ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Numerical Methods for the Shallow Water Equations with Bottom Topography

Bachelor Thesis

Richard Mueller

Monday 27th August, 2012

Advisors: Prof. Dr. R. Hiptmair
Department of Mathematics, ETH Zürich

Contents

Chapter 1

Introduction

1.1 The Shallow Water Equations

The shallow water equations model flows in lakes, rivers, near shore ocean, etc., especially in phenomena where the vertical scales are much smaller than the horizontal ones. I present a short sketch of the derivation of the one-dimensional shallow water equations with flat bottom topography. Consider a channel with unit width filled with water. It is assumed that the vertical velocity is negligible. The horizontal velocity is approximately constant throughout any cross section and it depends on x and t (space and time). The water is assumed to be incompressible, so its density $\bar{\rho}$ is constant. The height h of the water level changes in time and space, so $h = h(x,t)$. One assumes the conservation of mass and momentum throughout the channel. The mass in an arbitrary interval $[a,b]$ at time t is given by

$$\int_a^b \bar{\rho} h(x,t) \mathrm{d}x \ . \tag{1.1}$$

The mass flux is given by the momentum density $\bar{\rho}u$ integrated over the vertical scale. So the mass flux is given by

$$\bar{\rho} h(x,t) u(x,t) \ . \tag{1.2}$$

If (1.2) is smooth enough, the mass flux can be rewritten as

$$\bar{\rho} \int \frac{\partial}{\partial x} (u(x,t)h(x,t)) \mathrm{d}x \ . \tag{1.3}$$

The change of mass in time in the space interval $[a,b]$ only depends on the mass flux, so the following relation must hold

$$\frac{\mathrm{d}}{\mathrm{d}t} (\int_a^b \bar{\rho} h(x,t) \mathrm{d}x) + \bar{\rho} \int \frac{\partial}{\partial x} (u(x,t)h(x,t)) \mathrm{d}x = 0 \ . \tag{1.4}$$

If one assumes that h is "nice" enough (smooth and integrable), the integration and differentiation can be interchanged

$$\bar{\rho}\int_a^b \frac{\partial}{\partial t}h(x,t)\mathrm{d}x + \bar{\rho}\int \frac{\partial}{\partial x}(u(x,t)h(x,t))\mathrm{d}x = 0\ . \tag{1.5}$$

Formally, one has

$$h_t + (hu)_x = 0 \tag{1.6}$$

since $[a,b]$ was arbitrary and the equation (1.5) must hold for all such space intervals in the domain. The constant $\bar{\rho}$ drops out.
The momentum in an arbitrary interval is given by

$$\bar{\rho}\int_a^b h(x,t)u(x,t)\mathrm{d}x\ . \tag{1.7}$$

The momentum flux is given by (see [4])

$$\bar{\rho}h(x,t)(u(x,t))^2 + p(x,t) = \int \frac{\partial}{\partial x}(\bar{\rho}h(x,t)(u(x,t))^2 + p(x,t))\mathrm{d}x \tag{1.8}$$

where p is the pressure. In (1.8) it is assumed that h, u and p are "nice" enough so that the integration and differentiation can be interchanged. The change of momentum in time only depends on the momentum flux at the endpoint. Thus, one gets

$$\frac{\mathrm{d}}{\mathrm{d}t}(\int_a^b \bar{\rho}h(x,t)u(x,t)\mathrm{d}x) + \int_a^b \frac{\partial}{\partial x}(\bar{\rho}h(x,t)(u(x,t))^2 + p(x,t))\mathrm{d}x = 0\ . \tag{1.9}$$

In differential form the above equation (1.9) can be written as follows

$$(hu)_t + (hu^2 + p)_x = 0\ . \tag{1.10}$$

The pressure p can be determined by the hydrostatic law, which says that the pressure at $h-y$ below the water level is given by

$$\bar{\rho}g(h-y) \tag{1.11}$$

where g is the gravitational constant. Integrating (1.11) from $y=0$ to $y=h$ yields

$$\bar{\rho}g\int_0^h h - y\mathrm{d}y = \frac{1}{2}\bar{\rho}gh^2\ . \tag{1.12}$$

Replacing p in (1.10) by (1.12) and dropping the constant $\bar{\rho}$ yields

$$(hu)_t + (hu^2 + \frac{1}{2}gh^2)_x = 0\ . \tag{1.13}$$

Combining equation (1.6) and (1.13) gives the conservative form of the one-dimensional shallow water equations

$$\begin{aligned} h_t + (hu)_x &= 0 \\ (hu)_t + (hu^2 + \frac{1}{2}gh^2)_x &= 0 \ . \end{aligned} \tag{1.14}$$

The conservative form of the shallow water equations in two dimensions looks as follows

$$\begin{aligned} h_t + (hu)_x + (hv)_y &= 0 \\ (hu)_t + (hu^2 + \frac{1}{2}gh^2)_x + (huv)_y &= 0 \\ (hv)_t + (huv)_x + (hv^2 + \frac{1}{2}gh^2)_y &= 0 \ . \end{aligned} \tag{1.15}$$

In the above formulation I have neglected the eddy viscosity and assumed a flat bottom topography, therefore there is no source term. Eddy viscosity is the transfer of momentum from the large scales to the smaller scales of the motion. In the rest of the thesis I will never consider the eddy viscosity. But I introduce the shallow water equations with some given bottom topography.

$$\begin{aligned} h_t + (hu)_x + (hv)_y &= 0 \\ (hu)_t + (hu^2 + \frac{1}{2}gh^2)_x + (huv)_y &= -ghb_x \\ (hv)_t + (huv)_x + (hv^2 + \frac{1}{2}gh^2)_y &= -ghb_y \end{aligned} \tag{1.16}$$

h is the height of the water (column), (u, v) is the velocity field, b is the bottom topography and g is the gravitational constant, which is set to $g = 9.807\frac{m}{s^2}$. I assume that b is nice enough, i.e. the function is at least continuous and has some smoothness. In my numerical experiments b is always piecewise smooth and continuous. As mentioned in the above derivation, the mass and the momentum are conserved.
In the rest of the thesis, I am only going to work with the one-dimensional shallow water equations

$$\begin{aligned} h_t + (hu)_x &= 0 \\ (hu)_t + (hu^2 + \frac{1}{2}gh^2)_x &= -ghb_x \ . \end{aligned} \tag{1.17}$$

This can be formulated in terms of conservative variable, flux and source vector; so one gets the following system of balance laws

$$U_t + f(U)_t = -s(x, U) \tag{1.18}$$

where $U = (h, hu)^T$, $f = (hu, hu^2 + \frac{1}{2}gh^2)^T$, $s = (0, ghb_x)^T$.
It is known that solutions of hyperbolic PDEs like (1.17) can develop discontinuous shocks after some (finite) time, even if the initial data is smooth [4]. So, one can no longer speak about classical solutions and therefore, the notion of weak solution is introduced.

Definition 1.1 (weak solution) *Consider an arbitrary conservation law*

$$U_t + f(U)_x = 0$$

where f is an arbitrary flux function and U some unknown conservative variable. $U(x,t)$ is called a weak solution of the above conservation law given some initial data $\widetilde{U}(x,0)$, if

$$\int_0^\infty \int_{-\infty}^\infty (U\varphi_t + f(U)\varphi_x \mathrm{d}x\mathrm{d}t + \int_{-\infty}^\infty U(x,0)\varphi(x,0)\mathrm{d}x = 0 \tag{1.19}$$

holds for all $\varphi \in C_c^1(\mathbb{R}_{\geqslant 0} \times \mathbb{R})$.

A solution of (1.18) considered in the weak sense is well-defined as long as the source term remains uniformly bounded, see [2]. Since weak solutions are not necessarily unique, one needs some admissibility conditions to get unique solutions. In the case of the shallow water equations, one imposes the so-called entropy condition.

Remark 1.2 *Clearly, all smooth solutions satisfy* (1.19).

1.2 Entropy Condition

Let $E(U)$ be a convex function and $H = H(U)$, $J = (J_1(x,U), J_2(x,U))^T$ functions such that the following relations hold

$$\partial_U H = \langle \partial_U E, \partial_U f(U) \rangle, \quad \partial_x J_1 = \langle \partial_U E, s \rangle \tag{1.20}$$

where $\langle \cdot, \cdot \rangle$ is the standard scalar product and f is the true flux of a balance or conservation law. $E(U)$ is called *entropy function*, $H(U)$ and J are the so-called *entropy flux functions* associated to the entropy function.
Smooth solutions satisfy

$$E(U)_t + (H(U) + J_1)_x = 0 \,.$$

Since one expects discontinuous shock solutions of the balance law, the above equation is modified to an inequality, because there is energy should dissipation at shock waves. So one gets

$$E(U)_t + (H(U) + J_1)_x \leqslant 0 \,.$$

General systems of balance laws do not necessarily have an entropy function but the shallow water equations do.
The entropy function of the shallow water equations is the total energy

$$E := \frac{1}{2}(hu^2 + gh(h+b))$$

and the associated entropy fluxes are then

$$H(U) := \frac{1}{2}(hu^3 + gh^2u), \quad J := ghb(u,0)^T$$

plugging in these expressions into above yields

$$\frac{1}{2}(hu^2 + gh(h+b))_t + (\frac{1}{2}(hu^3 + gh^2u) + ghbu)_x \leqslant 0 \ .$$

The entropy variable V is then given by

$$V = \partial_U E = (gh - \frac{u^2}{2}, u)^T \ . \tag{1.21}$$

Remark 1.3 *The entropy variables and the energy variables denote the same quantity for the shallow water equations, because their entropy function is the total energy.*

Remark 1.4 *In contrast to systems of balance laws, one can show that scalar conservation and balance laws have an infinite number of entropy functions. In fact, any strictly convex function is an entropy function for the scalar case.*

1.3 Numerical Methods

Since explicit formulas for the solutions of the balance laws or conservation laws are either very hard to compute or often not available, one uses numerical schemes to approximate these solutions. I will use the finite volume method in this thesis. The framework of this method can be found in [4]. I outline the general idea for the one-dimensional case.
Consider the conservation law and a uniform Cartesian grid $\{x_i\}_i \subset \mathbb{R}$. Generally, the mesh size does not need to be uniform, but for the sake of simplicity I assume uniformity, so $\Delta x = x_{j+1/2} - x_{j-1/2}\ \forall i$. In the rest of the thesis I always assume a uniform grid. Let $C_i = [x_{j-1/2}, x_{j+1/2}]$ denote a cell so that $\{C_i\}_i$ is a partition of the domain.
In the finite volume method, one considers the approximate average of U over the cell C_i

$$U_i(t) \approx \frac{1}{\Delta x}\int_{C_i} U(x,t)\mathrm{d}x$$

and one updates these averages to the next time level by solving a Riemann problem at each edge of the cells. This is the so-called Godunov method.
I consider the semi-discrete form of a standard finite volume method. For systems of conservation laws the general scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i(t) = -\frac{1}{\Delta x}(F_{i+1/2}(t) - F_{i-1/2}(t)) \tag{1.22}$$

where $F_{i+1/2}$ is a Godunov type numerical flux (e.g. Roe, Rusanov, etc.).
Since I want to compute balance laws, the above relation is intuitively modified to the following one

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i(t) = -\frac{1}{\Delta x}(F_{i+1/2}(t) - F_{i-1/2}(t)) - S_i \tag{1.23}$$

where S_i is a suitable discretization of the source term. I will introduce an explicit discretization in the next section.
The equation (1.22) respectively (1.23) can be considered as ODEs. So I can use standard ODE solving methods to perform the time stepping. I use the *second order strong stability preserving Runge Kutta method (SSP RK2)*

$$\begin{aligned} U_j^{\star} &= U_j^n + \Delta t \mathcal{L}(U_i^n) \\ U_j^{\star\star} &= U_j^{\star} + \Delta t \mathcal{L}(U_i^{\star}) \\ U_j^{n+1} &= \frac{1}{2}(U_j^n + U_j^{\star\star}) \,. \end{aligned} \tag{1.24}$$

$\mathcal{L}$ is defined as follows $\mathcal{L}(U_i) := -\frac{1}{\Delta x}(F_{i+1/2} - F_{i-1/2}) - S_i$ and $U_j^n = U_j(t^n)$ the approximation of U at time t^n. I assume uniform time steps.

Remark 1.5 *A strong stability preserving (Runge Kutta) method is a TVD multistep method. TVD means "total variation diminishing" and this property is preserved in such methods. Recall the TVD property (see [4])*

$$\mathrm{TV}(U^{n+1}) \leqslant \mathrm{TV}(U^n)$$

where

$$\mathrm{TV}(U) := \sum_{i=-\infty}^{\infty} |U_i - U_{i-1}| \,.$$

In the thesis, I will present an energy conservative scheme introduced in [1] and based on this, I will derive additional schemes (e.g. by adding numerical diffusion) to handle the difficulties occuring with the energy conservative scheme.

1.4 Steady States

Steady states are physically trivial, time independent states, which one wants to preserve during the numerical computation. These steady states are often used as test cases. A simple steady state for the shallow water equations is the so-called *lake at rest*. This state is given by

$$h + b \equiv constant, \quad u \equiv 0 \, . \tag{1.25}$$

Although there is in fact nothing to compute, in a numerical point of view, preserving this steady state can be difficult. Standard numerical schemes with a naive bottom discretization produce unphysical oscillations in the non-flat bottom region. If there are dry areas, i.e. $h(x) + b(x) = b(x)$ in some region, then schemes may completely fail. For the moment, I assume that there are no dry areas in the domain.
A more general steady state is the so-called *moving equilibrium state*, which is given by

$$m \equiv constant, \quad p \equiv constant \tag{1.26}$$

where

$$m := hu, \quad p := \frac{u^2}{2} + g(h + b) \tag{1.27}$$

m and p are called the *equilibrium variables*. The vector $P := (m, p)^T$ is simply called the *equilibrium variable*.
The *lake at rest* is just a special case of this steady state with $u \equiv 0$, because

$$u \equiv 0 \Rightarrow m \equiv 0 \text{ and } p = g(h + b) \equiv const. \Rightarrow h + b \equiv const.$$

A scheme which preserves a steady state (e.g. *lake at rest*) is called *well-balanced*.

Chapter 2

Well-Balanced Schemes

2.1 The Energy Conservative Scheme

I start with a basic energy conservative scheme. Its derivation is based on a result of *E. Tadmor*, which can be found in [1]. Before I formulate the theorem, I introduce the same standard notation used throughout the thesis.

$$[\![a]\!]_{i+1/2} := a_{i+1} - a_i, \quad \bar{a}_{i+1/2} := \frac{1}{2}(a_i + a_{i+1}) \tag{2.1}$$

which denote the jump and the arithmetic average where $a_i := a(x_i)$.

Theorem 2.1 (Tadmor [1]) *Consider the one-dimensional system of conservation law*

$$U_t + f(U)_x = 0 \tag{2.2}$$

with entropy function $E(U)$, entropy variable $V = \partial_U E$, entropy flux H and entropy potential $\Psi := \langle V, f \rangle - H$. Let $F_{i+1/2}$ be a consistent numerical flux that satisfies

$$\langle [\![V]\!]_{i+1/2}, F_{i+1/2} \rangle = [\![\Psi]\!]_{i+1/2} \tag{2.3}$$

Then, the scheme

$$\frac{\mathrm{d}}{\mathrm{d}t} U_i = -\frac{1}{\Delta x}(F_{i+1/2} - F_{i-1/2})$$

satisfies the discrete entropy identity

$$\frac{\mathrm{d}}{\mathrm{d}t} E(U_i(t)) = -\frac{1}{\Delta x}(\widetilde{H}_{i+1/2} - \widetilde{H}_{i-1/2})$$

where $\widetilde{H}_{i+1/2} := \langle \bar{V}_{i+1/2}, F_{i+1/2} \rangle - \bar{\Psi}_{i+1/2}$ is the numerical entropy flux.
In particular, the scheme is energy preserving.

