Structure preserving finite volume methods for the shallow water equations

by

Ulrik Skre Fjordholm

THESIS
for the degree of

MASTER OF SCIENCE

(Master i Anvendt matematikk og mekanikk)

Department of Mathematics
Faculty of Mathematics and Natural Sciences
University of Oslo

May 2009

Preface

I wish to express my deepest gratitude to my supervisor Siddhartha Mishra for all his help and support through the past year. This thesis is a product of our cooperation. I am also grateful to all my friends in reading room B601 and B606 for making my time here at Blindern as enjoyable as it has been. Thanks in particular to Nils Henrik Risebro, Torstein Nilssen, John Christian Ottem, Nikolay Qviller and Robin Bjørnetun for proof reading the thesis.

Last, but not least, I thank Ingeborg for all the support and interest she has shown me.

Blindern, May 2009.

Contents

CHAPTER 1

Introduction

The most widely used model for fluid flows under the influence of gravity is the shallow water system

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

where h is the height of the water column, g is the gravitational constant and u and v are velocity in the x- and y-directions, respectively. This system of equations has applications in weather simulations, tidal waves, river and irrigation flows, tsunami prediction and more [**2, 25**]. The system is an example of a hyperbolic conservation law; others include the Buckley-Leverett equation of reservoir flow, the magneto-hydrodynamic equations of plasma physics and the Euler equations of gas dynamics. In most cases these are nonlinear and highly nontrivial to solve analytically, and as such, numerical methods must be applied to obtain approximate solutions. Due to the nonlinear nature of conservation laws, stability results of numerical methods are hard to obtain. Much work was done to this end in the 1980's by Osher, Harten, Lax and Tadmor, among others. Of the most fundamental concepts developed was that of entropy preservation and entropy stability. By asserting that schemes satisfy certain entropy inequalities, one obtains stability estimates and convergence towards the "correct" solution. In the case of the shallow water equations, the relevant entropy is the *energy* of the solution.

The main part of this thesis focuses on energy preserving and energy stable schemes for the shallow water equations. We develop a novel energy preserving, second-order accurate scheme that is very simple to implement, is computationally cheap and is stable compared to other existing energy preserving schemes [**11**]. To allow for a correct dissipation of energy in the vicinity of shocks, a novel numerical diffusion operator of the Roe type [**31, 32**] is designed. The energy preserving scheme, together with this diffusion operator, gives an energy stable scheme for the shallow water system. We apply a standard reconstruction procedure to obtain a second-order accurate scheme.

In the presence of a varying bottom topography, appropriate source terms must be added to the shallow water system. This gives the non-homogeneous set of equations

$$
\begin{aligned}
h_t + (hu)_x + (hv)_y &= 0, \\
(hu)_t + \left(hu^2 + \frac{1}{2} g h^2 \right)_x + (huv)_y &= -ghb_x, \\
(hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2} g h^2 \right)_y &= -ghb_y,
\end{aligned}
\tag{1.2}
$$

where $b = b(x, y)$ gives the elevation of the bottom from some absolute level. An important point of study for non-homogeneous conservation laws such as the above are steady states, solutions that are constant in time. The most natural steady state for (1.2) is the lake at rest, where there is no flow at all. More general, moving steady states also exist. As many interesting flows are merely perturbations of a steady state, the ability of numerical methods for correctly computing steady states is essential. Such schemes are called *well-balanced*, since the early works of LeRoux and coworkers [**16**]. With a compatible discretization of the source term in (1.2), we found that both our energy preserving and energy stable schemes are well-balanced [**12**].

In addition to well-balanced energy preserving and energy stable schemes, we consider a method for controlling vorticity errors in numerical approximations of the shallow water equations. The new method shows promise in problems where there is a strong interaction between the flow in the $x-$ and $y-$directions [**13**]. A similar approach has been considered earlier in the case of the Euler equations by Ismail and Roe [**20**].

This thesis is organized as follows. In the rest of the present chapter we derive the shallow water equations from basic principles. We review some theory on hyperbolic conservation laws, and in Section 1.3 we give a short introduction to finite volume methods.

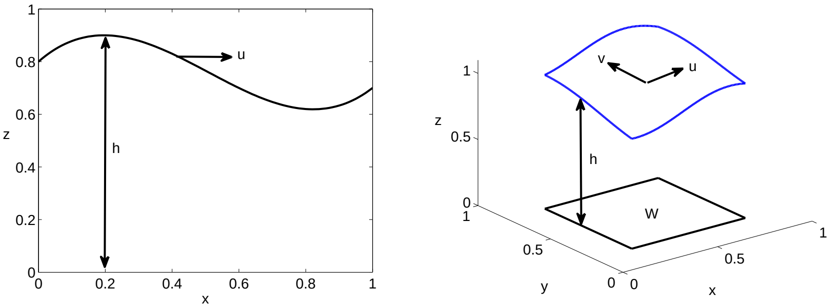
Chapter 2 details the construction of our new energy preserving and energy stable schemes. These are compared with other schemes in a number of numerical experiments. We also derive a second-order accurate version of the energy stable scheme. In Chapter 3, we investigate the well-balancing properties of the new schemes.

In Chapter 4, we derive a scheme for the correct prediction of flows with strong vortical forces.

All numerical methods presented in this thesis have been implemented in C++ using the linear algebra library Blitz++. The source code can be downloaded from `http://folk.uio.no/ulriksf/solver.zip`.

1.1. The shallow water equations

We model the flow of a fluid for which the depth is relatively small compared to the length scales, as is illustrated in Figure 1.1(b). If the variation in, say, the y-direction is small, then the problem is essentially one-dimensional, as in Figure 1.1(a). First, we derive the one-dimensional shallow water equations, and then provide the generalization to two spatial dimensions.



(a) One-dimensional

(b) Two-dimensional

Figure 1.1. Intersection of fluid profile.

Since the height of the fluid is small, we may assume that the velocity of the fluid is constant in the vertical direction, $u(x, z, t) \equiv u(x, t)$. Furthermore, we assume that the density ρ of the fluid is constant. The mass of the fluid over an interval $[x_1, x_2]$ is given by

$$\int_{x_1}^{x_2} \int_0^{h(x,t)} \rho \; dzdx = \int_{x_1}^{x_2} \rho h(x,t) \; dx.$$

If there is no creation or destruction of mass, then this quantity can only change due to fluid flow through the boundary of the interval. Denoting the rate of flow, the *flux*, over a point x by $f(x, t)$, then this means that

$$\frac{d}{dt} \int_{x_1}^{x_2} \rho h(x,t) \; dx = f(x_1, t) - f(x_2, t). \tag{1.3}$$

Since the fluid is transported, or *advected*, along the velocity field, the flux over a point x is given by the mass density times the velocity. Thus,

$$f(x,t) = \int_0^{h(x,t)} \rho \; dz \cdot u(x,t) = \rho h(x,t) u(x,t).$$

If h and u are differentiable functions of x and t, then (1.3) can be written as

$$\int_{x_1}^{x_2} \rho h_t(x,t) \; dx = - \int_{x_1}^{x_2} \rho (hu)_x(x,t) \; dx.$$

(We use the notational convention $h_t = \frac{\partial h}{\partial t}, h_x = \frac{\partial h}{\partial x}$, etc.) Since x_1 and x_2 were arbitrarily chosen, this implies that

$$h_t + (hu)_x = 0, \tag{1.4}$$

which we refer to as *conservation of mass*. We say that h is a *conserved variable*.

