gradients*, Journal of Computational Physics, 165 (2000), pp.89-125.

[24] G. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, Journal of Computational Physics, 126 (1996), pp.202-228.

[25] S. Jin, *A steady state capturing method for hyperbolic systems with geometrical source terms*, Mathematical Modelling and Numerical Analysis, 35 (2001), pp.631-646.

[26] S. Karni, E. Kirr, A. Kurganov and G. Petrova, *Compressible two-phase flows by central and upwind schemes*, Mathematical Modelling and Numerical Analysis, 38 (2004), pp.477-494.

[27] A. Kurganov and D. Levy, *Central-upwind schemes for the Saint-Venant system*, Mathematical Modelling and Numerical Analysis, 36 (2002), pp.397-425.

[28] R.J. LeVeque, *Balancing source terms and flux gradients on high-resolution Godunov methods: the quasi-steady wave-propagation algorithm*, Journal of Computational Physics, 146 (1998), pp.346-365.

[29] X.-D. Liu, S. Osher and T. Chan, *Weighted essentially nonoscillatory schemes*, Journal of Computational Physics, 115 (1994), pp.200-212.

[30] S. Noelle, N. Pankratz, G. Puppo and J.R. Natvig, *Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows*, Journal of Computational Physics, 213 (2006), pp.474-499.

[31] C. Pares and M.J. Castro, *On the well-balanced property of Roe's method for nonconservative hyperbolic sysytem: Applications to shallow water systems.* Mathematical Modelling and Numerical Analysis, 38 (2004), pp.821-852.

[32] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp.201-231.

[33] J. Qiu and C.-W. Shu, *Runge-Kutta discontinuous Galerkin method using WENO limiters*, SIAM Journal on Scientific Computing, 26 (2005), pp.907-929.

[34] T.C. Rebollo, A.D. Delgado and E.D.F. Nieto, *A family of stable numerical solvers for the shallow water equations with source terms*, Computer Methods in Applied Mechanics and Engineering, 192 (2003), pp.203-225.

[35] P.L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, Journal of Computational Physics, 43 (1981), pp.357-372.

[36] G. Russo, *Central schemes for balance laws*, Proceedings of the VIII International Conference on Nonlinear Hyperbolic Problems, Magdeburg, 2000.

[37] R. Saurel and R. Abgrall, *A multiphase Godunov method for compressible multifluid and multiphase flows*, Journal of Computational Physics, 150 (1999), pp.425-467.

[38] J. Shi, C. Hu and C.-W. Shu, *A technique of treating negative weights in WENO schemes*, Journal of Computational Physics, 175 (2002), pp.108-127.

[39] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Mathematics of Computation, 49 (1987), pp.105-121.

[40] 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, pp.325-432.

[41] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics, 77 (1988), pp.439-471.

[42] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes, II*, Journal of Computational Physics, 83 (1989), pp.32-78.

[43] M.E. Vazquez-Cendon, *Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry*, Journal of Computational Physics, 148 (1999), pp.497-526.

[44] S. Vukovic and L. Sopta, *ENO and WENO schemes with the exact conservation property for one-dimensional shallow water equations*, Journal of Computational Physics, 179 (2002), pp.593-621.

[45] S. Vukovic, N. Crnjaric-Zic and L. Sopta, *WENO schemes for balance laws with spatially varying flux*, Journal of Computational Physics, 199 (2004), pp.87-109.

[46] Y. Xing and C.-W. Shu, *High order finite difference WENO schemes with the exact conservation property for the shallow water equations*, Journal of Computational Physics, 208 (2005), pp.206-227.

[47] Y. Xing and C.-W. Shu, *High order well-balanced finite difference WENO schemes for a class of hyperbolic systems with source terms*, Journal of Scientific Computing, to appear.

[48] 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*, Journal of Computational Physics, 214 (2006), pp.567-598.

[49] 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*, Communication in Computational Physics, 1 (2006), pp.100-134.

[50] K. Xu, *A well-balanced gas-kinetic scheme for the shallow-water equations with source terms*, Journal of Computational Physics, 178 (2002), pp.533-562.

[51] J.G. Zhou, D.M. Causon, C.G. Mingham and D.M. Ingram, *The surface gradient method for the treatment of source terms in the shallow-water equations*, Journal of Computational Physics, 168 (2001), pp.1-25.

[52] T. Zhou, Y. Li and C.-W. Shu, *Numerical comparison of WENO finite volume and Runge-Kutta discontinuous Galerkin methods*, Journal of Scientific Computing, 16 (2001), pp.145-171.