For the one-dimensional shallow water equations, one has the following entropy variable and entropy potential

$$V = (gh - \frac{u^2}{2}, u)^T =: (V^{(1)}, V^{(2)})^T \,. \tag{2.4}$$

$$\Psi = \frac{1}{2} g u h^2 \tag{2.5}$$

So the corresponding jumps look as follows:
For V I have

$$\begin{aligned}
[\![V^{(1)}]\!]_{i+1/2} &= [\![gh - \frac{1}{2}u^2]\!]_{i+1/2} \\
&= g[\![h]\!]_{i+1/2} - \frac{1}{2}[\![u^2]\!]_{i+1/2} \\
&= g[\![h]\!]_{i+1/2} - \frac{1}{2}(u_{i+1}^2 - u_i^2) \\
&= g[\![h]\!]_{i+1/2} - \frac{1}{2}(u_{i+1} + u_i)(u_{i+1} - u_i) \\
&= g[\![h]\!]_{i+1/2} - \bar{u}_{i+1/2}[\![u]\!]_{i+1/2}
\end{aligned}$$

$$[\![V^{(2)}]\!]_{i+1/2} = [\![u]\!]_{i+1/2}$$

and for the entropy potential

$$\begin{aligned}
[\![\Psi]\!]_{i+1/2} &= \frac{1}{2} g [\![u h^2]\!]_{i+1/2} \\
&\overset{(*)}{=} \frac{1}{2} g (\bar{u}_{i+1/2}[\![h^2]\!]_{i+1/2} + \bar{h^2}_{i+1/2}[\![u]\!]_{i+1/2}) \\
&= \frac{g}{2}\bar{u}_{i+1/2}(h_{i+1} + h_i)(h_{i+1} - h_i) + \frac{g}{2}\bar{h^2}_{i+1/2}[\![u]\!]_{i+1/2} \\
&= g\bar{u}_{i+1/2}\bar{h}_{i+1/2}[\![h]\!]_{i+1/2} + \frac{g}{2}\bar{h^2}_{i+1/2}[\![u]\!]_{i+1/2} \,.
\end{aligned}$$

In $(*)$, I used the following identity

$$[\![ab]\!]_{i+1/2} = \bar{a}_{i+1/2}[\![b]\!]_{i+1/2} + \bar{b}_{i+1/2}[\![a]\!]_{i+1/2} \,.$$

The identity above is a straightforward calculation.
The numerical flux $F_{i+1/2}$ can be written as $F_{i+1/2} = (F^{(1)}_{i+1/2}, F^{(2)}_{i+1/2})^T$. The goal is to satisfy the assumption of *Tadmor's* theorem, i.e. (2.3) , so that the resulting flux is energy conservative.

$$\begin{aligned}
\langle [\![V]\!]_{i+1/2}, F_{i+1/2} \rangle &= [\![V^{(1)}]\!]_{i+1/2} F^{(1)}_{i+1/2} + [\![V^{(2)}]\!]_{i+1/2} F^{(2)}_{i+1/2} \\
&= F^{(1)}_{i+1/2} \cdot (g[\![h]\!]_{i+1/2} - \bar{u}_{i+1/2}[\![u]\!]_{i+1/2}) + F^{(2)}_{i+1/2}[\![u]\!]_{i+1/2} \\
&\overset{!}{=} g\bar{u}_{i+1/2}\bar{h}_{i+1/2}[\![h]\!]_{i+1/2} + \frac{g}{2}\bar{h^2}_{i+1/2}[\![u]\!]_{i+1/2} \\
&= [\![\Psi]\!]_{i+1/2}
\end{aligned}$$

By rearranging the above terms, the following should hold

$$[\![h]\!]_{i+1/2} g F^{(1)}_{i+1/2} + [\![u]\!]_{i+1/2}(F^{(2)}_{i+1/2} - \bar{u}_{i+1/2}F^{(1)}_{i+1/2}) \overset{!}{=} g\bar{u}_{i+1/2}\bar{h}_{i+1/2}[\![h]\!]_{i+1/2} + \frac{g}{2}\bar{h^2}_{i+1/2}[\![u]\!]_{i+1/2}$$

and by equating the coefficients in the jumps of h and u, I get the following system of equations fot the two unknowns $F^{(1)}_{i+1/2}$, $F^{(2)}_{i+1/2}$

$$\begin{aligned} gF^{(1)}_{i+1/2} &= g\bar{u}_{i+1/2}\bar{h}_{i+1/2} \\ F^{(2)}_{i+1/2} - \bar{u}_{i+1/2}F^{(1)}_{i+1/2} &= \frac{g}{2}\overline{h^2}_{i+1/2} \,. \end{aligned}$$

This system can easily be solved and I get the components of the energy conservative flux

$$\begin{aligned} F^{(1)}_{i+1/2} &= \bar{u}_{i+1/2}\bar{h}_{i+1/2} \\ F^{(2)}_{i+1/2} &= \frac{g}{2}\overline{h^2}_{i+1/2} + (\bar{u}_{i+1/2})^2\bar{h}_{i+1/2} \,. \end{aligned}$$

Based on this calculation, I can define the *energy conservative (EC)* flux as:

$$F^{EC}(U_i, U_{i+1}) = F^{EC}_{i+1/2} = \begin{pmatrix} \bar{u}_{i+1/2}\bar{h}_{i+1/2} \\ \frac{g}{2}\overline{h^2}_{i+1/2} + (\bar{u}_{i+1/2})^2\bar{h}_{i+1/2} \end{pmatrix} \tag{2.6}$$

This flux is clearly consistent.
Since I do not consider the conservation law but the balance law, I have to implement the source term. The corresponding entropy variable V is then slightly different than before, namely $V = (g(h+b) - \frac{u^2}{2}, u)^T$. The energy conservation statement from *Tadmor's* theorem then has a slightly different assumption. It is stated in the next lemma, which can be found in [2].

Lemma 2.2 *A numerical flux $F_{i+1/2}$ is energy conservative if*

$$\langle [\![V]\!]_{i+1/2}, F_{i+1/2} \rangle = [\![\Psi]\!]_{i+1/2} + g\bar{u}_{i+1/2}\bar{h}_{i+1/2}[\![b]\!]_{i+1/2} \,. \tag{2.7}$$

The corresponding finite volume scheme then satisfies the energy conservation statement

$$\frac{\mathrm{d}}{\mathrm{d}t}E_i = -\frac{1}{\Delta x}(\widetilde{H}_{i+1/2} - \widetilde{H}_{i-1/2})$$

where the numerical energy flux is given by

$$\widetilde{H}_{i+1/2} = \langle \bar{V}_{i+1/2}, F_{i+1/2} \rangle - \bar{\Psi}_{i+1/2} - \frac{g}{4}\bar{h}_{i+1/2}[\![u]\!]_{i+1/2}[\![b]\!]_{i+1/2} \,.$$

In particular, the total energy is preserved: $\sum_i E_i(t)\Delta x = \sum_i E_i(0)\Delta x$, $\forall t > 0$

A proof can also be found in [2]. It is a straightforward calculation.

Remark 2.3 *One can easily see that F^{EC}, $V = (g(h+b) - \frac{u^2}{2}, u)^T$, Ψ satisfy the identity* (2.7)*, since all the terms are the same as before except the jump of the bottom topography. This jump is multiplied by g and $F^{(1)}_{i+1/2} = \bar{u}_{i+1/2}\bar{h}_{i+1/2}$ which is nothing else than the term added to $[\![\Psi]\!]_{i+1/2}$ on the right-hand side of* (2.7)*. Thus, the total energy is preserved with the EC flux.*

I now introduce the source discretization, which is chosen in such a way that the steady state is conserved. It differs slightly from a naive discretization.

$$S_i := \begin{pmatrix} 0 \\ \frac{g}{2\Delta x}(\bar{h}_{i+1/2}[\![b]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![b]\!]_{i-1/2}) \end{pmatrix} \tag{2.8}$$

So, the following finite volume scheme is called the energy conservative scheme, or short *EC scheme*

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i = -\frac{1}{\Delta x}(F^{EC}_{i+1/2} - F^{EC}_{i-1/2}) - S_i \ . \tag{2.9}$$

or if one writes the scheme componentwise

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}h_i &= -\frac{1}{\Delta x}(\bar{h}_{i+1/2}\bar{u}_{i+1/2} - \bar{h}_{i-1/2}\bar{u}_{i-1/2}) && (2.10) \\ \frac{\mathrm{d}}{\mathrm{d}t}(h_i u_i) &= -\frac{1}{\Delta x}(\bar{h}_{i+1/2}(\bar{u}^2_{i+1/2} + \frac{g}{2}\bar{h^2}_{i+1/2} - \bar{h}_{i-1/2}(\bar{u}^2_{i-1/2} + \frac{g}{2}\bar{h^2}_{i-1/2}) \\ &\quad - \frac{g}{2\Delta x}(\bar{h}_{i+1/2}[\![b]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![b]\!]_{i-1/2}) \ . && (2.11) \end{aligned}$$

Some main properties of the *EC* scheme (2.9) are summerized in the following theorem

Theorem 2.4 ([2]) *The EC scheme* (2.9) *satisfies the following properties.*

1. *It is second order accurate.*
2. *It is energy conservative.*
3. *It is well-balanced for the lake at rest steady state, i.e. for given initial data*

$$u_i \equiv 0, \quad h_i + b_i \equiv constant \quad \forall i$$

then

$$\frac{\mathrm{d}}{\mathrm{d}t}h_i \equiv 0, \quad \frac{\mathrm{d}}{\mathrm{d}t}(h_i u_i) \equiv 0 \quad \forall i$$

Proof

1. This follows by truncation error analysis.
2. This property follows by lemma 2.2 and remark 2.3.
3. This is the most interesting part. For the first component (2.10), I have

$$\begin{aligned} u_i \equiv 0 \ \forall i \quad &\Rightarrow -\frac{1}{\Delta x}(\bar{h}_{i+1/2}\bar{u}_{i+1/2} - \bar{h}_{i-1/2}\bar{u}_{i-1/2}) \equiv 0 \\ &\Leftrightarrow \frac{\mathrm{d}}{\mathrm{d}t}h_i \equiv 0 \ . \end{aligned}$$

For the second component, I use the following identity

$$\bar{a^2}_{i+1/2} - \bar{a^2}_{i-1/2} = \bar{a}_{i+1/2}[\![a]\!]_{i+1/2} + \bar{a}_{i-1/2}[\![a]\!]_{i-1/2} \tag{2.12}$$

and the easy fact that

$$h + b \equiv constant \Rightarrow [\![h+b]\!]_{i+1/2} \equiv 0 \ \forall i \ .$$

Since $u_i \equiv 0 \ \forall i$ the second equation (2.11) becomes

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(h_i u_i) &= -\frac{g}{2\Delta x}(\bar{h^2}_{i+1/2} - \bar{h^2}_{i-1/2} + \bar{h}_{i+1/2}[\![b]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![b]\!]_{i-1/2}) \\
&\overset{(2.12)}{=} -\frac{g}{2\Delta x}(\bar{h}_{i+1/2}[\![h]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![h]\!]_{i-1/2} + \bar{h}_{i+1/2}[\![b]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![b]\!]_{i-1/2}) \\
&= -\frac{g}{2\Delta x}(\bar{h}_{i+1/2}\underbrace{[\![h+b]\!]_{i+1/2}}_{\equiv 0} + \bar{h}_{i-1/2}\underbrace{[\![h+b]\!]_{i-1/2}}_{\equiv 0}) \\
&\equiv 0 \ .
\end{aligned}$$

□

So, I have an energy conservative scheme which I now test in two numerical experiments.

2.2 Numerical Experiments for EC Scheme

If not stated differently, throughout the thesis I use transmissive (open Neumann) boundary condition, $CFL = 0.4$ and number of grid points $N = 200$.

2.2.1 Lake at rest

Theorem 2.4 ensures that this steady state should be preserved in theory. The setting of the experiment is the following:
Given the domain $\Omega = [0, 20]$ and a bottom topography

$$b(x) = \begin{cases} \frac{4-(x-10)^2}{20} & \text{if } x \in [8, 12] \\ 0 & \text{else .} \end{cases} \tag{2.13}$$

Initial data is given by

$$u_i \equiv 0, \quad h_i + b_i \equiv 1 \qquad \forall i \ .$$

I set the endtime $T = 100$, since the lake at rest should be preserved at large time scales. The EC scheme preserves the lake at rest, even at large time scale. One can observe that in the beginning of the time stepping there are few very small oscillations around the non-flat bottom topography region. These tiny oscillations become more after a short time period but they are

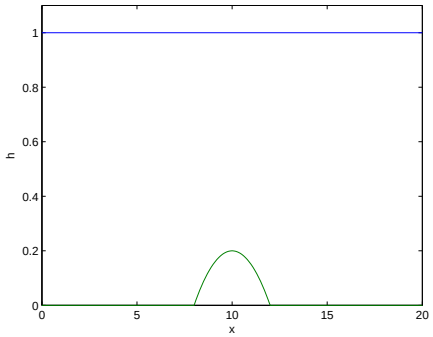

Figure 2.1: Lake at rest, $T = 100$

very small (single precision error is zero). So, the lake at rest is preserved (up to some very small error). Probably, these oscillations are an effect induced by the time stepping routine, in my case the (*SSP RK2*). The scheme works as long as the domain stays flooded. If I introduce some dry areas in the initial data, the scheme fails to compute the lake at rest. The approximate solution starts oscillating, first at small amplitude, then it grows until it explodes. This happens after some very short time. In chapter 3, I am going to do further analysis about dry areas.
With a little "trick", I can achieve the preservation of the lake at rest with initial dry area: If I weaken the condition that h, the height of the water, is always non-negative. So, h can also be negative. I change the bottom topography (2.13) so that I have a dry area with the same initial data as above

$$\widetilde{b}(x) := \begin{cases} \frac{4-(x-10)^2}{3.5} & \text{if } x \in [8,12] \\ 0 & \text{else .} \end{cases}$$

So, I have $\widetilde{b} > 1$ at $x = 10$. Together with $h_i + \widetilde{b}_i \equiv 1$, I get that $h < 0$ at $x = 10$. If I implement it this way allowing negative height, the lake at rest is preserved even with dry areas. This approach only works if the lake stays at rest. Computations of small perturbation completely fail due to the negative height. In addition, $h < 0$ physically makes no sense, so I do not want to use this approach any further in this section.

2.2.2 Dam break

Now, I consider a simple dam break problem with flat bottom topography in the domain $\Omega = [-1, 1]$. I have the following initial data

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

The derivation of the exact solution can be found in [4]. The exact solution consists of a left-going rarefaction wave and a right-going shock wave. As

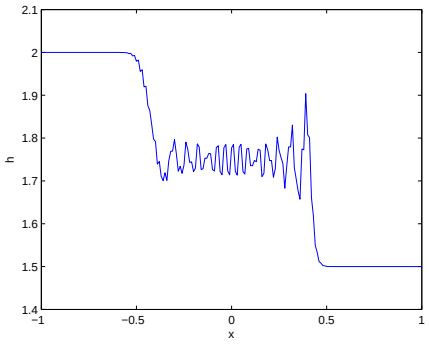

Figure 2.2: Dam break, $T = 0.15$, $N = 200$.