Momentum is also advected by the velocity u, so the flux of momentum due to advection is

$$(\rho h u) \cdot u = \rho h u^2.$$

In addition to advection, momentum is affected by the pressure p in the fluid. Therefore the flux of momentum past a point x is

$$f(x,t) = \rho h(x,t)u(x,t)^2 + p(x,t).$$

By the assumption of *hydrostatic balance*, the pressure at a point (x,z) is simply the force of gravity exerted upon the water above that point. Thus, the total amount of pressure above x is

$$p(x,t) = \int_0^{h(x,t)} \rho g\big(h(x,t) - z\big)\ dz = \frac{1}{2}\rho g h(x,t)^2.$$

Assuming differentiability as before, we get *conservation of momentum*

$$(hu)_t + \left(\frac{1}{2}gh^2 + hu^2\right) = 0.$$

Combining this with (1.4), we obtain the one-dimensional *shallow water equations*

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + \left(\frac{1}{2}gh^2 + hu^2\right)_x &= 0. \end{aligned} \tag{1.5}$$

We will always assume that h is positive, as negative height is meaningless. Actually, it may be shown that if h is nonnegative at the initial time, then it will stay nonnegative at all times.

The derivation of (1.5) can be generalized to the two-dimensional setup seen in Figure 1.1(b). We only give the derivation of conservation of mass. Let $W \subset \mathbb{R}^2$ be any open, connected set. Then the rate of change of mass over W exactly balances the flux through the boundary ∂W. Thus, denoting the flux over a point $\mathbf{x} \in \mathbb{R}^2$ by $\mathbf{f}(\mathbf{x},t) = (f(\mathbf{x},t),\ g(\mathbf{x},t))$, we have

$$\frac{d}{dt}\int_W \rho h(\mathbf{x},t)\ d\mathbf{x} = -\int_{\partial W} \mathbf{f}(\mathbf{x},t)\cdot \nu\ dS(\mathbf{x}),$$

where $\nu = (\nu^x, \nu^y)$ is the unit normal vector on ∂W. Integrating by parts, this identity is equivalent to

$$\frac{d}{dt}\int_W \rho h(\mathbf{x},t)\ d\mathbf{x} = -\int_W f_x(\mathbf{x},t) + g_y(\mathbf{x},t)\ dS(\mathbf{x}).$$

Since W was arbitrary, this implies that

$$\rho h_t + f_x + g_y = 0.$$

As for the one-dimensional setup, mass is advected with the fluid, so writing $\mathbf{u} = (u,v)$ for the velocity field, we have $\mathbf{f} = \rho h\mathbf{u}$. Thus,

$$h_t + (hu)_x + (hv)_y = 0.$$

Applying a similar argument to the momentum in each direction, we obtain the two-dimensional shallow water system (1.1).

1.2. Conservation laws

The flux function of the one-dimensional shallow water equations (1.5) depends on x and t only through h and u. Hence, the system can be written more compactly as a *conservation law* [**19**], which in general is of the form

$$U_t + f(U)_x = 0 \qquad \text{for } x \in \mathbb{R} \text{ and } t > 0. \tag{1.6}$$

The solution $U = U(x,t) : \mathbb{R}\times[0,\infty) \to \mathbb{R}^n$ is referred to as the *conserved variables.* In the case of the shallow water equations (1.5), we have $U = [h, hu]^\top$ and

$$f(U) = \begin{bmatrix} hu \\ \frac{1}{2}gh^2 + hu^2 \end{bmatrix}.$$

Integrating the conservation law (1.6) over an interval (x_1, x_2), we see that the total amount of quantity over the interval can only change due to the flux through the endpoints:

$$\frac{d}{dt}\int_{x_1}^{x_2} U(x,t)\; dx = f(U(x_1,t)) - f(U(x_2,t)). \tag{1.7}$$

Thus, conservation laws model quantities – in this case mass and momentum – that are preserved over time. Any change in the quantity in a domain is caused solely by the flux through the boundary.

The conservation law (1.6) can be written in the quasi-linear form

$$U_t + f'(U)U_x = 0, \tag{1.8}$$

where $f'(U)$ denotes the Jacobian of f. We say that the conservation law is *hyperbolic* if $f'(U)$ has n real distinct eigenvalues [**19**]. The shallow water system is hyperbolic; indeed, $f'(U)$ has eigenvalues

$$\lambda_1 = u - \sqrt{gh} \qquad \text{and} \qquad \lambda_2 = u + \sqrt{gh},$$

which are real and distinct since h is positive. Other examples of hyperbolic conservation laws include the wave equation, the magneto-hydrodynamics equations, the Einstein equation and the Euler equations [**26**].

The property of hyperbolicity implies a finite speed of propagation: Information travels at a finite speed given by the eigenvalues of $f'(U)$. This is a fundamental property of these PDE. As a consequence of this property, the task of solving an initial value problem can be tackled by dividing the spatial domain into smaller domains, solving the resulting *independent* problems and then patching these together. This may be contrasted with parabolic equations such as the heat equation, where information travels at an infinite speed. Hence, any local change in initial data implies an immediate global change in the solution.

The two-dimensional shallow water system (1.1) can be written in the similar form

$$U_t + f(U)_x + g(U)_y = 0, \tag{1.9}$$

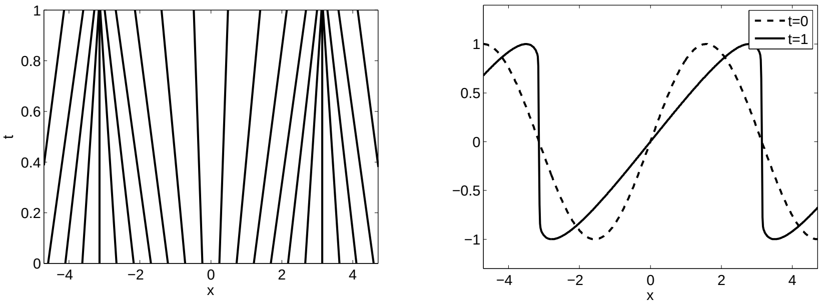
for $U = [h, hu, hv]^\top$. In one spatial dimension, the condition of hyperbolicity of a conservation law is simply that $f'(U)$ has real and distinct eigenvalues. For multidimensional conservation laws such as (1.9), we require the same for $f'(U)$ and $g'(U)$, but in addition we require that all nontrivial linear combinations of $f'(U)$ and $g'(U)$ have real and distinct eigenvalues. It is a straight-forward calculation to show that (1.1) is hyperbolic in this sense [**26**]. Specifically, $f'(U)$ and $g'(U)$ have eigenvalues

$$\begin{aligned} \lambda_1 &= u - \sqrt{gh}, & \lambda_2 &= u, & \lambda_3 &= u + \sqrt{gh}, \\ \mu_1 &= v - \sqrt{gh}, & \mu_2 &= v, & \mu_3 &= v + \sqrt{gh}. \end{aligned} \tag{1.10}$$

We will first develop the theory for the one-dimensional conservation law and then generalize it to the multi-dimensional case.