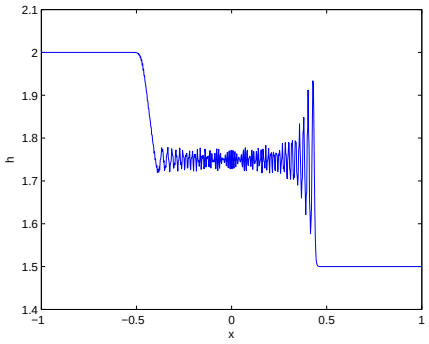

Figure 2.3: Dam break, $T = 0.15$, $N = 800$.

one can observe in figure 2.2 and 2.3, there are large post-shock oscillations. If the number of grid cells is raised (e.g. $N = 800, \ldots$), the number of oscillating parts in the approximate solution increases, as one can see in figure 2.3. But the oscillations seem to be bounded. This effect is due to the energy conservation property of the EC scheme, since one expects that the energy should be dissipated at shocks. So, the EC scheme preserves this energy in form of these unphysical post-shock oscillations.
There are several ways to encounter this difficulty of the EC scheme. E.g. introducing eddy viscosity would be a possibility. But the results are not very satisfying, because the eddy viscosity is a small quantity and thus, only has a limited effect on these oscillations. With eddy viscosity, the number of grid points needs to be very high to get rid of these oscillations. For further readings, I refer to [1]. Another possibility is to introduce numerical diffusion. This is done in the next sections.

2.3 Energy Stable Schemes

2.3.1 First Order Energy Stable Scheme

The first step of designing an energy stable scheme is the computation of the *Roe* diffusion matrix $Q^{Roe}_{i+1/2}$. For this purpose, the eigenvalues and eigenvectors of $f'(U)$ need to be computed, where $f'(U)$ is the Jacobian matrix of the true flux of the shallow water equations $f(U) = (hu, hu^2 + \frac{1}{2}gh^2)^T$. $f'(U)$ is given by

$$f'(U) = \begin{pmatrix} 0 & 1 \\ -u^2 + gh & 2u \end{pmatrix}.$$

The eigenvalues $\lambda_\pm$ and eigenvectors r_i look as follows

$$\lambda_\pm = u \pm \sqrt{gh}, \quad r_{1,2} = \begin{pmatrix} 1 \\ \lambda_\pm \end{pmatrix} .$$

The *Roe* diffusion matrix is then given by

$$Q^{Roe}_{i+1/2} = \widetilde{R}_{i+1/2} |\widetilde{\Lambda}_{i+1/2}| \widetilde{R}^{-1}_{i+1/2} \tag{2.14}$$

where

$$\widetilde{R}_{i+1/2} := \frac{1}{\sqrt{2g}} \begin{pmatrix} 1 & 1 \\ \lambda_- & \lambda_+ \end{pmatrix}, \qquad |\widetilde{\Lambda}_{i+1/2}| = \begin{pmatrix} |\lambda_-| & 0 \\ 0 & |\lambda_+| \end{pmatrix} .$$

The *Roe* flux is defined as follows

$$F^{Roe}_{i+1/2} := F^{Roe}(U_i, U_{i+1}) := \frac{1}{2}(F(U_i) + F(U_{i+1})) - Q^{Roe}_{i+1/2} [\![U]\!]_{i+1/2}$$

where one puts the Roe average $\widehat{h}$, $\widehat{u}$ into $\lambda_\pm$.These Roe averages are given by

$$\widehat{h}_{i+1/2} := \frac{h_i + h_{i+1}}{2}, \quad \widehat{u}_{i+1/2} := \frac{\sqrt{h_i} u_i + \sqrt{h_{i+1}} u_{i+1}}{\sqrt{h_i} + \sqrt{h_{i+1}}} .$$

As in [1], I modify the *Roe* flux by replacing the central part of the flux by the EC flux and the diffusion matrix by a slightly different one, since the Roe diffusion matrix may not preserve the lake at rest steady state. This new diffusion matrix should preserve the lake at rest steady state.
I need some preliminary statements before designing the diffusion matrix. The following lemma one can be found in [1]. But first, I define the $R_{i+1/2}$ and $|\Lambda_{i+1/2}|$ matrix in terms of arithmetic averages:

$$R_{i+1/2} := \frac{1}{\sqrt{2g}} \begin{pmatrix} 1 & 1 \\ \bar{\lambda}_- & \bar{\lambda}_+ \end{pmatrix}, \qquad |\Lambda_{i+1/2}| := \begin{pmatrix} |\bar{\lambda}_-| & 0 \\ 0 & |\bar{\lambda}_+| \end{pmatrix} \tag{2.15}$$

and

$$\bar{\lambda}_\pm := \bar{u}_{i+1/2} \pm \sqrt{g \bar{h}_{i+1/2}} . \tag{2.16}$$

Lemma 2.5 ([1]) *Consider the conservative form of shallow water equations, i.e.* (1.17) *with $b \equiv$ constant, the total energy $E(U) = \frac{1}{2}(gh^2 + hu^2)$ and the energy variable $\partial_U E = V = (gh - \frac{u^2}{2}, u)^T$. Let U_i, U_{i+1} be given. The following properties hold:*

1.

$$[\![U]\!]_{i+1/2} = (U_V)_{i+1/2} [\![V]\!]_{i+1/2} \tag{2.17}$$

where

$$(U_V)_{i+1/2} = \frac{1}{g}\begin{pmatrix} 1 & \bar{u}_{i+1/2} \\ \bar{u}_{i+1/2} & (\bar{u}_{i+1/2})^2 + g\bar{h}_{i+1/2} \end{pmatrix}$$

is the change of variable matrix in terms of averages.

2. *Let $R_{i+1/2}$ be as above, then*

$$R_{i+1/2}R_{i+1/2}^T = (U_V)_{i+1/2}\,. \tag{2.18}$$

The proof are just straightforward calculations and can be found in [1].
So by this lemma, one gets the following relation

$$\begin{aligned} Q^{Roe}_{i+1/2}[\![U]\!]_{i+1/2} &= R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}[\![U]\!]_{i+1/2} \\ &\overset{(2.17)}{=} R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(U_V)_{i+1/2}[\![V]\!]_{i+1/2} \\ &\overset{(2.18)}{=} R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}R_{i+1/2}R^T_{i+1/2}[\![V]\!]_{i+1/2} \\ &= R_{i+1/2}|\Lambda_{i+1/2}|R^T_{i+1/2}[\![V]\!]_{i+1/2}\,. \end{aligned}$$

I now define the first order energy stable flux (*ES1*) based on the calculation done above

$$F^{ES1}_{i+1/2} := F^{EC}_{i+1/2} - \frac{1}{2}R_{i+1/2}|\Lambda_{i+1/2}|R^T_{i+1/2}[\![V]\!]_{i+1/2}\,. \tag{2.19}$$

Notation: $D^{ES1}_{i+1/2} := R_{i+1/2}|\Lambda_{i+1/2}|R^T_{i+1/2}$.
This *ES1* flux is consistent, because F^{EC} is consistent and the jump in the energy variable vanishes.

$$\begin{aligned} F^{ES1}(U,U) &= F^{EC}(U,U) - \frac{1}{2}D^{ES1}_{i+1/2}\begin{pmatrix} gh - \frac{u^2}{2} - gh - \frac{u^2}{2} \\ u - u \end{pmatrix} \\ &= f(U) \end{aligned}$$

where f is the true flux of the balance law (1.18).
The resulting *ES1* scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i = -\frac{1}{\Delta x}(F^{ES1}_{i+1/2} - F^{ES1}_{i-1/2}) - S_i\,. \tag{2.20}$$

One can show that the *ES1* scheme (2.20) is first order accurate and satisfies a discrete energy identity which implies energy dissipation over time. For further reading see [1]. Another property is stated in the next lemma.

Lemma 2.6 ([2]) *The* ES1 *scheme preserves the lake at rest steady state. Hence, it is well-balanced.*

Proof Given the lake at rest initial data.

$$u_i \equiv 0, \qquad h_i + b_i \equiv constant \qquad \forall i$$

It is enough to show that the jump $[\![V]\!]_{i+1/2} \equiv 0 \quad \forall i$, since F^{EC} is well-balanced.

$$\begin{aligned}
[\![V^{(1)}]\!]_{i+1/2} &= [\![g(h+b) - \frac{u^2}{2}]\!]_{i+1/2} \\
&= g[\![h+b]\!]_{i+1/2} - \frac{1}{2}[\![u^2]\!]_{i+1/2} \\
&= 0
\end{aligned}$$

$$[\![V^{(2)}]\!]_{i+1/2} = [\![u]\!]_{i+1/2} = 0$$ □

Before testing this scheme in some numerical experiments I want to introduce two second order extensions of the ES1 scheme.

2.3.2 Second Order Energy Stable Scheme

It is standard procedure to replace the piecewise constant cell average by a piecewise linear non-oscillatory reconstruction to get second order accuracy. In a first approach, the reconstruction step is done in the energy variable V_i. I define the piecewise linear reconstruction by

$$\widetilde{V}_i(x) := V_i + V_i' \cdot (x - x_i), \qquad x \in [x_{i-1/2}, x_{i+1/2}]$$

where

$$\begin{aligned}
V_i' &:= \frac{1}{\Delta x}\text{minmod}(V_{i+1} - V_i, V_i - V_{i-1}) \\
\text{minmod}(a,b) &:= \begin{cases} \text{sgn}(a)\min\{|a|,|b|\} & \text{if sgn}(a) = \text{sgn}(b) \\ 0 & \text{else .} \end{cases}
\end{aligned}$$

The reconstructed values at the cell interface edge are given by

$$\begin{aligned} V_i^r := \widetilde{V}_i(x_{i+1/2}) &= V_i + V_i' \cdot (x_{i+1/2} - x_i) \\ &= V_i + \frac{\Delta x}{2} V_i' \\ &= V_i + \frac{1}{2}\text{minmod}(V_{i+1} - V_i, V_i - V_{i-1}) \end{aligned}$$

$$\begin{aligned} V_{i+1}^l := \widetilde{V}_{i+1}(x_{i+1/2}) &= V_{i+1} + V_{i+1}' \cdot (x_{i+1/2} - x_{i+1}) \\ &= V_{i+1} - \frac{\Delta x}{2} V_{i+1}' \\ &= V_{i+1} - \frac{1}{2}\text{minmod}(V_{i+2} - V_{i+1}, V_{i+1} - V_i) \end{aligned}$$

where the superscript in $V_i^{r,l}$ denotes the right respective left edge value. I compute the corresponding reconstructed conservative variable U_{i+1}^l,U_i^r. I.e. for

$$U_{i+1}^l = \begin{pmatrix} h_{i+1}^l \\ h_{i+1}^l u_{i+1}^l \end{pmatrix}, \qquad U_i^r = \begin{pmatrix} h_i^r \\ h_i^r u_i^r \end{pmatrix} .$$

The diffusion matrix is then given by

$$D_{i+1/2}^{ES2} := R_{i+1/2} |\Lambda_{i+1/2}| R_{i+1/2}^T \tag{2.21}$$

where $R_{i+1/2}$ and $|\Lambda_{i+1/2}|$ are defined as above and $\lambda_\pm$ is adapted to the reconstructed values

$$\lambda_\pm := \frac{1}{2}(u_{i+1}^l + u_i^r) \pm \sqrt{\frac{g}{2}(h_{i+1}^l + h_i^r)} .$$

The second order accurate energy stable flux, short *ES2* flux, is then given by

$$F_{i+1/2}^{ES2} := F_{i+1/2}^{EC} - \frac{1}{2} D_{i+1/2}^{ES2} (V_{i+1}^l - V_i^r) . \tag{2.22}$$

It is easy to see that the flux is consistent. The resulting scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t} U_i = -\frac{1}{\Delta x}(F_{i+1/2}^{ES2} - F_{i-1/2}^{ES2}) - S_i . \tag{2.23}$$

The *ES2* scheme preserve the lake at rest and is second order accurate.
A second approach to a second order extension of the ES1 scheme is the following (cf. [1]).
Instead of reconstructing the energy variable V_i, I directly reconstruct the conservative variable U_i. Define $\widetilde{U}$ as the piecewise linear reconstruction.

$$\widetilde{U}_i(x) := U_i + U_i' \cdot (x - x_i), \qquad x \in [x_{i-1/2}, x_{i+1/2}] \tag{2.24}$$

where U_i' is given by

$$U_i' := \frac{1}{\Delta x}\text{minmod}(U_{i+1} - U_i, \frac{1}{2}(U_{i+1} - U_{i-1}), U_i - U_{i-1})$$

where in this case the minmod function is defined as follows

$$\text{minmod}(a,b,c) := \begin{cases} \text{sgn}(a)\min\{|a|,|b|,|c|\} & \text{if } \text{sgn}(a) = \text{sgn}(b) = \text{sgn}(c) \\ 0 & \text{else} \end{cases}$$

Define the edge values U_{i+1}^l, U_i^r:

$$\begin{aligned} U_i^r &:= \widetilde{U}_i(x_{i+1/2}) \\ &= U_i + U_i' \cdot (x_{i+1/2} - x_i) \\ &= U_i + \frac{1}{2}\text{minmod}(U_{i+1} - U_i, \frac{1}{2}(U_{i+1} - U_{i-1}), U_i - U_{i-1}) \end{aligned}$$

$$\begin{aligned} U_{i+1}^l &:= \widetilde{U}_{i+1}(x_{i+1/2}) \\ &= U_{i+1} - \frac{1}{2}\text{minmod}(U_{i+2} - U_{i+1}, \frac{1}{2}(U_{i+2} - U_i), U_{i+1} - U_i)\ . \end{aligned}$$

The diffusion matrix is again defined in terms of the reconstructed values U_i^r, U_{i+1}^l; I denote it by $D_{i+1/2}^{ES2,2}$. Then, this second order energy stable flux (ES2,2) is simply given by

$$F_{i+1/2}^{ES2,2} := F_{i+1/2}^{EC} - \frac{1}{2}D_{i+1/2}^{ES2,2}(V_{i+1}^l - V_i^r) \tag{2.25}$$

where V_{i+1}^l, V_i^r are the energy variables of U_i^r, U_{i+1}^l. The flux is obviously consistent. The numerical scheme is then given by

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i = -\frac{1}{\Delta x}(F_{i+1/2}^{ES2,2} - F_{i-1/2}^{ES2,2}) - S_i\ . \tag{2.26}$$

As above, the *ES2,2* scheme is second order accurate.

Remark 2.7 *Although the* ES1 *scheme is energy stable, its second order extensions may not be. The question, whether the second order schemes are energy conservative is not answered yet.*

2.4 Numerical Experiments for Energy Stable Schemes

2.4.1 Lake at rest

As for the EC scheme, I test if the three energy stable schemes preserve the *lake at rest* steady state. The setting is the same as in section 2.2.1. As expected, all three schemes preserve the state. In figure 2.4 or better in figure 2.6 one can observe that the ES2,2 scheme slightly oscillates around the non-flat bottom region. The ES1 and ES2 schemes preserve the lake at rest far better than the ES2,2, as one can see in figure 2.7 and 2.6.

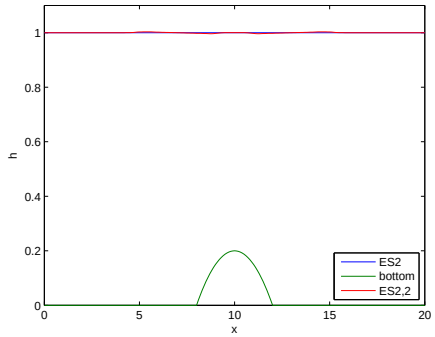

Figure 2.4: ES2 schemes, $T = 1$

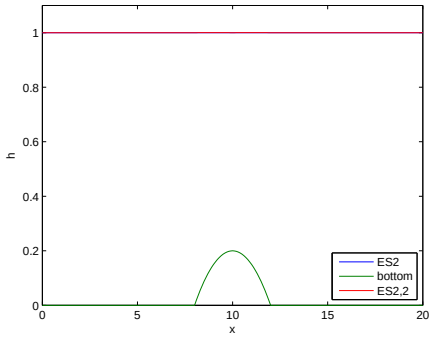

Figure 2.5: ES2 schemes, $T = 50$

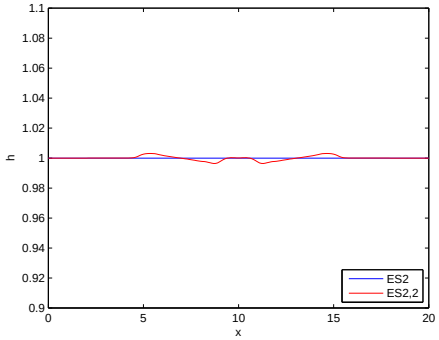

Figure 2.6: ES2 schemes, $T = 1$. Zoomed in.