Of the most interesting properties of hyperbolic conservation laws is the appearance of discontinuities or *shocks* in even the simplest of problems. To demonstrate this, we will solve an initial value problem for the canonical scalar, hyperbolic conservation law called Burgers' equation [**26**]

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \tag{1.11}$$



(a) A selection of characteristics.

(b) The solution at $t = 1$.

FIGURE 1.2. The solution of Burgers' equation with initial data $u_0(x) = \sin(x)$.

with initial data $u(x, 0) = u_0(x)$. Carrying out differentiation of $f(u)$ in (1.11), we obtain the equivalent PDE

$$u_t + uu_x = 0. \tag{1.12}$$

We solve this equation by the method of characteristics [**19**]. Let u solve (1.12), and for each $x_0 \in \mathbb{R}$, let the *characteristic* $x(t)$ solve

$$\begin{cases} x'(t) = f'(u(x(t), t)) = u(x(t), t) & \text{for } t > 0 \\ x(0) = x_0. \end{cases} \tag{1.13}$$

Then $\frac{d}{dt}u(x(t), t) = 0$, so u is constant along characteristics. Hence information is transported along the characteristics with speed equal to $u(x(t), t) = u_0(x_0)$, the eigenvalue of $f'(u_0(x_0))$. The solution of (1.13) is therefore

$$x(t) = tu_0(x_0) + x_0.$$

Solving this equation for x_0 then gives a solution

$$u(x, t) = u_0(x_0(x, t))$$

of (1.12).

The main assumption of this derivation is that there is a unique characteristic through each given point (x, t) in the x-t-plane. This is an incorrect assumption; indeed, solving $x(t) = \widetilde{x}(t)$ for two characteristics x and $\widetilde{x}$ gives a solution

$$t = -\frac{1}{\min\limits_{x} u_0'(x)}.$$

Hence, if $u_0'(x_0) < 0$ for some x_0, then there will be an intersection of characteristics. At the point of intersection the solution is multi-valued; it is both equal to $u_0(x(0))$ and $u_0(\widetilde{x}(0))$. This is illustrated in the following example. We solve Burgers' equation with initial data $u_0(x_0) = \sin(x_0)$. A selection of characteristics is plotted in Figure 1.2(a). Along each characteristic $x(t)$, the solution is constant, equal to u_0 at the point $x(0)$. At time $t = 1$, there is an intersection of characteristics, and at this point, the solution is both equal to 1 and −1.

The problem of multi-valuedness is resolved by the formation of a shock – a discontinuity in the solution. This is illustrated in Figure 1.2(b), where the solution of Burgers' equation at $t = 1$ is shown. To allow for discontinuous solutions of a differential equation, one considers the equation in a weak sense.

DEFINITION 1.1. A function $U \in L^1_{\mathrm{loc}}(\mathbb{R} \times [0,\infty))$ is a *weak solution* of (1.6) if

$$\int_0^\infty \int_{\mathbb{R}} U\varphi_t + f(U)\varphi_x \; dxdt + \int_{\mathbb{R}} \varphi(x,0)U(x,0) \; dx = 0$$

for all $\varphi \in C_c^\infty(\mathbb{R} \times [0,\infty))$.

While this formulation allows the presence of discontinuities, uniqueness of solutions may not be possible without adding further *admissibility* criteria.

1.2.1. Existence and uniqueness. The mechanism that counters the formation of large gradients in nature is *viscosity.* Viscosity is taken into account in the more fundamental equation

$$U_t^\varepsilon + f(U^\varepsilon)_x = \varepsilon U_{xx}^\varepsilon \tag{1.14}$$

for an $\varepsilon > 0$ [**14**]. Under certain assumptions, this equation has a unique smooth solution. Letting $\varepsilon \to 0$ one may obtain, using a compactness argument, a solution of (1.6). A solution that is the limit of such functions is called a *vanishing viscosity solution* and is considered to be the physically relevant solution of the conservation law [**14**].

It may be hard to verify that a given weak solution is the vanishing viscosity solution. However, it is known that viscosity is closely related to the concept of entropy, a measure of disorder in an isolated system. By the second law of thermodynamics, entropy is nondecreasing in time; by a difference in sign, mathematical entropy is non*increasing* in time. In the context of conservation laws, entropy should be preserved in smooth regions of the solution and be dissipated (decrease) around shocks. Thus, for a smooth solution, the entropy should be a conserved variable. We formalize these ideas as follows.

Let some function $E = E(U) : \mathbb{R}^n \to \mathbb{R}$ be given as a measure of the entropy present in a solution U of the conservation law. Under what conditions does E satisfy its own conservation law

$$E(U)_t + Q(U)_x = 0 \tag{1.15}$$

for some entropy flux function Q? By left-multiplying the equivalent form of the conservation law (1.8) by $E'(U)^\top$, we obtain

$$E(U)_t + E'(U)^\top f'(U)U_x = 0.$$

($E'(U)^\top$ denotes the transpose of the column vector $E'(U)$, so left-multiplying by $E'(U)^\top$ is the same as taking the Euclidean inner product with $E'(U)$.) Hence, (1.15) may be satisfied if and only if there exists a Q such that

$$Q'(U)^\top = E'(U)^\top f'(U).$$

DEFINITION 1.2. An *entropy pair* for the conservation law (1.6) is a pair (E, Q) consisting of a convex function $E : \mathbb{R}^n \to \mathbb{R}$ and a function $Q : \mathbb{R}^n \to \mathbb{R}$ such that

$$Q'(U)^\top = E'(U)^\top f'(U) \qquad \text{for all } U \in \mathbb{R}^n.$$

Recall that convexity of a function $E : \mathbb{R}^n \to \mathbb{R}$ means that the matrix $E''(U)$, the Hessian of E, is a positive matrix. The condition of convexity will soon be justified.

If the conservation law is scalar, then any convex differentiable function $E : \mathbb{R} \to \mathbb{R}$ gives rise to an entropy pair; just let

$$Q(U) = \int_0^U E'(s)f'(s) \; ds.$$

For instance, for the Burgers' equation (1.11), the entropy $E(u) = \frac{u^2}{2}$ gives an entropy flux $Q(U) = \int_0^u s^2 \; ds = \frac{u^3}{3}$. For systems, the process of finding entropy pairs is more difficult, although many conservation laws are equipped with rather natural entropies. In the linear case $f(U) = AU$ for a symmetric matrix A, one entropy pair is the *energy* $E(U) = \frac{1}{2}U^\top U$, with entropy flux $Q(U) = \frac{1}{2}U^\top AU$.