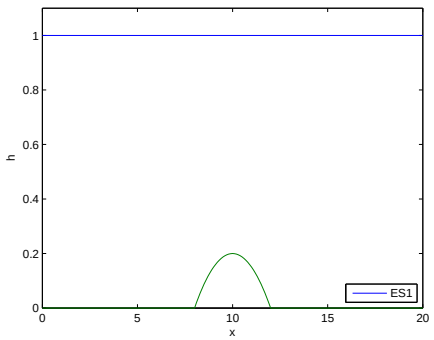

Figure 2.7: ES1 scheme, T = 100

2.4.2 Dam break

The energy stable schemes should resolve the *dam break problem* more accurate than the EC scheme. In particular, the approximate solution should be physically meaningful. The setting is the same as in 2.2.2. Recall the initial data

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

As expected, the *dam break problem* is solved more accurate by the ES schemes. One can also observe in figure 2.8 that the second order extensions ES2 and ES2,2 are less diffusive than the ES1 scheme. They capture the shock wave more sharply. There are no post-shock oscillations as in the EC scheme. In figure 2.9, I test whether the two different second order extensions differ much. The approximate solution of ES2 and ES2,2 are almost the same except at the beginning of the shock wave and rarefaction wave. There, the ES2 scheme produces small artifacts.

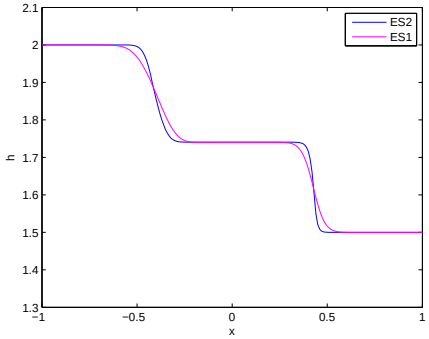

Figure 2.8: The ES2 vs ES1 scheme, $T = 0.1$

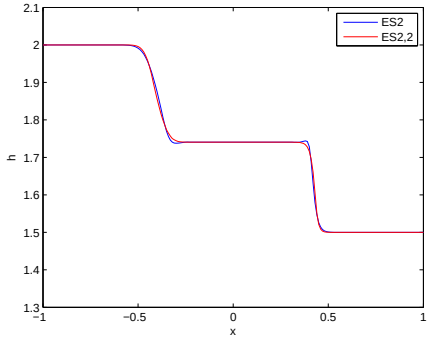

Figure 2.9: The ES2 vs ES2,2 scheme, $T = 0.1$

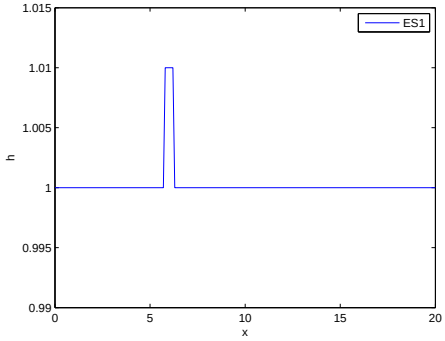

Figure 2.10: This is the initial height h.

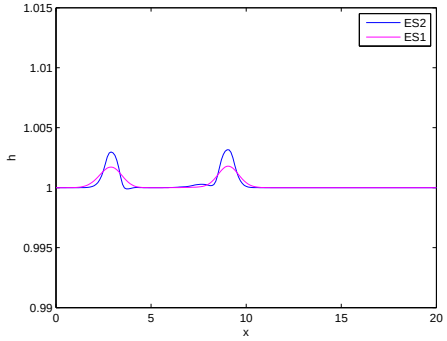

Figure 2.11: ES2 vs. ES1 scheme, $T = 1$.

2.4.3 Perturbated lake at rest

Consider the *lake at rest* with some non-flat bottom topography. But this time, there is some small initial perturbation given as small height differences in a small area.
The idea of this test case is to check if this initial perturbation remains small after some time and if the pertubation does not cause oscillations in an unphysical manner.
Let the domain and the bottom be as in 2.2.1. The initial data is given by:

$$h(x,0) = \begin{cases} 1.01 - b(x) & \text{if } |x-6| < \frac{1}{4} \\ 1 - b(x) & \text{else} \end{cases}, \qquad u(x,0) \equiv 0\ .$$

The solution should be a left-going and a right-going wave. Figure 2.10 shows the initial height of the water (note that the figure is zoomed in so the bottom topography is not visible). Figure 2.11 shows the water height at $T = 1$. The approximate solution consists of a left- and right-going wave as the true solution. The second order extension ES2 performs better in computing the perturbation than the ES1 scheme, since the waves are less

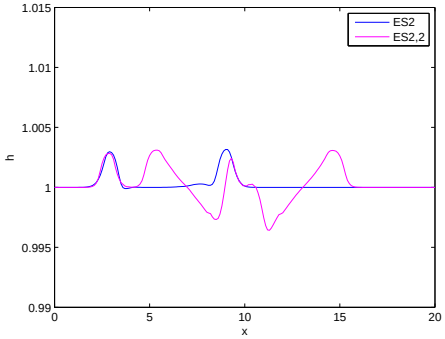

Figure 2.12: ES2 vs. ES2,2 scheme, $T = 1$.

diffusive. In contrast to the ES2 scheme, the ES2,2 scheme fails to capture the small perturbation, as one can observe in figure 2.12. The approximate solution of ES2,2 captures the left-going wave quite well but is nowhere near the true solution for the right-going wave. Hence, this scheme is not useful for the shallow water equations with bottom topography. If the bottom topography is flat, the scheme works fine, as we have seen in section 2.4.2.

2.5 Well-Balanced Schemes For The General Steady State

Up to here, I have only considered the lake at rest steady state. But how to handle more general steady states? Recall from section 1.4 the *moving equilibrium* steady state:

$$m \equiv constant\ , \qquad p \equiv constant \tag{2.27}$$

where

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

Recall from section 1.4 $P := (m, p)^T$. m, p are called the *equilibrium variables*. I define a discrete version of the classical momentum m, the *staggered momentum* $M_{i+1/2}$

$$M_{i+1/2} := \bar{h}_{i+1/2}\bar{u}_{i+1/2}\ . \tag{2.28}$$

The EC scheme preserves the discrete version of (2.27), i.e. instead of the classical momentum, the staggered momentum is constant in time. This property is formulated in the next lemma:

Lemma 2.8 ([2]) *Let $M_{i+1/2}$ be the staggered momentum and $p_i = \frac{u_i^2}{2} + g(h_i + b_i)$. Then, the EC scheme preserves the state*

$$M_{i+1/2} \equiv C_1\ , \qquad p_i \equiv C_2 \qquad \forall i \tag{2.29}$$

for some constants C_1, C_2.

Proof

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}h_i &= -\frac{1}{\Delta x}(\bar{h}_{i+1/2}\bar{u}_{i+1/2} - \bar{h}_{i-1/2}\bar{u}_{i-1/2}) \\ &= -\frac{1}{\Delta x}(M_{i+1/2} - M_{i-1/2}) \\ &= 0\ .\end{aligned}$$

For the second equation, I use the following unhandy but useful identity

$$\begin{aligned}&\frac{1}{2}(\bar{h}_{i+1/2}[\![p]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![p]\!]_{i-1/2}) + u_i(M_{i+1/2} - M_{i-1/2}) \\ &= \bar{h}_{i+1/2}(\bar{u}_{i+1/2})^2 + \frac{g}{2}\bar{h^2}_{i+1/2} - \bar{h}_{i-1/2}(\bar{u}_{i-1/2})^2 + \frac{g}{2}\bar{h^2}_{i-1/2} + \frac{g}{2}(\bar{h}_{i+1/2}[\![b]\!]_{i+1/2} - \bar{h}_{i-1/2}[\![b]\!]_{i-1/2}).\end{aligned} \tag{2.30}$$

This identity implies that the second equation (2.11) in the EC scheme can be written as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}h_iu_i = -\frac{1}{\Delta x}(\frac{1}{2}(\bar{h}_{i+1/2}[\![p]\!]_{i+1/2} + \bar{h}_{i-1/2}[\![p]\!]_{i-1/2}) + u_i(M_{i+1/2} - M_{i-1/2}))\ .$$

But $[\![p]\!]_{i+1/2} \equiv 0$ since $p_i \equiv constant$, and $M_{i+1/2} - M_{i-1/2} = 0 \quad \forall i$. This implies

$$\frac{\mathrm{d}}{\mathrm{d}t}h_iu_i \equiv 0\ .$$

Thus, the state (2.29) is preserved.
The above identity (2.30) can be proved by a straightforward calculation. □

As in the previous section, I need to design a diffusion matrix, since the EC scheme produces large oscillations at shockwaves. The energy stable schemes do not necessarily preserve the *moving equilibrium* steady state, so a new diffusion matrix is needed.

2.5.1 Well-Balanced First Order Scheme

Instead of acting on the jump of the energy variable V_i like in the ES schemes, I want to design a diffusion matrix that acts on the *equilibrium variable* $P = (m, p)^T$, respectively on its jump.
Therefore, consider the following change of variable matrix stated in the lemma below.

Lemma 2.9 *Given the conservative variables U_i, U_{i+1} and their corresponding equilibrium variables P_i, P_{i+1} where P_i is definded as*

$$P_i := (m_i, p_i)^T, \quad m_i = h_iu_i, \quad p_i = \frac{u_i^2}{2} + g(h_i + b_i)\ .$$

Then the following identity holds

$$[\![U]\!]_{i+1/2} = (U_P)_{i+1/2}[\![P]\!]_{i+1/2} \tag{2.31}$$

where

$$(U_P)_{i+1/2} := \begin{pmatrix} \frac{\bar{u}_{i+1/2}}{\alpha_{i+1/2}} & -\frac{\bar{h}_{i+1/2}}{\alpha_{i+1/2}} \\ 1 & 0 \end{pmatrix}, \quad \alpha_{i+1/2} := (\bar{u}_{i+1/2})^2 - g\bar{h}_{i+1/2} \, .$$

The proof is a tedious but straightforward calculation.
Since the EC scheme is well-balanced with respect to the staggered momentum, I define a staggered equilibrium variable $\widetilde{P}_i$ as follows

$$\widetilde{P}_i := (\frac{1}{2}(M_{i+1/2} + M_{i-1/2}), p_i)^T \, .$$

Note that $\widetilde{P}_i \approx P_i$; they differ of order Δx. For the scheme, I am going to use the *staggered equilibrium variable* rather than the *classical equilibrium variable*, because EC scheme may not preserve the classical momentum.
Consider now the standard *Roe* diffusion matrix that acts on the jump of the conservative variable. Then, based on the previous consideration, one gets the following relation

$$\begin{aligned} Q^{Roe}_{i+1/2}[\![U]\!]_{i+1/2} &= R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}[\![U]\!]_{i+1/2} \\ &\overset{(2.31)}{=} R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(U_P)_{i+1/2}[\![P]\!]_{i+1/2} \\ &\approx R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(U_P)_{i+1/2}[\![\widetilde{P}]\!]_{i+1/2} \, . \end{aligned} \tag{2.32}$$

$R_{i+1/2}$ and $|\Lambda_{i+1/2}|$ is defined as in (2.15).
In the $(U_P)_{i+1/2}$ matrix, $\alpha_{i+1/2}$ might be equal to zero for some i. To avoid dividing by zero, I define $(\widetilde{U}_P)_{i+1/2}$ as follows

$$(\widetilde{U}_P)_{i+1/2} := \begin{pmatrix} \frac{\bar{u}_{i+1/2}}{\widetilde{\alpha}_{i+1/2}} & -\frac{\bar{h}_{i+1/2}}{\widetilde{\alpha}_{i+1/2}} \\ 1 & 0 \end{pmatrix}, \quad \widetilde{\alpha}_{i+1/2} := \begin{cases} \max\{\alpha_{i+1/2}, \varepsilon\} & \text{if } \alpha_{i+1/2} \geqslant 0 \\ \min\{\alpha_{i+1/2}, -\varepsilon\} & \text{if } \alpha_{i+1/2} \leqslant 0 \end{cases} .$$

where $\varepsilon > 0$ is some small tolerance value (e.g. 10^{-6}).
Then the relation (2.32) becomes

$$\begin{aligned} Q^{Roe}_{i+1/2}[\![U]\!]_{i+1/2} &\approx R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(U_P)_{i+1/2}[\![\widetilde{P}]\!]_{i+1/2} \\ &\approx R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(\widetilde{U}_P)_{i+1/2}[\![\widetilde{P}]\!]_{i+1/2} \, . \end{aligned} \tag{2.33}$$

I then define the first order well-balanced diffusion matrix

$$D^{WB1}_{i+1/2} := R_{i+1/2}|\Lambda_{i+1/2}|R^{-1}_{i+1/2}(\widetilde{U}_P)_{i+1/2} \, . \tag{2.34}$$

The corresponding flux is given by

$$F^{WB1}_{i+1/2} := F^{EC}_{i+1/2} - \frac{1}{2} D^{WB1}_{i+1/2} [\![\widetilde{P}]\!]_{i+1/2} \,. \tag{2.35}$$

And the *well-balanced first order* scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t} U_i = -\frac{1}{\Delta x} (F^{WB1}_{i+1/2} - F^{WB1}_{i-1/2}) - S_i \,. \tag{2.36}$$

The scheme is consistent, since F^{EC} is consistent and the jump $[\![\widetilde{P}]\!]_{i+1/2}$ in the diffusion term vanishes if one tests consistency.

Lemma 2.10 *The WB1 scheme* (2.36) *preserves the moving equilibrium steady state for the staggered equilibrium variable. So, if*

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

for some constants C_1, C_2.
Then

$$\frac{\mathrm{d}}{\mathrm{d}t} h_i \equiv 0, \quad \frac{\mathrm{d}}{\mathrm{d}t} h_i u_i \equiv 0 \qquad \forall i$$

Proof F^{EC} preserves this steady state as shown in lemma 2.8. The jump $[\![\widetilde{P}]\!]_{i+1/2}$ vanishes, since

$$\begin{aligned} [\![\widetilde{P}^{(1)}]\!]_{i+1/2} &= \frac{1}{2}(M_{i+3/2} + M_{i+1/2}) - \frac{1}{2}(M_{i+1/2} + M_{i-1/2}) \\ &= 0 \\ [\![\widetilde{P}^{(2)}]\!]_{i+1/2} &= p_{i+1} - p_i = 0 \end{aligned}$$

where the superscript in $\widetilde{P}^{(1),(2)}$ denotes the first, respectively second component of $\widetilde{P}$. □

It can be showed that the *WB1* scheme is first order accurate. As in the previous section about energy stable schemes, I introduce a second order extension for the *well-balanced* scheme. It is done in a similar way as with the energy stable schemes.