For the shallow water equations, the entropy of interest is the *total energy.* This is given by $E(U) = \frac{1}{2}(hu^2 + gh^2)$, and is the sum of kinetic and potential energy. It is readily verified that

$$E(h, hu) = \frac{1}{2}(hu^2 + gh^2) \qquad \text{and} \qquad Q(h, hu) = \frac{1}{2}hu^3 + guh^2 \tag{1.16}$$

is an entropy pair for the shallow water equations (1.5).

The calculations carried out to obtain the entropy equality (1.15) are not valid when discontinuities are present in the solution. To obtain a similar result for weak solutions, we must realize the solution as the limit of solutions of the viscous problem (1.14). Assume that an entropy pair (E, Q) of the conservation law is given. By multiplying the equation (1.14) by $E'(U^\varepsilon)^\top$ on both sides, we get

$$\begin{aligned} E(U^\varepsilon)_t + Q(U^\varepsilon)_x &= \varepsilon E'(U^\varepsilon)^\top U^\varepsilon_{xx} \\ &= \varepsilon \left(E(U^\varepsilon)_{xx} - (U^\varepsilon_x)^\top E''(U^\varepsilon)U^\varepsilon_x\right) \\ &\leq \varepsilon E(U^\varepsilon)_{xx}, \end{aligned}$$

the inequality following from the positivity of $E''(U^\varepsilon)$. In the limit $\varepsilon \to 0$, the right-hand side vanishes (formally), and we obtain the *entropy inequality*

$$E(U)_t + Q(U)_x \leq 0. \tag{1.17}$$

This is to be interpreted in the sense of distributions:

$$\int_0^\infty \int_{\mathbb{R}} E(U)\varphi_t + Q(U)\varphi_x \; dxdt + \int_{\mathbb{R}} E(x, 0)\varphi(x, 0) \; dx \geq 0$$

for all $\varphi \in C_c^\infty(\mathbb{R} \times [0, \infty))$ with $\varphi \geq 0$.

Definition 1.3. We say that a weak solution U of the conservation law satisfies the *entropy condition* for an entropy pair (E, Q) if (1.17) holds.

In domains where the solution is smooth, the entropy equality (1.15) must necessarily hold. The entropy inequality (1.17) is only relevant near discontinuities in U or its derivatives.

The concept of entropy has proven important for existence and uniqueness of conservation laws. In [**23**], Kruzkov showed that for a scalar equation, even in several dimensions, there is a unique weak solution that satisfies the entropy condition for a specific class of entropy pairs. A more general result for systems of conservation laws in one dimension was proved by Lax [**24**]. He showed that a solution satisfies the entropy condition for a strictly convex entropy function whenever it is the physically correct solution. More importantly, the entropy condition is a sufficient condition whenever the initial data is “sufficiently small”.

1.3. Finite volume methods

As the task of solving a conservation law explicitly is tractable only in the simplest of cases, we must employ numerical methods when solving more realistic problems. Of the available approaches, the finite volume methods are the most successful [**26**]. These methods rely on partitioning the computational domain into a finite set of control volumes, over which the conserved variables are averaged. This average form allows the presence of discontinuities in the solution – contrary to finite difference schemes, which rely on a pointwise approximation of the PDE.

The domain is divided into intervals

$$I_i = \left[x_{i-1/2},\ x_{i+1/2}\right) \qquad \text{for } i \in \mathbb{Z}$$

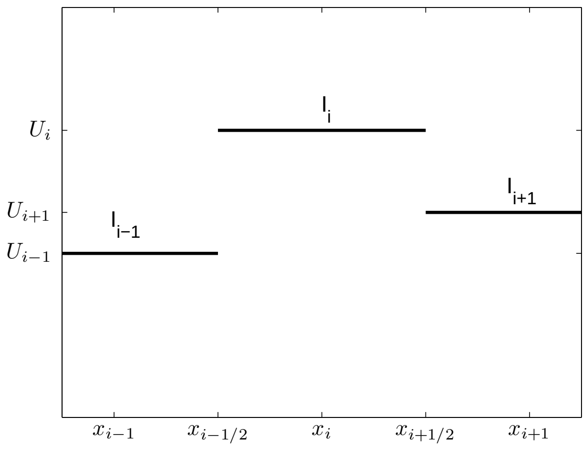


FIGURE 1.3. Division of $\mathbb{R}$ into control sets I_i.

of length $\Delta x_i = x_{i+1/2} - x_{i-1/2}$, and $U(x,t)$ is approximated by a piecewise constant function

$$U_i(t) = \frac{1}{\Delta x_i} \int_{I_i} U(x,t)\, dx$$

(see Figure 1.3). By (1.7), U_i satisfies

$$\frac{d}{dt} U_i(t) = -\frac{1}{\Delta x} \big(f(U(x_{i+1/2},t)) - f(U(x_{i-1/2},t)) \big). \tag{1.18}$$

Integrating this equation from t to $t + \Delta t$ for some time step $\Delta t > 0$, we find an expression for U_i at the next time step,

$$U_i(t+\Delta t) = U_i(t) - \frac{1}{\Delta x} \int_t^{t+\Delta t} \big(f(U(x_{i+1/2},s)) - f(U(x_{i-1/2},s)) \big)\, ds.$$

However, this expression relies explicitly on the exact solution U, and so must be approximated appropriately.

The following algorithm is due to Godunov [**15**]. At the initial time $t = 0$, we average the initial data over each grid cell I_i, obtaining a piecewise constant approximation $U_i(0)$. At each cell interface $x_{i+1/2}$ we solve the *Riemann problem*

$$U^{\mathrm{RP}}(x,0) = \begin{cases} U_i & \text{if } x < x_{i+1/2} \\ U_{i+1} & \text{if } x > x_{i+1/2}. \end{cases} \tag{1.19}$$

This is a well-defined initial value problem that may be solved exactly. We obtain an approximate solution at the next time

$$U_i(\Delta t) = U_i(0) - \frac{1}{\Delta x} \int_0^{\Delta t} \Big(f\big(U^{\mathrm{RP}}(x_{i+1/2},s)\big) - f\big(U^{\mathrm{RP}}(x_{i-1/2},s)\big) \Big)\, ds. \tag{1.20}$$

The process described above is repeated iteratively until t has reached some end time $t_{\max}$.

The solution of the Riemann problem consists of a set of at most n waves emanating from $x = x_{i+1/2}$. The speed of propagation of each of these waves in the x-t-plane is bounded by the maximum eigenvalue of $f'(U_i)$ and $f'(U_{i+1})$. To keep waves from different Riemann problems from interacting with each other, we select Δt so that the *CFL condition*

$$c\frac{\Delta t}{\Delta x} \leq 1,$$

where

$$c = \max_{\substack{k \\ x\in\mathbb{R}}} |\lambda_k(U(x,0))|,$$

is satisfied. The quantity $c\frac{\Delta t}{\Delta x}$ is denoted as the *CFL number* in the remainder; the CFL condition requires the CFL number to be less than 1.