2.5.2 Well-Balanced Second Order Scheme

I replace the piecewise constant cell averages in terms of the equilibrium variable $\widetilde{P}_i$ by a non-oscillatory piecewise linear reconstruction.
Define

$$\widehat{P}_i(x) := \widetilde{P}_i + \widetilde{P}'_i \cdot (x - x_i), \qquad x \in [x_{i-1/2}, x_{i+1/2}]$$

where

$$\widetilde{P}_i' := \frac{1}{\Delta x}\text{minmod}(\widetilde{P}_{i+1} - \widetilde{P}_i, \widetilde{P}_i - \widetilde{P}_{i-1}) \ .$$

Define the right and the left reconstructed edge values as follows

$$\begin{aligned}
\widetilde{P}_i^r := \widehat{P}_i(x_{i+1/2} &= \widetilde{P}_i + \widetilde{P}_i' \cdot (x_{i+1/2} - x_i) \\
&= \widetilde{P}_i + \frac{\Delta x}{2}\widetilde{P}_i' \\
&= \widetilde{P}_i + \frac{1}{2}\text{minmod}(\widetilde{P}_{i+1} - \widetilde{P}_i, \widetilde{P}_i - \widetilde{P}_{i-1}) \\
\widetilde{P}_{i+1}^l := \widehat{P}_{i+1}(x_{i+1/2} &= \widetilde{P}_{i+1} + \frac{1}{2}\text{minmod}(\widetilde{P}_{i+2} - \widetilde{P}_{i+1}, \widetilde{P}_{i+1} - \widetilde{P}_i) \ .
\end{aligned}$$

Define the *well-balanced second order (WB2)* flux as follows

$$F_{i+1/2}^{WB2} := F_{i+1/2}^{EC} - \frac{1}{2}R_{i+1/2}|\Lambda_{i+1/2}|R_{i+1/2}^{-1}(U_P)_{i+1/2}(\widetilde{P}_{i+1}^l - \widetilde{P}_i^r) \ . \tag{2.37}$$

The diffusion matrix is the same as before, only the jump is reconstructed. The corresponding WB2 scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i = -\frac{1}{\Delta x}(F_{i+1/2}^{WB2} - F_{i-1/2}^{WB2}) - S_i \ . \tag{2.38}$$

As before, the WB2 scheme is consistent, second order accurate and preserves the moving equilibrium steady state.

Remark 2.11 *The question whether the well-balanced schemes are energy stable or not, is not answered yet, either. But numerical experiments indicate that the schemes perform well.*

2.6 Numerical Experiments for the Well-Balanced Schemes

The well-balanced schemes are designed to preserve the moving equilibrium state. First, I consider an example with the sub- and supersonic steady state.

2.6.1 The Subsonic and Supersonic Steady State

The domain is given by $\Omega = [0, 20]$ and the bottom topography b as above in (2.13). I impose the following initial condition in terms of the staggered equilibrium variables

$$p_i = 22.07, \quad M_{i+1/2} = 4.42 \qquad \forall i. \tag{2.39}$$

Now to get the conservative variables h, hu, I simply take

$$hu \approx M_{i+1/2} = 4.42 =: C$$

and with $p_i = \frac{u_i^2}{2} + g(h_i + b_i)$ I have the following equation

$$\begin{aligned} 22.07 &\overset{!}{=} \frac{u_i^2}{2} + g(h_i + b_i) \\ &\approx \frac{C^2}{2h_i^2} + g(h_i + b_i) \end{aligned}$$

which is equivalent to the following third order equation in h_i

$$gh_i^3 + (gb_i - 22.07)h_i^2 + \frac{C^2}{2} \overset{!}{=} 0 \,.$$

Generally, this equation has three dinstict solutions for h_i. Note that the solutions depend on b_i (g and C are constant). In my case, there are always two positive and one negative solution, so I can sort out the negative solution. The other two solutions correspond to the *subsonic* respectively *supersonic* state. Figure 2.13 shows a supersonic steady state. Supersonic means that the *Froude number* $\mathrm{Fr} = \frac{|u|}{\sqrt{gh}}$ is greater than 1. Figure 2.14 shows the subsonic steady state. In this case the, I have $\mathrm{Fr} < 1$.
These two states are clearly preserved with the well-balanced schemes, as one can observe in 2.13 and 2.14 and. Even at large time scale, the steady state is preserved. In the figures, I have only plotted the solution of the WB2 scheme, since the solution of the WB1 and WB2 schemes are indistinguishable in the plots.

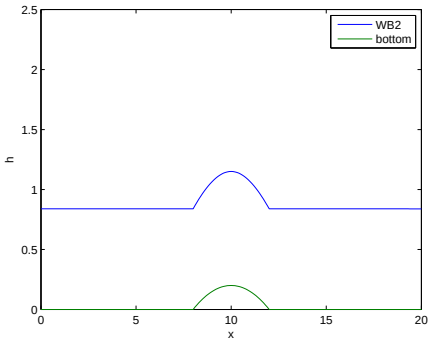

Figure 2.13: The supersonic case, $T = 20$

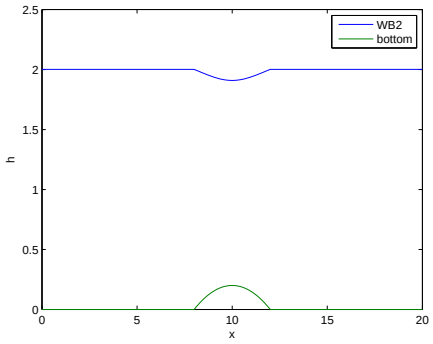

Figure 2.14: The subsonic case, $T = 20$

2.6.2 Perturbated States

As in the lake at rest experiment in section 2.4.3, I want to test whether a small initial perturbation may cause problems with the well-balanced schemes. The initial condition of this experiment is the same as in the previous subsection 2.6.1, except in the region $|x - 6| < \frac{1}{4}$, I add 0.01 the the

initial height. Mathematically formulated:
First, I have the initial values as in (2.39). From this information, I get the two admissible height found in the previous experiment. I focus on the subsonic case, denote the discrete height (without perturbation) values as $\widetilde{h}_i = \widetilde{h}(x_i)$. So, the perturbated initial height h_i is given by

$$h_i = h(x_i) = \begin{cases} \widetilde{h}_i + 1.01 & \text{if } |x_i - 6| < \frac{1}{4} \\ \widetilde{h}_i & \text{else} \end{cases} .$$

Note that, $M_{i+1/2}$, the staggered momentum, is kept constant.
I use the WB2 scheme to compute the approximate solutions. In contrast to the previous perturbated states experiment, I cannot say how the true solution looks like. Figure 2.15 and 2.16 show the supersonic steady state with the initial perturbation. The created wave moves to the right side. As soon as it dissappears in the hump, no real effect can be seen. When the wave moves outward the hump (figure 2.15) it has lost its original shape and looks like a small "cavity". There are also small oscillations in the region where the perturbation has started.
Figure 2.17 and 2.18 show the subsonic steady state with the initial perturbation. Here, the approximate solution consists of two waves: a left-going and a right-going one. The right-going wave passes the cavity in the middle without changing a lot (apparently), as one can see in figure 2.18. As before, there are small oscillations in the regions where the perturbation originated (figure 2.18).

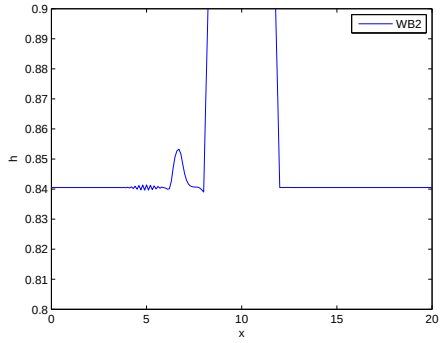

Figure 2.15: Perturbated supersonic state, $T = 0.3$.

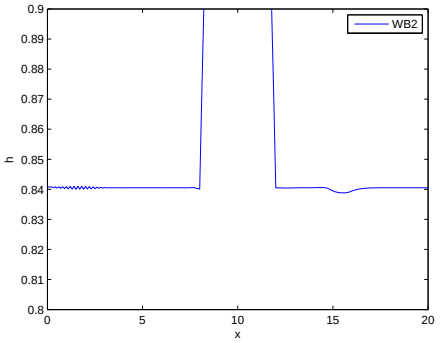

Figure 2.16: Perturbated supersonic state, $T = 1.2$.

Remark 2.12 *I receive some tiny osciallations in my plots, this is probably due to my implementation.*

Since I do not have any true solution for this experiment I compare my results to a reference solution, in this case the WB2 scheme on a refined

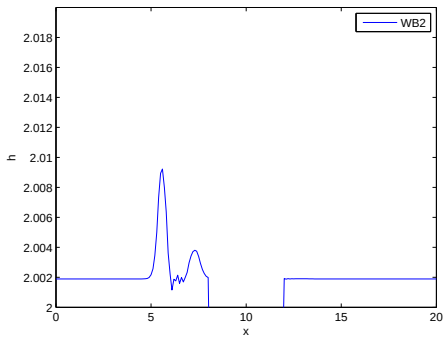

Figure 2.17: Perturbated subsonic state, $T = 0.2$.

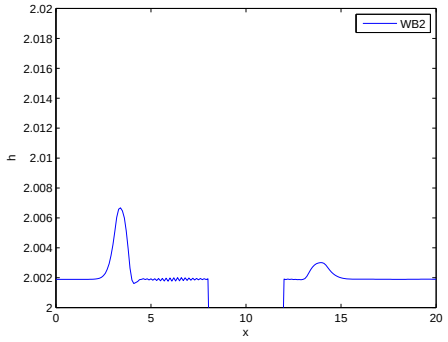

Figure 2.18: Perturbated subsonic state, $T = 1.2$.

mesh ($N = 1600$). The reference solutions in the figures 2.19, 2.20, 2.21 and 2.21 show the approximate solution at the same endtime T as the solutions computed on the coarse mesh ($N = 200$). The oscillations mentioned in remark 2.12 are damped to very small "noise". The small cavity in figure 2.16 is much less diffusive in the reference solution (figure 2.20). As for all waves, they are a slightly diffusive when computed on the coarse mesh ($N = 200$). But the approximate solutions on the coarse mesh capture the reference solutions quite well.

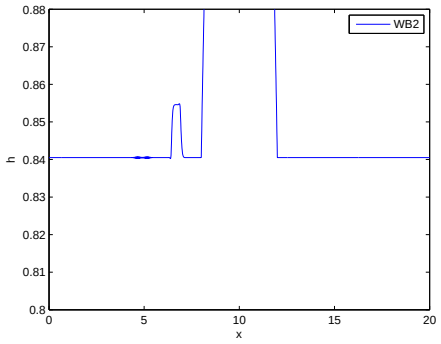

Figure 2.19: Reference solution, $T = 0.3$.

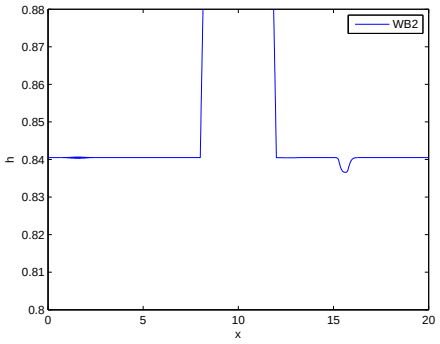

Figure 2.20: Reference solution, $T = 1.2$.

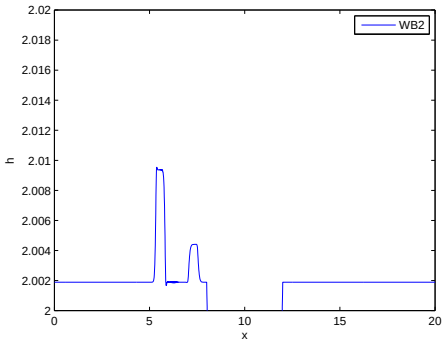

Figure 2.21: Reference solution, $T = 0.2$.

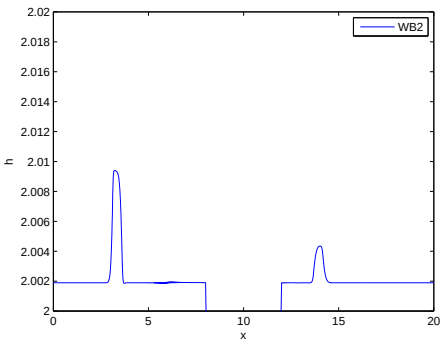

Figure 2.22: Reference solution, $T = 1.2$.

Chapter 3

Dry Areas

3.1 Basic Examples

Up to here, I have assumed that there are no dry areas in the initial situation of an experiment and the domain stays flooded during the whole time stepping. But what does happen with the introduced schemes (EC, ES, WB), if there are dry areas included.
I first consider the following simple example to show the limitation of the schemes derived in the previous section.

Example 3.1 *In this example, there are no dry areas in the initial setup, but after some time there will be a dry area. The number of grid cells and the boundary conditions are as usual. In this example I use a different CFL number, namely* $CFL = 0.1$. *The bottom topography is flat and the domain is* $\Omega = [-1, 1]$. *The initial condition is given by*

$$h(x,0) \equiv 0.7, \quad (hu)(x,0) = \begin{cases} -7 & \text{if } x < 0 \\ 7 & \text{if } x \geqslant 0 \end{cases}, \quad b \equiv 0\,. \tag{3.1}$$

I use the WB2 scheme to compute the approximate solution. With this initial data, the middle region is going to be dry after some short time period. As one can see in figure 3.1, the approximate solution does not remain positive in the middle region; it is slightly negative at $T = 0.0055$. *A bit later at* $T = 0.0089$, *see figure 3.2, the height of the approximate solution is just right before blowing up. All the other schemes (EC, ES) presented in the previous chapter also show this phenomenon.*

The problem in this very simple example is the negative height. None of the schemes are positivity preserving with respect to the height, as example 3.1 showed us.
Another easy example has already been introduced in section 2.2.1.

Example 3.2 *In this example, there are dry areas in the initial setup. The domain is given by* $\Omega = [0, 20]$. *CFL and number of grid points are as usual. The initial*

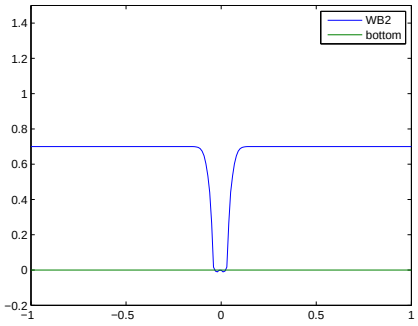

Figure 3.1: WB2 scheme, $T = 0.0055$

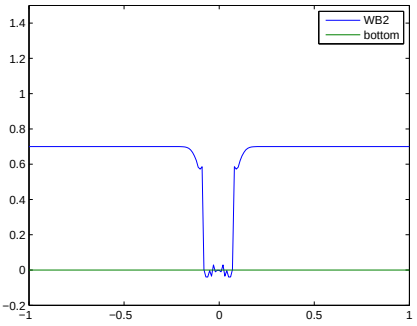

Figure 3.2: WB2 scheme, $T = 0.0089$

condition is given by

$$u(x,0) \equiv 0, \quad h(x) + b(x) = \max\{1, b(x)\} \tag{3.2}$$

where the bottom is defined as

$$b(x) := \begin{cases} \frac{4-(x-10)^2}{3.5} & \text{if } x \in [8,12] \\ 0 & \text{else.} \end{cases} \tag{3.3}$$

This is the so-called lake at rest with dry areas. Figure 3.3 shows the initial setup. The WB schemes fail immediately, since the $R_{i+1/2}$ matrix is singular for some i's and so the inverse matrix $R^{-1}_{i+1/2}$ in (2.34) does not exist.
The energy stable schemes perform only slightly better. They start oscillating right in the beginning and the height h_i is negative for some i. After a short time, this approximate solution blows up. In figure 3.4, one can see the approximate solution computed with the EC scheme. It starts oscillating and the water seems to "drain out" of the domain. At $T = 18$, the water level is already significantly lowered. At some later time, this approximate solution performed by the EC scheme also blows up like the other schemes. An observation, which one cannot see out of the plot, is that the approximate height h_i is negative for a small number of i's.
Of course, all of the above mentioned observations are physical nonsense, since the lake should stay at rest even if there is a dry area. So, none of these schemes can preserve the lake at rest with dry area although all of these schemes can preserve the standard lake at rest.

The problem of the EC, ES and WB schemes is that they are not positivity preserving with respect to h. Due to real world or physical observation, it is implicit to require positive water height. So, in the next section 3.2, I present a positivity preserving flux limiter. In section 3.4, I choose another scheme which is positivity preserving by its own to perform some numerical experiments.

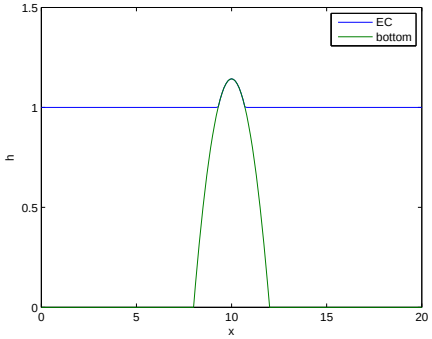

Figure 3.3: Lake at rest with dry area, $T = 0$

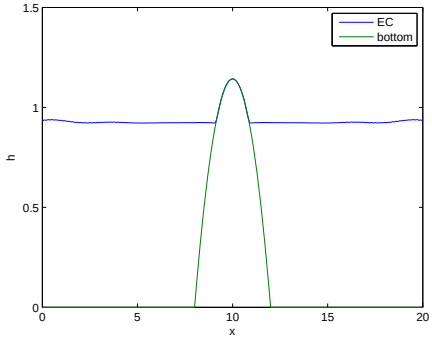

Figure 3.4: Lake at rest with dry area, $T = 18$

3.2 Cut-Off Flux Limiter

The idea of the cut-off flux limiter is very easy. It simply cuts off the numerical flux as soon as the scheme renders a positive value to a negative one. It is based on a cut-off flux limiter for gas dynamics and can be found in [3]. It is originally designed to keep the density in gas dynamics positive.
If I have a water height $h(x_i, t^n) = h_i^n \geqslant 0 \quad \forall i$ I want to preserve this positivity after one evolution time step, i.e. $h_i^{n+1} \geqslant 0 \quad \forall i$. The superscript denotes the index of the time level t ($t^n = n \cdot \Delta t$, I assume a uniform time grid) and not the n-th power of it.
For the beginning, I assume the conservation law of the shallow water equations. Consider the discrete time stepping of the finite volume scheme, i.e.

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2} - F_{i-1/2}) \tag{3.4}$$

where U_i^n is the conservative variable at time t^n and $F_{i+1/2} = (F_{i+1/2}^{(1)}, F_{i+1/2}^{(2)})^T$ some numerical flux.
I multiply (3.4) by 2 and rewrite it as follows

$$\begin{aligned} 2U_i^{n+1} &= 2U_i^n + 2\frac{\Delta t}{\Delta x}(F_{i-1/2} - F_{i+1/2}) \\ &= (U_i^n + 2\frac{\Delta t}{\Delta x}F_{i-1/2}) + (U_i^n - 2\frac{\Delta t}{\Delta x}F_{i+1/2}) \\ &= U_i^- + U_i^+ \end{aligned} \tag{3.5}$$

If U_i^- and U_i^+ both are positive, then so is U_i^{n+1}.
Let $\varepsilon > 0$ be some small tolerance, then the cut-off flux limiter $\theta_{i+1/2}$ is defined as follows:

1. Set $\theta_{i+1/2}^+ = \theta_{i+1/2}^- = 1 \quad \forall i$.

2. If $h(U_i^+) < \varepsilon$, i.e. if

$$h_i - 2\frac{\Delta t}{\Delta x} F_{i+1/2}^{(1)} < \varepsilon \tag{3.6}$$

then, the following equation is solved for the unknown $\theta_{i+1/2}^+$

$$\begin{aligned} h_i - 2\frac{\Delta t}{\Delta x}\theta_{i+1/2}^+ F_{i+1/2}^{(1)} &\stackrel{!}{=} \varepsilon \\ \Rightarrow \quad \theta_{i+1/2}^+ &= \frac{h_i - \varepsilon}{2\frac{\Delta t}{\Delta x} F_{i+1/2}^{(1)}} \, . \end{aligned} \tag{3.7}$$

3. If $h(U_{i+1}^-) < \varepsilon$, i.e. if

$$h_{i+1} + 2\frac{\Delta t}{\Delta x} F_{i+1/2}^{(1)} < \varepsilon \tag{3.8}$$

then, the following equation is solved for the unknown $\theta_{i+1/2}^-$

$$\begin{aligned} h_{i+1} + 2\frac{\Delta t}{\Delta x}\theta_{i+1/2}^- F_{i+1/2}^{(1)} &\stackrel{!}{=} \varepsilon \\ \Rightarrow \quad \theta_{i+1/2}^- &= \frac{\varepsilon - h_{i+1}}{2\frac{\Delta t}{\Delta x} F_{i+1/2}^{(1)}} \, . \end{aligned} \tag{3.9}$$

4. Define $\theta_{i+1/2} := \min\{\theta_{i+1/2}^-, \theta_{i+1/2}^+\}$. Define the limited flux $F_{i+1/2}^\star$ as

$$F_{i+1/2}^\star := \theta_{i+1/2} F_{i+1/2} \, . \tag{3.10}$$

Remark 3.3 *It may happen that one has $h_i < \varepsilon$ (e.g. if $h_i = 0$ in the initial condition). To prevent from occuring, I define $\widetilde{h}_i$ as follows*

$$\widetilde{h}_i := \max(h_i, \varepsilon)$$

and I consider this height $\widetilde{h}$ for the computation.
Otherwise, if for instance $\exists i$ such that $h_i < \varepsilon$ and $F_{i+1/2}^{(1)} = 0$, then (3.6) *and* (3.8) *hold. Hence I redefine $\theta_{i+1/2}^\pm$ as in* (3.7) *and* (3.9)*, but this would fail (because of the division by zero).*

Remark 3.4 *The original cut-off flux limiter in [3] is slightly different. U_i^- and U_i^+ are defined as follows:*

$$2U_i^{n+1} = (U_i^n + 2\frac{\Delta t}{\Delta x}(F_{i-1/2} - f_i)) + (U_i^n + 2\frac{\Delta t}{\Delta x}(f_i - F_{i+1/2})) =: U_i^- + U_i^+$$

where f_i is the true flux of the shallow water equations (or the true flux of any considered conservation or balance law). In this case, $\theta_{i+1/2}$ also depends on the

true flux. This flux limiter is easily proved to be consistent and accuracy preserving. However, for the cut-off flux limiter presented in section 3.2, I cannot give any proof of consistency or accuracy preservation. So, numerical experiments might give some information about the performance of this flux limiter.
I use the flux limiter derived in this section, because numerical experiments have not been succesful with the original flux.

Remark 3.5 *In fact, the tolerance ε is a lower bound for the height. So, if $h = \varepsilon$ in some region (I assume that h is defined as in remark 3.3), then this region is called a* dry area.

3.3 Numerical Experiments with the Flux Limiter

In this section, I test the schemes introduced in chapter 2 *augmented with the cut-off flux limiter* to several experiments. I use the usual CFL number ($CFL = 0.4$), number mesh points ($N = 200$) and boundary condition (open Neumann) if I do not state differently. The tolerance ε for the cut-off flux limiter is set to 10^{-6}.

3.3.1 Initial Momentum with Flat Bottom Topography

Basically, the setup of this experiment is the same as in example 3.1. The CFL number is 0.1 and I use the WB2 scheme. The approximate solution stays positive, as one can see in figure 3.5 and 3.6, and it looks physically meaningful. The dry area in the middle in figure 3.6 is not really dry, i.e. it has a thin layer of water with height ε. If the experiment runs for a longer time, at some time there will be no water left in the domain.

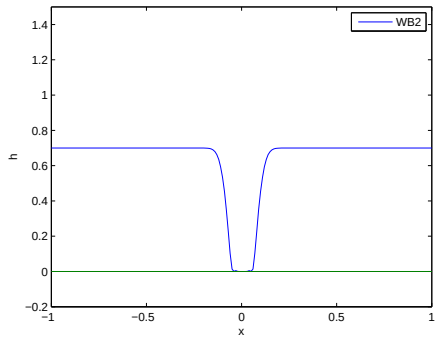

Figure 3.5: WB2 scheme, $T = 0.0089$

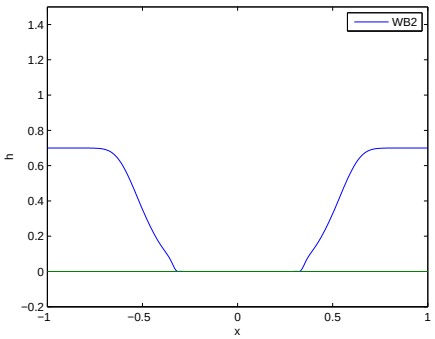

Figure 3.6: WB2 scheme, $T = 0.05$

3.3.2 Initial Momentum with Continuous Bottom Topography

In this experiment, I have a non-flat bottom topography. b is defined as follows

$$b(x) = \begin{cases} -20x^2 + 0.2 & \text{if } x \in (-0.1, 0.1) \\ 0 & \text{else.} \end{cases}$$

I consider two different initial conditions. The first one is given by

$$h(x,0) = 1 - b(x), \qquad (hu)(x,0) = \begin{cases} -7 & \text{if } x < 0 \\ 7 & \text{if } x \geqslant 0 \end{cases}. \tag{3.11}$$

And the second initial condition is given by

$$h(x,0) = 1 - b(x), \qquad (hu)(x,0) = \begin{cases} -7(1 - b(x)) & \text{if } x < 0 \\ 7(1 - b(x)) & \text{if } x \geqslant 0 \end{cases}. \tag{3.12}$$

The difference between these two initial conditions is that in the first one the whole momentum hu is kept constant and in the second one only the speed u is constant. I use the WB2 scheme with the cut-off flux limiter to compute the approximate solution and I take a different CFL number, namely 0.1. Figure 3.7 and 3.8 show the approximate solution for the initial

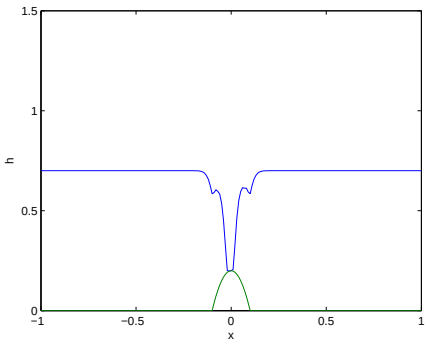

Figure 3.7: WB2 scheme, $T = 0.005$

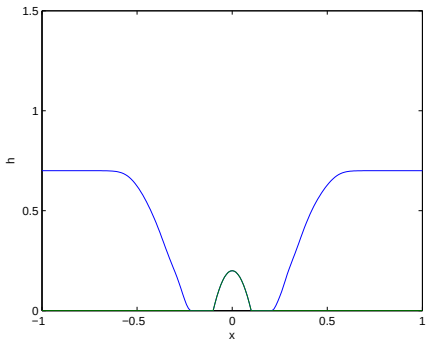

Figure 3.8: WB2 scheme, $T = 0.05$

data (3.12) computed with a coarse mesh ($N = 200$). As one can observe, the region in the middle becomes dry after some time and the scheme is able to manage the dry region. Since I lack true solutions I compare the approximate solution to a reference solution. In this case, it is the solution computed with the WB2 on a refined mesh ($N = 1600$). The figures 3.9 and 3.10 show the reference solution at the same time as the solution computed on the coarse mesh. The approximate solution is slightly more diffusive than the reference solution, as one can observe for instance in figure 3.9 in comparison to 3.7. Apart from this, they do not differ much.

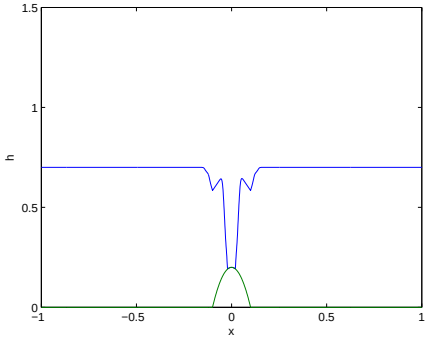

Figure 3.9: Reference solution, $T = 0.005$

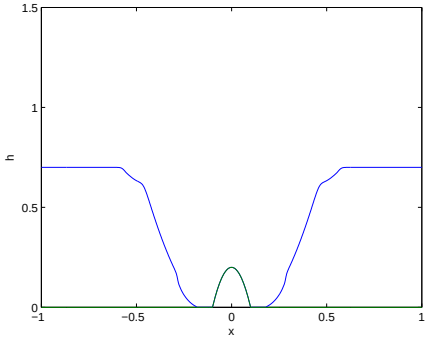

Figure 3.10: Reference solution, $T = 0.05$

Figure 3.11 and 3.12 show the approximate solution for the second initial data (3.12). As above, I use the WB2 scheme with the cut-off flux limiter and as reference solution I compute the approximate solution on a refined mesh ($N = 1600$), see figure 3.13 and 3.14. The approximate solution is slightly

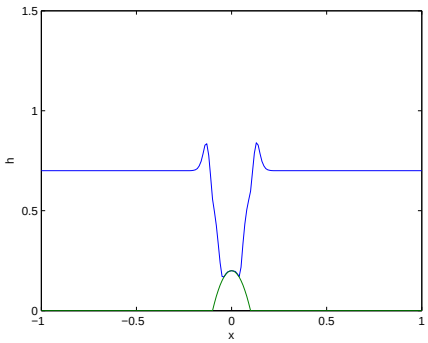

Figure 3.11: WB2 scheme, $T = 0.005$

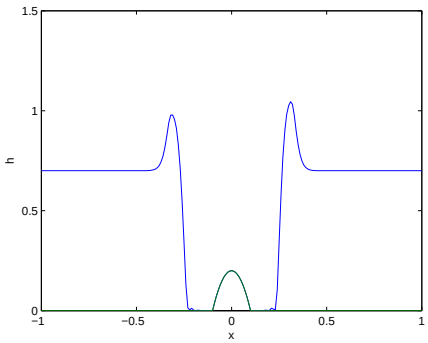

Figure 3.12: WB2 scheme, $T = 0.02$

diffusive in comparison to the reference solution. As one may notice, in figure 3.12 or in 3.11 the wave on the right hand-side is higher than on the left hand-side. This effect is probably due to the discretization of the initial condition (3.12) and the bottom topography. The point $x = 0$ contributes to the left hand-side in the initial condition and the bottom topography is symmetric around zero. So, the "splitting" is not exactly done in the middle but a bit on the right, what may lead to the observed height difference of the waves. If I refine the mesh, these different wave heights vanish, see figure 3.14.