In general, it is hard to solve the Riemann problem exactly. Instead, one considers so-called approximate Riemann solvers. In the formula (1.20), we replace $f(U^{\mathrm{RP}}(x_{i+1/2}, s))$ by a *numerical flux function* $F_{i+1/2} = F(U_i, U_{i+1})$. The semi-discrete form of the scheme is then

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(F_{i+1/2} - F_{i-1/2}\right). \tag{1.21}$$

Note that this updating formula only depends on $U(\cdot, t_k)$ in the cells I_{i-1}, I_i and I_{i+1}; thus, the scheme is a *3-point scheme.* More general $(2p+1)$-point schemes exist [**14**], and will indeed be considered later when we increase the order of accuracy of the scheme.

The specific form of the semi-discrete formula (1.21) is not incidental. A finite volume method that may be written in this form is called *conservative* [**26**]. Conservation implies that the flow into a grid cell I_{i+1} from the left exactly balances the flow out of the cell I_i to the right; they are both equal to $F_{i+1/2}$. Thus, a discrete conservation of the quantity U is obtained. Indeed, by summing (1.21) over $i = k, \dots, l$, we obtain

$$\frac{d}{dt}\left(\Delta x \sum_{i=k}^{l} U_i\right) = F_{k+1/2} - F_{l-1/2},$$

a discrete version of (1.7).

There lies great freedom in choosing the numerical flux $F_{i+1/2}$, but to ensure that the scheme approximates the correct equation, a certain consistency criterion is imposed. F is termed *consistent* with the conservation law (1.6) if $F(U,U) = f(U)$ for all $U \in \mathbb{R}^n$. Thus, when the solution in two neighboring grid cells are equal, the flux through the cell interface should exactly equal the flux given by the conservation law.

By integrating (1.21) over $t \in (t_k, t_k + \Delta t_k)$ and applying any numerical method of integration, we find a solution $U_i(t_{k+1})$ at the next time step. This integral may be discretized with a number of methods. Perhaps the simplest one is the first-order accurate forward Euler method. Writing $U_i^k = U_i(t_k)$ and $\mathcal{L}(U_i^k)$ for the right-hand side of (1.21), this method is given by

$$U_i^{k+1} = U_i^k + \Delta t_k \mathcal{L}(U_i^k).$$

In this thesis we also use the second- and third-order accurate, non-oscillatory Runge-Kutta methods [**17**]

$$\begin{aligned} U_i^* &= U_i^k + \Delta t_k \mathcal{L}(U_i^k), \\ U_i^{**} &= U_i^* + \Delta t_k \mathcal{L}(U_i^*), \\ U_i^{k+1} &= \frac{1}{2}(U_i^k + U_i^{**}) \end{aligned} \tag{1.22}$$

and

$$\begin{aligned} U_i^* &= U_i^k + \Delta t_k \mathcal{L}(U_i^k), \\ U_i^{**} &= \frac{3}{4}U_i^k + \frac{1}{4}U_i^* + \frac{\Delta t}{4}\mathcal{L}(U_i^*), \\ U_i^{k+1} &= \frac{1}{3}U_i^k + \frac{2}{3}U_i^{**} + \frac{2\Delta t}{3}\mathcal{L}(U_i^{**}). \end{aligned} \tag{1.23}$$

1.3.1. Two-dimensional conservation laws. The numerical method outlined above may be generalized to conservation laws in two (and more) spatial dimensions (1.9). The computational domain is divided into grid cells

$$I_{i,j} = \left[x_{i-1/2}, x_{i+1/2}\right) \times \left[y_{j-1/2}, y_{j+1/2}\right)$$

with $x_{i+1/2} - x_{i-1/2} \equiv \Delta x$ and $y_{j+1/2} - y_{j-1/2} \equiv \Delta y$. The solution U is approximated by the cell average

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

and the numerical method is written in the form

$$\frac{d}{dt} U_{i,j} = -\frac{1}{\Delta x}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) - \frac{1}{\Delta y}\left(G_{i,j+1/2} - G_{i,j-1/2}\right) \tag{1.24}$$

for numerical fluxes $F_{i+1/2,j} = F(U_{i,j}, U_{i+1,j})$ and $G_{i,j+1/2} = G(U_{i,j}, U_{i,j+1})$. These are approximate Riemann solvers of the problems

$$\begin{cases} U_t + f(U)_x = 0 \\ U(x,0) = \begin{cases} U_{i,j} & \text{if } x < x_{i+1/2} \\ U_{i+1,j} & \text{if } x > x_{i+1/2} \end{cases} \end{cases}$$

and

$$\begin{cases} U_t + g(U)_y = 0 \\ U(y,0) = \begin{cases} U_{i,j} & \text{if } y < y_{j+1/2} \\ U_{i,j+1} & \text{if } y > y_{j+1/2}, \end{cases} \end{cases}$$

respectively. F and G are said to be *consistent* with the conservation law (1.9) if we have $F(U,U) = f(U)$ and $G(U,U) = g(U)$ for all $U \in \mathbb{R}^n$.

1.3.2. Two standard schemes. Two well-known finite volume fluxes are the Rusanov and the Roe fluxes. We include them here for reference. See also [**26**].

Rusanov: The Rusanov (or Local Lax-Friedrichs) flux is

$$F^{\text{Rus}}_{i+1/2} = \frac{1}{2}\left(f(U_i) + f(U_{i+1})\right) - \frac{1}{2} c_{i+1/2}(U_{i+1} - U_i), \tag{1.25}$$

where

$$c_{i+1/2} = \max_k \{|\lambda_k(U_i)|, |\lambda_k(U_{i+1})|\}$$

is an estimate of the local wave speed.

Roe: The Roe flux relies on a linearization of the conservation law around a point $\widetilde{U}$. It is given by

$$F^{\text{Roe}}_{i+1/2} = \frac{1}{2}\left(f(U_i) + f(U_{i+1})\right) - \frac{1}{2} R|\Lambda|R^{-1}(U_{i+1} - U_i), \tag{1.26}$$

where R is the matrix of eigenvectors of $f'(\widetilde{U})$ and

$$|\Lambda| = \operatorname{diag}\left(|\lambda_1(\widetilde{U})|, \dots, |\lambda_n(\widetilde{U})|\right).$$

The state $\widetilde{U}$ suggested in Roe's paper [**32**] called the *Roe average* ensures that isolated single shocks are resolved exactly. For the shallow water equations, this state is given by $\widetilde{U} = (\widetilde{h}, \widetilde{h}\widetilde{u})$ with

$$\widetilde{h} = \frac{h_i + h_{i+1}}{2} \qquad \text{and} \qquad \widetilde{u} = \frac{\sqrt{h_i}u_i + \sqrt{h_{i+1}}u_{i+1}}{\sqrt{h_i} + \sqrt{h_{i+1}}}. \tag{1.27}$$