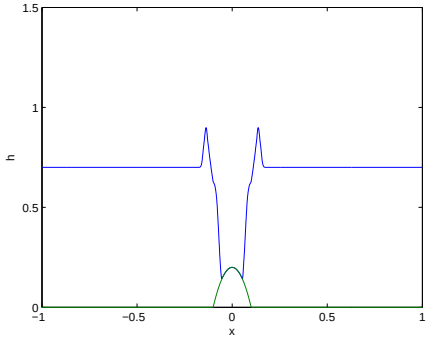

Figure 3.13: Reference solution, $T = 0.005$

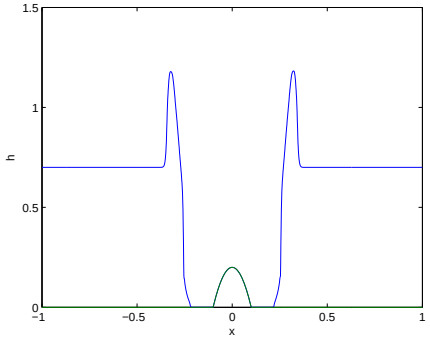

Figure 3.14: Reference solution, $T = 0.02$

3.3.3 Initial Momentum with Discontinuous Bottom Topography

This experiment can be found in [5] (section 5.3). I use the WB2 scheme with the cut-off flux limiter. I take the CFL number 0.1. The domain is $\Omega = [0, 25]$. The bottom topography is given by

$$b(x) = \begin{cases} 1 & \text{if } x \in [\frac{25}{3}, \frac{25}{2}] \quad (\approx [8.33, 12.5]) \\ 0 & \text{else.} \end{cases} \tag{3.13}$$

The initial condition is given by

$$h(x,0) = 10 - b(x), \qquad (hu)(x,0) = \begin{cases} -350 & \text{if } x < \frac{50}{3} (\approx 16.67) \\ 350 & \text{if } x \geqslant \frac{50}{3} \end{cases}. \tag{3.14}$$

Figure 3.16 shows the approximate solution computed on a refined mesh

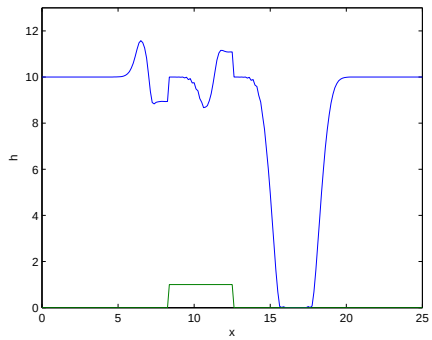

Figure 3.15: WB2 scheme, $T = 0.05$

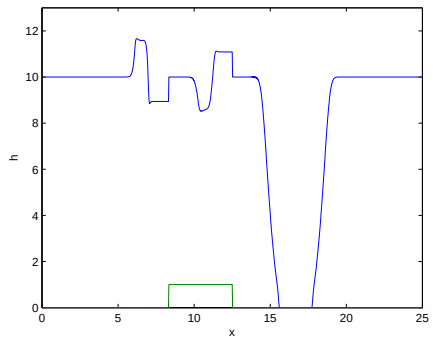

Figure 3.16: Refinded mesh $N = 800$, $T = 0.05$

($N = 800$). This approximate solution looks almost the same as the solution in figure 3.15. The approximate solution computed on the coarse mesh (the

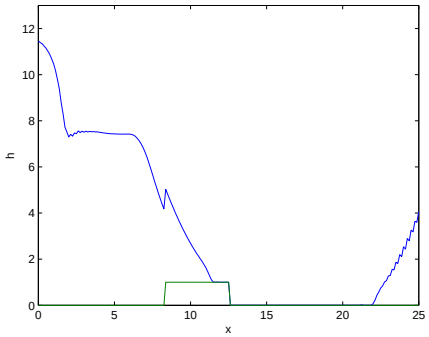

Figure 3.17: WB2 scheme, $T = 0.25$

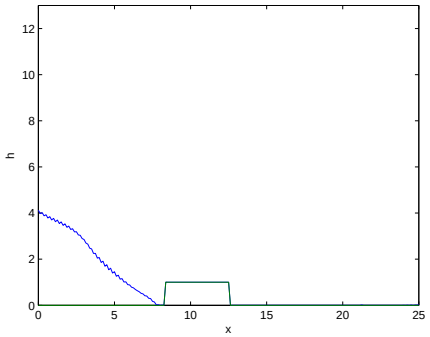

Figure 3.18: WB2 scheme, $T = 0.45$

usual $N = 200$) presents some more oscillations and is more diffusive than the one computed on the refined mesh. Overall, the approximate solution in figure 3.15, 3.17 and 3.18 matches quite well with the numerical result obtained in [5].

3.3.4 Lake at Rest with Initial Dry Area

I now consider a simple lake at rest with a dry area in the initial setup. This is a steady state and should be preserved during the time stepping. I use the *EC* scheme with the cut-off flux limiter. I take the usual CFL number 0.4. The domain is $\Omega = [0, 20]$ and the bottom topography is the same as in example 3.2. Recall:

$$b(x) = \begin{cases} \frac{4-(x-10)^2}{3.5} & \text{if } x \in [8, 12] \\ 0 & \text{else.} \end{cases}$$

The initial condition is given by

$$u(x, 0) \equiv 0, \qquad h(x, 0) = \begin{cases} 1 - b(x) & \text{if } b(x) < 1 - 10^{-6} \\ 10^{-6} & \text{else.} \end{cases} \tag{3.15}$$

A thin water layer is put on the dry area, cf. remark 3.3. Figure 3.19 shows the initial height. The scheme fails to preserve this steady state, since the approximate solution oscillates and creates unphysical waves, as one can observe in figure 3.20. So, the schemes presented so far cannot handle this example.
If there are no initial dry areas, the EC and the WB2 schemes perform well and give a reasonable approximate solution and a relatively coarse mesh ($N = 200$). But as soon as there are initial dry areas, the previous schemes fail to compute reasonable approximate solutions. Due to the above mentioned observation, simulating the flooding of dry areas with the schemes may also fail.

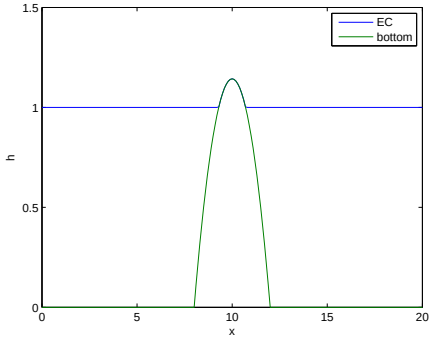

Figure 3.19: Initial height, $T = 0$

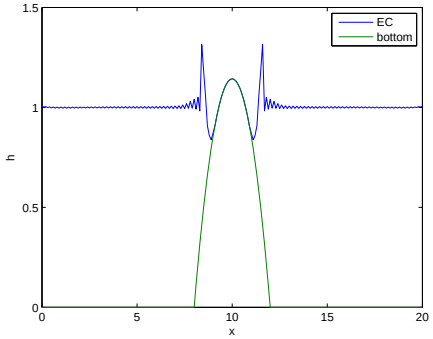

Figure 3.20: EC scheme, $T = 3.3$

3.3.5 Problematic Flux Tolerance

The cut-off flux limiter needs a small tolerance value ε. It seems that the experiments may be very sensitive to this tolerance. If I take a smaller tolerance value than 10^{-8} for the flux limiter, the approximate solution of the numerical experiment in section 3.3.2 blows up after some time.
However, in the experiment of section 3.3.3, taking ε smaller than 10^{-8}, e.g. 10^{-10}, does not affect the approximate solution, i.e. the solution does not explode. I would need further analysis of the flux limiter and its tolerance, but this lies beyond the scope of the thesis.

Remark 3.6 *The tolerance value of the diffusion matrix in* (2.34) *and the tolerance value of the flux limiter do not seem to be correlated.*

3.4 Rusanov Scheme

The schemes derived in chapter 2 are not positivity preserving and therefore have difficulties computing numerical experiments with dry areas. The cut-off flux limiter introduced in section 3.2 works with these schemes as long as there is no initial dry areas. I this section, I introduce the standard *Rusanov scheme* to encounter the problems of the previous schemes.
The general Rusanov flux is given by

$$F^{Rus}_{i+1/2} := \frac{1}{2}(f(U_i) + f(U_{i+1})) - \frac{1}{2}\max\{|\lambda^{\pm}_i|, |\lambda^{\pm}_{i+1}|\}(U_{i+1} - U_i) \tag{3.16}$$

where f is the true flux of the conservation/balance law and λ is given by

$$\lambda^{\pm}_i := u_i \pm \sqrt{gh_i}\,. \tag{3.17}$$

The Rusanov flux is energy stable (see [1]). The resulting scheme looks as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}U_i = -\frac{1}{\Delta x}(F^{Rus}_{i+1/2} - F^{Rus}_{i-1/2}) - S_i\,. \tag{3.18}$$

It is known that the Rusanov scheme is first order accurate.
First, consider the discrete time stepping of the conservation law (this is equivalent to the semidiscrete scheme with Forward Euler time stepping).

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2}^{Rus} - F_{i-1/2}^{Rus}) \,. \tag{3.19}$$

For the shallow water equations, I have the following relation for the first component of U_i:

$$\begin{aligned} h_i^{n+1} = \; & h_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2}^{Rus,(1)} - F_{i-1/2}^{Rus,(1)}) \\ =: \; & h_i^n - G(U_{i-1}^n, U_i^n, U_{i+1}^n) \,. \end{aligned}$$

If $G(U_{i-1}^n, U_i^n, U_{i+1}^n) \leqslant h_i^n$, then the scheme is positivity preserving since $h_i^n - G \geqslant 0$. For convenience, I drop the superscript n in the rest of the thesis.

Lemma 3.7 *The Rusanov scheme is positivity preserving for the conservation law. I.e. if $h_i^n \geqslant 0 \; \forall i$ then $h_i^{n+1} \geqslant 0 \; \forall i$.*

Proof As stated above, it is enough to show: $G \leqslant h_i$.

$$\begin{aligned} G = & \frac{\Delta t}{\Delta x}(\frac{1}{2}(h_i u_i + h_{i+1}u_{i+1}) - \frac{1}{2}\max\{|\lambda_i^+|, |\lambda_{i+1}^+|\}(h_{i+1} - h_i) \\ & - \frac{1}{2}(h_{i-1}u_{i-1} + h_i u_i) + \frac{1}{2}\max\{|\lambda_{i-1}^+|, |\lambda_i^+|\}(h_i - h_{i-1}) \\ = & \frac{\Delta t}{2\Delta x}(h_i(\max\{|\lambda_i^+|, |\lambda_{i+1}^+|\} + \max\{|\lambda_{i-1}^+|, |\lambda_i^+|\}) \\ & + h_{i+1}(u_{i+1} - \max\{|\lambda_i^+|, |\lambda_{i+1}^+|\}) \\ & - h_{i-1}(u_{i-1} + \max\{|\lambda_{i-1}^+|, |\lambda_i^+|\})) \\ =: & \; (\star) \end{aligned}$$

With the (restricted) CFL condition

$$\frac{\Delta t}{\Delta x}\max\{|u + \sqrt{gh}|\} \leqslant \frac{1}{2}$$

I have

$$\frac{\Delta t}{2\Delta x}(h_i(\max\{|\lambda_i^+|, |\lambda_{i+1}^+|\} + \max\{|\lambda_{i-1}^+|, |\lambda_i^+|\}) \leqslant \frac{1}{2}h_i$$

Define $\gamma := \frac{\Delta t}{2\Delta x}$. So, I have the following upper bound for $(\star)$:

$$\begin{aligned} (\star) \leqslant & \frac{1}{2}h_i + \gamma h_{i+1} \underbrace{(u_{i+1} - \max\{|\lambda_i^+|, |\lambda_{i+1}^+|\})}_{\leqslant 0} - \gamma h_{i-1} \underbrace{(u_{i-1} + \max\{|\lambda_{i-1}^+|, |\lambda_i^+|\})}_{\geqslant 0} \\ \leqslant & \; h_i \end{aligned}$$

Since $\gamma > 0$ and $h_i \geqslant 0 \; \forall i$.
Hence, the scheme is positivity preserving. □

For the balance law, the scheme remains positivity preserving as long as the source term is "nice" enough.

3.5 Numerical Experiments with the Rusanov Scheme

For the numerical experiments, I use the Rusanov method with the cut-off flux limiter to compute the approximate solutions. I use the flux limiter, although the Rusanov scheme is positivity preserving, because numerical errors may lead to very small negative heights h_i, which cause the scheme to fail. In contrast to the previous section, I use a very small tolerance for the flux limiter, namely I set $\varepsilon = 10^{-15}$. The CFL number varies in each experiment. The boundary condition and the mesh point are as usual if not stated differently. First, I consider a standard experiment.

3.5.1 Lake at Rest

In this section, I test the Rusanov scheme to two different lake at rest steady states. The main difference is the bottom topography.
In the first experiment, I consider the following bottom topography in the domain $\Omega = [0, 12]$

$$b(x) := \frac{1}{18}x^2 - \frac{2}{3}x + 2\,. \tag{3.20}$$

The CFL number is 0.1918 and the initial condition is given by

$$u(x,0) \equiv 0, \qquad h(x,0) = \begin{cases} 0.3 - b(x) & \text{if } b(x) < 0.3 - 10^{-15} \\ 10^{-15} & \text{else.} \end{cases} \tag{3.21}$$

I have initial dry areas and a smooth bottom topography. The Rusanov

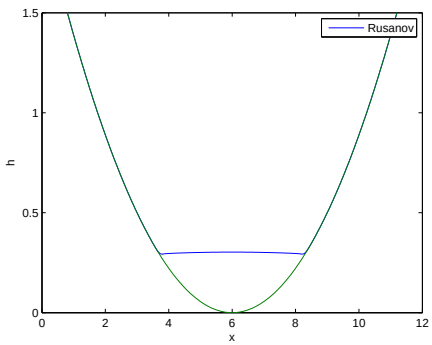

Figure 3.21: Lake at rest, $T = 10$

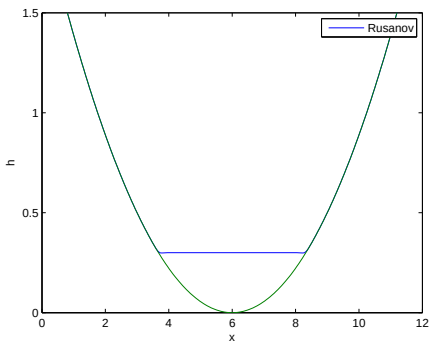

Figure 3.22: Lake at rest, $T = 50$

scheme produces very small oscillation in the beginning of time stepping, as one can see in figure 3.21. There are also some artifacts at the wet/dry

interface. But all these effects are rather small, since after some time the water height is as in the initial setup, see figure 3.22. The momentum is not perfectly preserved, due to this small perturbation in the beginning.
The second setup for the lake at rest is similar to the one from section 2.2.1. The bottom topography in the domain $\Omega = [0, 20]$ is given by

$$b(x) = \begin{cases} \frac{4-(x-10)^2}{4.1} & \text{if } x \in [8, 12] \\ 0 & \text{else.} \end{cases} \tag{3.22}$$

The CFL number is 0.1 and the initial condition is given by

$$u(x, 0) \equiv 0, \qquad h(x, 0) = 1 - b(x) \; . \tag{3.23}$$

I have no initial dry areas. Figure 3.23 shows rather strong oscillations of

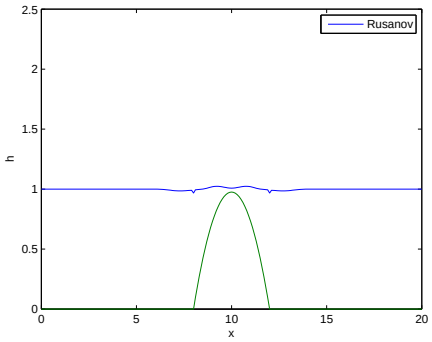

Figure 3.23: Lake at rest, $T = 0.4$

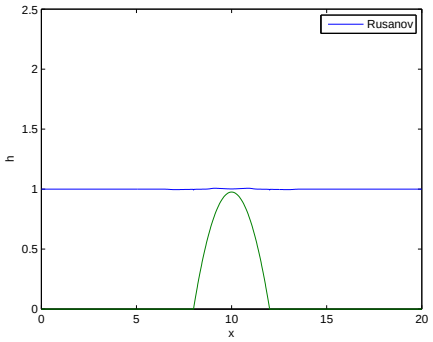

Figure 3.24: Refined mesh $N = 800$, $T = 0.4$

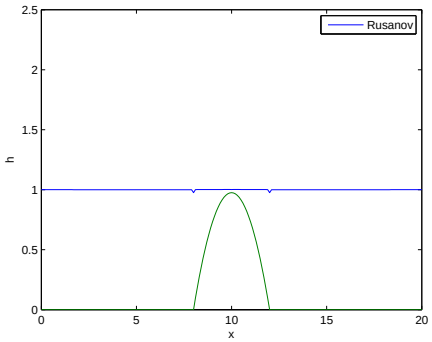

Figure 3.25: Lake at rest, $T = 10$

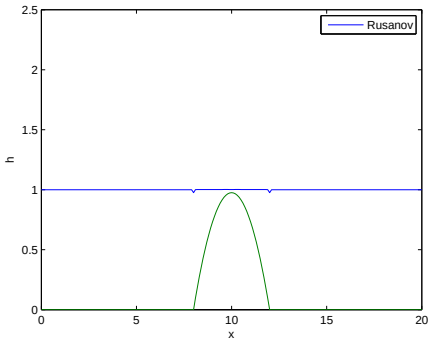

Figure 3.26: Lake at rest, $T = 50$

the approximate solution around the non-flat bottom region. In particular, there are two artifacts where the bottom is not smooth anymore but only

continuous. Both of these effects can be reduced by increasing the meshpoints, see figure 3.24 with $N = 800$. After some time, the height of the water remains constant and the state stays preserved, see figure 3.25 and 3.25. The momentum hu, though, is not preserved perfectly due to the oscillation in the beginning.
The Rusanov scheme does not preserve to lake at rest very well. But the deviation are only very small and may be neglected in practice. One could increase the accuracy by increasing the number of meshpoints.