1.3.3. Further notation. For a grid function u_i we define the difference and average operators

$$[\![u]\!]_{i+1/2} = u_{i+1} - u_i \qquad \text{and} \qquad \overline{u}_{i+1/2} = \frac{u_i + u_{i+1}}{2}.$$

The following identities are readily verified:

$$\begin{aligned} [\![uv]\!]_{i+1/2} &= \overline{u}_{i+1/2}[\![v]\!]_{i+1/2} + \overline{v}_{i+1/2}[\![u]\!]_{i+1/2}, \\ \overline{u^2}_{i+1/2} - \overline{u^2}_{i-1/2} &= \overline{u}_{i+1/2}[\![u]\!]_{i+1/2} + \overline{u}_{i-1/2}[\![u]\!]_{i-1/2}. \end{aligned} \tag{1.28}$$

For a two-dimensional grid function $u_{i,j}$ we write

$$\begin{aligned} [\![u]\!]_{i+1/2,j} &= u_{i+1,j} - u_{i,j}, & [\![u]\!]_{i,j+1/2} &= u_{i,j+1} - u_{i,j}, \\ \overline{u}_{i+1/2,j} &= \frac{u_{i,j} + u_{i+1,j}}{2}, & \overline{u}_{i,j+1/2} &= \frac{u_{i,j} + u_{i,j+1}}{2}. \end{aligned}$$

CHAPTER 2

Energy preserving and energy stable schemes

2.1. Introduction

The two most fundamental questions for any numerical approximation is: Does the scheme converge as the grid is refined, and if it does, is the limit the correct solution? In our framework, those questions are: Does the solutions obtained by a finite volume method (1.21) converge to a weak solution of the conservation law (1.6), and furthermore, is that limit the physically relevant solution? The first question was answered by Lax and Wendroff [**14**].

Theorem 2.1 (Lax-Wendroff). *If a sequence $\{U^{(k)}\}_k$ of solutions computed by a consistent and conservative scheme with grid sizes $\Delta t^{(k)}$ and $\Delta x^{(k)}$ converges boundedly a.e. as $\Delta t^{(k)}, \Delta x^{(k)} \to 0$, then the limit is a weak solution.*

Thus consistency and conservation is essential to ensure that the limit is a solution of the conservation law. Note that this theorem does *not* address the question of whether the scheme will at all converge, only that if it does, then the limit is a weak solution.

The second question is answered in the form of a discrete version of the entropy condition. Given an entropy pair (E, Q), we look for numerical methods whose solutions satisfy a discrete form of the entropy inequality (1.17),

$$\frac{d}{dt}E(U_i(t)) + \frac{1}{\Delta x}\left(\widehat{Q}_{i+1/2} - \widehat{Q}_{i-1/2}\right) \leq 0 \qquad \text{for all } i \in \mathbb{Z} \text{ and } t > 0 \tag{2.1}$$

for a numerical entropy flux $\widehat{Q}_{i+1/2} = \widehat{Q}\big(U_i(t), U_{i+1}(t)\big)$ that is consistent with Q. The numerical entropy flux $\widehat{Q}$ is assumed to be continuous as a function $\widehat{Q} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. The following theorem – whose idea and proof are very similar to Lax' and Wendroff's – shows that the discrete entropy condition is sufficient for the limit function to be the physically correct solution.

Theorem 2.2 (Osher [**31**]). *Let $\{U_i^{(k)}(t)\}_k$ be a sequence of functions computed by a consistent and conservative scheme that converges boundedly a.e. to a function U. Assume that there is an entropy flux $\widehat{Q}$, consistent with Q, such that* (2.1) *is satisfied for every $U^{(k)}$. Then U is a weak solution of* (1.6) *that satisfies the entropy condition for (E, Q).*

Proof. The fact that the limit is a weak solution follows from the Lax-Wendroff theorem.

Let $k \in \mathbb{N}$; we write $U_i = U_i^{(k)}$. Let $\varphi \in C_c^\infty(\mathbb{R} \times [0, \infty))$ with $\varphi \geq 0$ and denote $\widetilde{\varphi}_i = \widetilde{\varphi}_i(t) = \varphi(x_i, t)$. By multiplying (2.1) by $\Delta x \widetilde{\varphi}_i$, summing over $i \in \mathbb{Z}$ and integrating over $t \in [0, \infty)$, we obtain

$$\Delta x \sum_{i=-\infty}^{\infty} \int_0^\infty \left(\frac{d}{dt}E(U_i)\widetilde{\varphi}_i + \frac{1}{\Delta x}\left(\widehat{Q}_{i+1/2} - \widehat{Q}_{i-1/2}\right)\widetilde{\varphi}_i\right) dt \leq 0.$$

Integrating and summing by parts, this becomes

$$\Delta x \sum_{i=-\infty}^{\infty} E(U_i(0))\widetilde{\varphi}_i(0) + \Delta x \sum_{i=-\infty}^{\infty} \int_0^\infty \left(E(U_i)\frac{d}{dt}\widetilde{\varphi}_i + \widehat{Q}_{i+1/2}\frac{\widetilde{\varphi}_{i+1} - \widetilde{\varphi}_i}{\Delta x}\right) dt \geq 0.$$

We may view U_i and $\widetilde{\varphi}_i$ as step functions on $\mathbb{R} \times [0, \infty)$, letting $\widetilde{U}(x) = U_i$ and $\widetilde{\varphi}(x) = \widetilde{\varphi}_i$ for $x \in I_i$. Then the above can be written as

$$\begin{aligned}\int_{\mathbb{R}} E\big(\widetilde{U}(x,0)\big)\widetilde{\varphi}(x,0)\;dx + \int_{\mathbb{R}}\int_0^\infty E\big(\widetilde{U}(x,t)\big)\frac{d}{dt}\widetilde{\varphi}(x,t)\;dtdx \\ + \int_{\mathbb{R}}\int_0^\infty \widehat{Q}\left(\widetilde{U}(x), \widetilde{U}(x+\Delta x)\right)\frac{\widetilde{\varphi}(x+\Delta x)-\widetilde{\varphi}(x)}{\Delta x}\;dtdx \geq 0.\end{aligned} \tag{2.2}$$

Both $\widetilde{U}$ and $\widetilde{\varphi}$ converge boundedly a.e. to U and φ, respectively, so by Lebesgue's dominated convergence theorem, the first two terms converge to

$$\int_{\mathbb{R}} E(U(x,0))\varphi(x,0)\;dx + \int_{\mathbb{R}}\int_0^\infty E(U(x,t))\varphi_t(x,t)\;dtdx$$

as $k \to \infty$. Since $\widehat{Q}$ is continuous and consistent with Q, the third term converges to

$$\int_{\mathbb{R}}\int_0^\infty Q(U(x,t))\varphi_x(x,t)\;dtdx.$$

Thus the left-hand side of (2.2) converges to

$$\int_{\mathbb{R}}\int_0^\infty E(U)\varphi_t + Q(U)\varphi_x\;dtdx + \int_{\mathbb{R}} E(U(x,0))\varphi(x,0)\;dx,$$

thus proving the inequality. □

This result gives a motivation for looking for schemes that satisfy the discrete entropy inequality (2.1). However, the theorem is not a constructive one; it gives no hint to how such a scheme may be designed. This problem was tackled by Tadmor in 1984 for scalar equations and was generalized to systems in 1987 [**34, 35**]. Given an entropy pair (E, Q), a finite volume scheme was termed *entropy preserving* if it satisfies the discrete entropy equality

$$\frac{d}{dt}E(U_i(t)) + \frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\right) = 0 \tag{2.3}$$

for a numerical entropy flux $\widetilde{Q}$ consistent with Q. By summing over $i = k, \dots, l$ we see that

$$\frac{d}{dt}\left(\Delta x\sum_{i=k}^{l} E(U_i(t))\right) = \widetilde{Q}_{k+1/2} - \widetilde{Q}_{l-1/2}$$

for all choices of endpoints k, l. Hence, if the energy flux through the boundary is zero, then the right-hand side vanishes, and so the total amount of entropy in the interval $[x_k, x_l]$ is preserved over time.

A scheme was termed *entropy stable* if it satisfies the discrete entropy inequality (2.1). It was found in [**34, 35**] that a scheme is entropy stable precisely when it contains more numerical diffusion than an entropy preserving scheme. Moreover, an explicit condition was found that ensures that a scheme is entropy preserving. Thus, by designing an entropy preserving scheme, one may determine entropy stability in a scheme by *comparison.*

These notions will be made more precise in the next few sections. In Section 2.3 we design a novel entropy preserving scheme for the shallow water system (1.5) and study it through a series of numerical experiments. In Section 2.4 we add numerical diffusion to obtain an entropy stable scheme. In Section 2.5 we generalize these ideas to the full two-dimensional shallow water system (1.1).

2.2. Entropy preserving schemes

A solution of the conservation law (1.6) should satisfy the entropy inequality (1.17). Specifically, this implies the stability estimate

$$\frac{d}{dt}\int_{\mathbb{R}} E(U)\ dx \leq 0$$

on the solution. By requiring that numerical approximations satisfy the discrete entropy inequality (2.1), we ensure that the solutions do not converge towards entropy violating weak solutions. What is more, we get a stability estimate

$$\frac{d}{dt}\left(\Delta x \sum_i E(U_i)\right) \leq 0$$

in the approximate solutions.

Recall that the entropy equality (1.15) was obtained by multiplying (1.6) by $E'(U)^\top$ on both sides. Doing this to the semi-discrete scheme (1.21), we obtain

$$\frac{d}{dt}E(U_i(t)) = -\frac{1}{\Delta x}E'(U_i(t))^\top \left(F_{i+1/2} - F_{i-1/2}\right). \tag{2.4}$$

If the right-hand side were equal to $-\frac{1}{\Delta x}\big(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\big)$ for some $\widetilde{Q}_{i+1/2} = \widetilde{Q}\big(U_i, U_{i+1}\big)$ that is consistent with Q, then (2.4) would indeed be an entropy equality, and the scheme would be entropy preserving.

Given the importance of the quantity $E'(U)$, it is termed the vector of *entropy variables* and is denoted by $V = E'(U)$ [**34**]. As E is strictly convex, we have $E''(U) = \partial_U V > 0$, and so the mapping $V = V(U)$ from conserved variables to entropy variables is differentiable and injective. As a consequence, we may work with conserved variables and entropy variables interchangeably.

Given U_i, we write $V_i = E'(U_i)$. We wish to find a more direct criterion on F for the identity

$$V_i^\top \big(F_{i+1/2} - F_{i-1/2}\big) = \widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}$$

to hold. The next theorem gives precisely this. The *entropy potential* is defined as $\psi(U) = V(U)^\top f(U) - Q(U)$.

THEOREM 2.3 (Tadmor [**35**]). *Assume that a consistent numerical flux $F_{i+1/2}$ satisfies*

$$[\![V]\!]_{i+1/2}^\top F_{i+1/2} = [\![\psi]\!]_{i+1/2}. \tag{2.5}$$

Then solutions computed by the scheme with numerical flux $F_{i+1/2}$ satisfy the discrete entropy equality (2.1) *with numerical entropy flux*

$$\widetilde{Q}_{i+1/2} = \overline{V}_{i+1/2}^\top F_{i+1/2} - \overline{\psi}_{i+1/2}. \tag{2.6}$$

Hence, $\widetilde{Q}$ is consistent with Q.

PROOF. Taking the inner product of (1.21) with $V_i = E'(U_i)$, we get

$$\frac{d}{dt}E(U_i) = -\frac{1}{\Delta x}\left(V_i^\top F_{i+1/2} - V_i^\top F_{i-1/2}\right)$$

(adding and subtracting)

$$\begin{aligned} = -\frac{1}{\Delta x}\bigg(&\Big(\overline{V}_{i+1/2}^\top F_{i+1/2} - \frac{1}{2}[\![V]\!]_{i+1/2}^\top F_{i+1/2}\Big) \\ &- \Big(\overline{V}_{i-1/2}^\top F_{i-1/2} + \frac{1}{2}[\![V]\!]_{i-1/2}^\top F_{i-1/2}\Big)\bigg)\end{aligned}$$

(by (2.5) and the definition of $\widetilde{Q}$)

$$
\begin{aligned}
&= -\frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} + \overline{\psi}_{i+1/2} - \frac{1}{2}[\![\psi]\!]_{i+1/2} - \widetilde{Q}_{i-1/2} - \overline{\psi}_{i-1/2} - \frac{1}{2}[\![\psi]\!]_{i-1/2}\right) \\
&= -\frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\right).
\end{aligned}
$$

For consistency of $\widetilde{Q}$, we see that if $U_i = U_{i+1}$, then

$$
\begin{aligned}
\widetilde{Q}_{i+1/2} &= \overline{V}_{i+1/2}^{\top} F_{i+1/2} - \overline{\psi}_{i+1/2} \\
&= V_i^{\top} f_i - (V_i^{\top} f_i - Q_i) \\
&= Q_i.
\end{aligned}
$$

□

In the scalar case $n = 1$, the equation (2.5) has the unique solution

$$F_{i+1/2} = \frac{[\![\psi]\!]_{i+1/2}}{[\![V]\!]_{i+1/2}}.$$

As mentioned earlier, Burgers' equation has an entropy pair $E(u) = u^2/2$, $Q(u) = u^3/3$. The corresponding entropy variables and potential are $V(u) = u$ and $\psi(u) = u^3/6$, so the entropy preserving flux for Burgers' equation is

$$F_{i+1/2} = \frac{1}{6}\frac{[\![u^3]\!]_{i+1/2}}{[\![u]\!]_{i+1/2}} = \frac{u_i^2 + u_i u_{i+1} + u_{i+1}^2}{6}.$$