3.5.2 Sloping Beach

In this experiment, I consider a beach and an incoming wave. The CFL number is 0.131, the boundary condition is as usual and the domain is $\Omega = [0, 12]$. The bottom topography of the beach is given by

$$b(x) = \begin{cases} \frac{6}{35}x - \frac{6}{7} & \text{if } x \geqslant 5 \\ 0 & \text{else.} \end{cases} \tag{3.24}$$

The initial condition is as follows

$$u(x,0) \equiv 0, \quad h(x,0) = \begin{cases} \max\{\max\{0.7, b(x)\} - b(x), 10^{-15}\} & \text{if } x \in [3.5, 4.5] \\ \max\{\max\{0.3, b(x)\} - b(x), 10^{-15}\} & \text{else.} \end{cases} \tag{3.25}$$

The figures 3.27-3.34 show the incoming wave hitting the beach and flowing

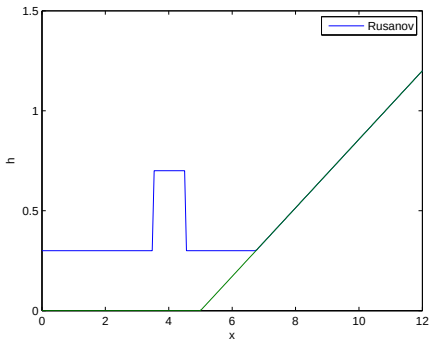

Figure 3.27: Initial height, $T = 0$

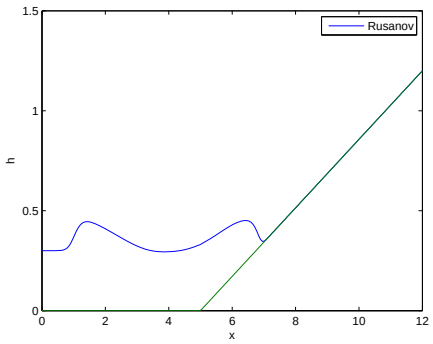

Figure 3.28: Sloping beach, $T = 1$

back. The wave might be a bit diffusive, since I use the Rusanov flux. But despite of this the approximate solution seems to be physically reasonable. The next experiment even goes a step further.

3.5.3 Parabolic Bowls

This experiment consists of two parabolic bowls which are separated by some small distance. They are connected with a quadratic curve, see figure 3.35.

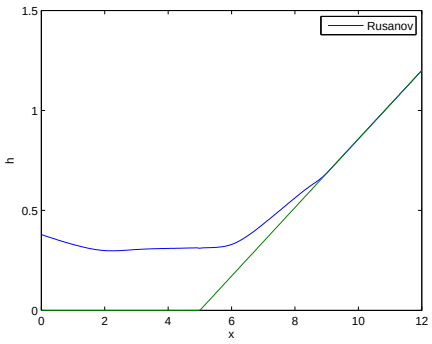

Figure 3.29: Sloping beach, $T = 2$

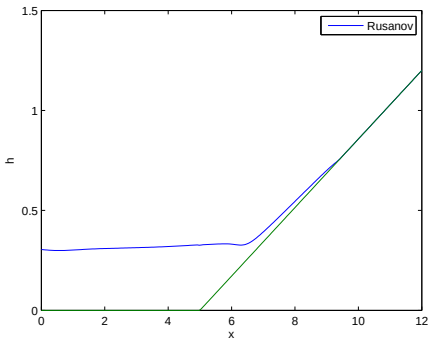

Figure 3.30: Sloping beach, $T = 3$

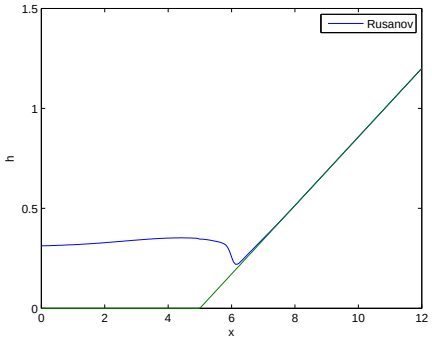

Figure 3.31: Sloping beach, $T = 4.5$

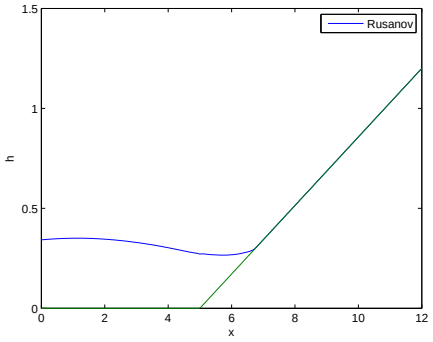

Figure 3.32: Sloping beach, $T = 6$

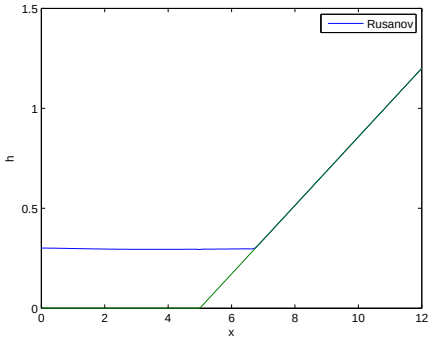

Figure 3.33: Sloping beach, $T = 12$

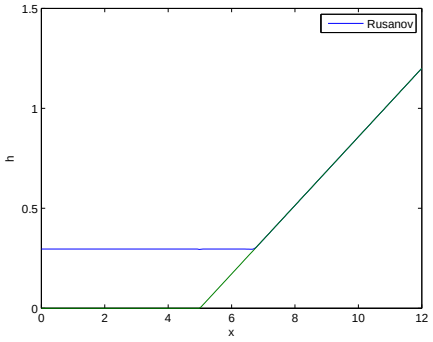

Figure 3.34: Sloping beach, $T = 20$

Both bowls contain some water. In the first bowl, I create two waves so that one of the waves spills over into the other bowl. The bottom topography (bowls) in the domain $\Omega = [0, 25]$ is given by

$$b(x) = \begin{cases} \frac{1}{18}x^2 - \frac{2}{3}x + 2 & \text{if } x \in [0, 12] \\ -x^2 + 25x - 154 & \text{if } x \in (12, 13) \\ \frac{1}{18}x^2 - \frac{19}{9}x + \frac{361}{18} & \text{if } x \in [13, 25] \ . \end{cases} \tag{3.26}$$

The CFL number is 0.3284 and the initial condition is given by

$$u(x, 0) \equiv 0, \quad h(x, 0) = \begin{cases} \max\{\max\{4.4, b(x)\} - b(x), 10^{-15}\} & \text{if } x \in [5.25, 6.75] \\ \max\{\max\{0.9, b(x)\} - b(x), 10^{-15}\} & \text{else.} \end{cases} \tag{3.27}$$

The figures 3.35-3.42 show the water height at different times. For the two times $T = 2$ and $T = 4.5$, I have computed a reference solution (with the Rusanov scheme on a refinded mesh $N = 1600$) to compare, see figure 3.43 and 3.44. This comparison shows that the Rusanov is indeed diffusive, but the further I go in time the less important is this fact, if one compares 3.44 to 3.39. In a physical point of view, the Rusanov scheme performs well even at the slopover. The waves in the two bowls move for a long time even at $T = 100$ there is no rest (figure 3.42). This is an expected observation, since in reality water in such bowls behaves in that way (maybe not exactly in this way, but similarly).

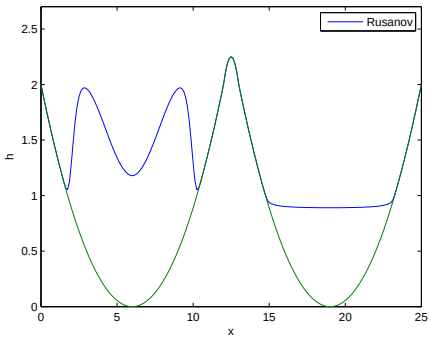

Figure 3.35: Parabolic bowls, $T = 0.5$

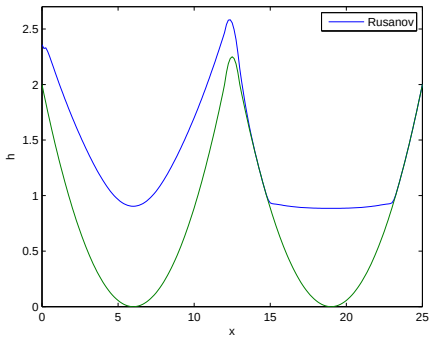

Figure 3.36: Parabolic bowls, $T = 1$

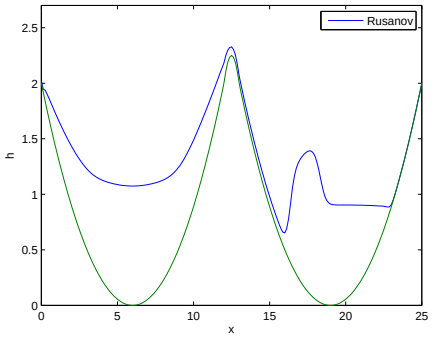

Figure 3.37: Parabolic bowls, $T = 2$

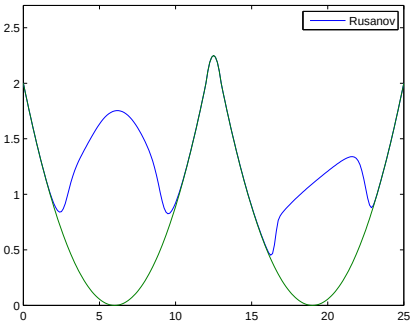

Figure 3.38: Parabolic bowls, $T = 3$

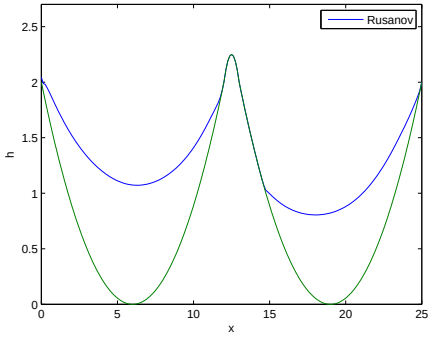

Figure 3.39: Parabolic bowls, $T = 4.5$

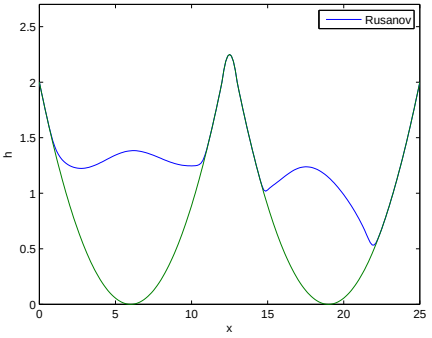

Figure 3.40: Parabolic bowls, $T = 6$

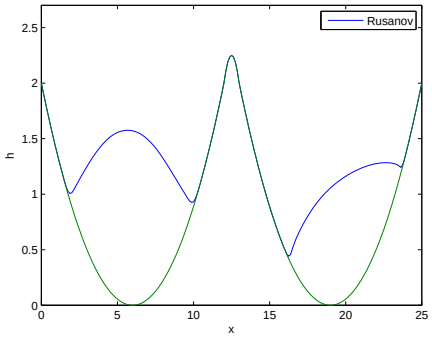

Figure 3.41: Parabolic bowls, $T = 10$

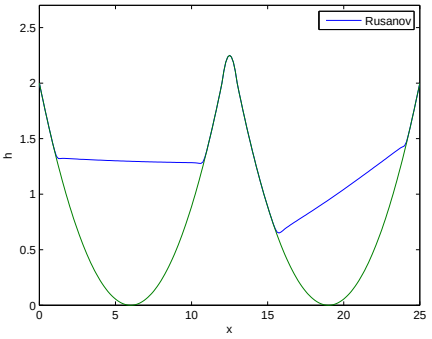

Figure 3.42: Parabolic bowls, $T = 100$

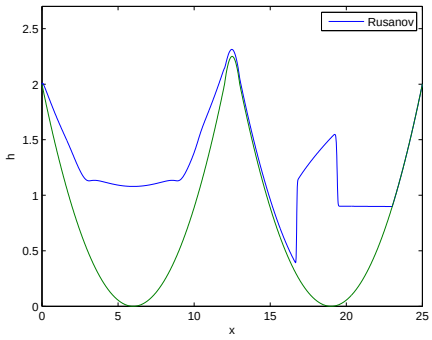

Figure 3.43: Reference solution, $T = 2$

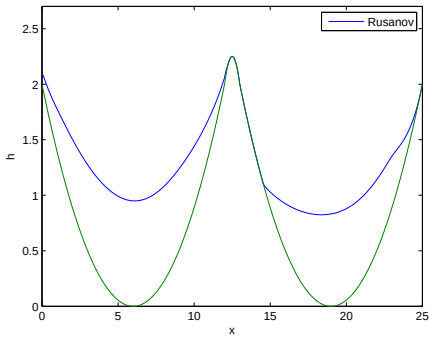

Figure 3.44: Reference solution, $T = 4.5$

Bibliography

[1] U. S. Fjordholm, S. Mishra, and E. Tadmor. Energy preserving and energy stable schemes for the shallow water equations. *Foundations of Computational Mathematics, Hong Kong 2008*, pages 93–139, 2009.

[2] U. S. Fjordholm, S. Mishra, and E. Tadmor. Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *Journal of Computational Physics*, 230(14):5587 – 5609, 2011.

[3] X. Y. Hu, N. A. Adams, and C.-W. Shu. Positivity-preserving flux limiters for high-order conservative schemes. *ArXiv e-prints*, March 2012.

[4] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[5] Y. Xing and C.-W. Shu. High-order finite volume weno schemes for the shallow water equations with dry states. *Advances in Water Resources*, v34:1026 – 1038, 2011.