The more general case of a system of conservation laws (1.6) is harder, but some existence results exist. The flux originally proposed in [**35**] is in the form of a path integral,

$$F_{i+1/2} = \int_0^1 f(V(\xi))\, d\xi, \tag{2.7}$$

where $V(\xi) = V_i + \xi[\![V]\!]_{i+1/2}$. It is a straight-forward calculation that F is consistent and satisfies (2.5); see [**35**]. To see how this numerical flux is applied, we consider the conservation law with flux $f(U) = AU$ for a symmetric matrix A. As previously mentioned, an entropy pair for this conservation law is $E(U) = \frac{1}{2}U^{\top}U$, $Q(U) = \frac{1}{2}U^{\top}AU$, with entropy variables $V = E'(U) = U$. Inserting into (2.7) and evaluating the integral gives the entropy preserving flux

$$F_{i+1/2} = \frac{1}{2}\left(f(U_i) + f(U_{i+1})\right).$$

However, for more general, nonlinear flux functions, the integral (2.7) might be hard to calculate explicitly.

2.3. Energy preservation

Our aim is to design *energy stable* schemes for the shallow water equations. These are schemes whose solutions satisfy the discrete entropy inequality

$$\frac{d}{dt}E(U_i(t)) + \frac{1}{\Delta x}\left(\widehat{Q}_{i+1/2} - \widehat{Q}_{i-1/2}\right) \leq 0$$

with $E = \frac{1}{2}(hu^2 + gh^2)$ the energy of the shallow water system and $\widehat{Q}$ any numerical entropy flux function that is consistent with $Q = \frac{1}{2}hu^3 + guh^2$. The first step towards designing energy stable schemes will be to design an energy *preserving* scheme – a scheme satisfying the discrete entropy equality (2.3) for the energy. Preservation of energy in the shallow water equations has been studied extensively by, among others, Arakawa et. al [**1, 2, 3**], and has been deemed important especially in long-term meteorological simulations. We go on to present three energy preserving schemes: The average flux (2.7), the *PEC* scheme presented in [**36**], and a novel, much simpler flux.

2.3.1. The AEC scheme. The integral (2.7) can be evaluated explicitly [**11**] to obtain the energy preserving flux $F_{i+1/2} = \left[F^{(1)}_{i+1/2}, F^{(2)}_{i+1/2}\right]^\top$, with

$$\begin{aligned} F^{(1)}_{i+1/2} &= \frac{h_i u_i}{3} + \frac{h_{i+1}u_{i+1}}{3} + \frac{h_i u_{i+1}}{6} + \frac{h_{i+1}u_i}{6} \\ &\quad - \frac{u_i^3}{24} - \frac{u_{i+1}^3}{24} + \frac{u_i u_{i+1}^2}{24} + \frac{u_{i+1}u_i^2}{24} \\ F^{(2)}_{i+1/2} &= \frac{1}{12}h_i u_i^2 + \frac{1}{12}h_{i+1}u_{i+1}^2 + \frac{1}{6}h_i u_{i+1}^2 + \frac{1}{6}h_{i+1}u_i^2 \\ &\quad + \frac{1}{4}h_i u_i u_{i+1} + \frac{1}{4}h_{i+1}u_i u_{i+1} + \frac{7g}{24}h_i^2 + \frac{7g}{24}h_{i+1}^2 \\ &\quad - \frac{g}{12}h_i h_{i+1} + \frac{1}{96}u_i^4 + \frac{1}{96}u_{i+1}^4 - \frac{1}{48}u_i^2 u_{i+1}^2. \end{aligned} \tag{2.8}$$

We denote this as the *AEC* (Averaged Energy Conservative) scheme. It is readily shown that the AEC flux is consistent with the shallow water system.

2.3.2. The PEC scheme. An explicit solution of (2.5) was found by Tadmor in [**36**]. In this flux, the path integral in (2.7) is replaced by a piecewise linear path along orthogonal directions in $\mathbb{R}^n$. Let $\{l_k, r_k\}_{k=1}^n$ be an orthogonal eigensystem in $\mathbb{R}^n$. Define

$$V^{[0]} = V_i \qquad \text{and} \qquad V^{[k]} = V^{[k-1]} + \left(\llbracket V \rrbracket^\top_{i+1/2} l_k\right) r_k \qquad \text{for } k = 1, \ldots, n.$$

Note that $V^{[n]} = V_{i+1}$. Let

$$F^{[k]} = \frac{\psi\left(V^{[k]}\right) - \psi\left(V^{[k-1]}\right)}{\llbracket V \rrbracket^\top_{i+1/2} l_k} l_k \qquad \text{for } k = 1, \ldots, n. \tag{2.9}$$

The entropy preserving flux is given by

$$F_{i+1/2} = \sum_{k=1}^n F^{[k]}. \tag{2.10}$$

To see that this flux satisfies (2.5), multiply by $\llbracket V \rrbracket^\top_{i+1/2}$ to get

$$\llbracket V \rrbracket^\top_{i+1/2} F_{i+1/2} = \sum_{k=1}^n \psi\left(V^{[k]}\right) - \psi\left(V^{[k-1]}\right) = \llbracket \psi \rrbracket_{i+1/2}.$$

In the flux (2.10), we follow Tadmor and Zhong [**37**] and choose the system of eigenvectors as that of the Jacobian of the flux function f, evaluated at the Roe average (1.27). We denote the resulting scheme as the *PEC* (Pathwise Energy Conservative) scheme.

2.3.3. The EEC scheme. The AEC and PEC schemes have several disadvantages. Both fluxes contain complex expressions that require a large number of floating point operations at each flux evaluation. Moreover, the PEC has some numerical stability issues, most notably that a division by zero may occur in the expression for $F^{[k]}$ (2.9). The third scheme that we will consider avoids these problems by applying (2.5) in a more direct manner. To separate this flux from the others, we denote it by $\widetilde{F}$.

The entropy variables and entropy potential of the energy of the shallow water equations are

$$V = E'(U) = \begin{bmatrix} gh - \frac{u^2}{2} \\ u \end{bmatrix} \qquad \text{and} \qquad \psi = \frac{1}{2}guh^2. \tag{2.11}$$

An energy preserving scheme with numerical flux $\widetilde{F}_{i+1/2} = \left[\widetilde{F}^{(1)}_{i+1/2}, \ \widetilde{F}^{(2)}_{i+1/2}\right]^\top$ must satisfy (2.5), which in this case is

$$[\![gh]\!]\widetilde{F}^{(1)}_{i+1/2} - [\![u^2/2]\!]_{i+1/2}\widetilde{F}^{(1)}_{i+1/2} + [\![u]\!]_{i+1/2}\widetilde{F}^{(2)}_{i+1/2} = \frac{1}{2}g[\![uh^2]\!]_{i+1/2}.$$

By the first identity in (1.28), this can be written as

$$\begin{aligned} g[\![h]\!]_{i+1/2}\widetilde{F}^{(1)}_{i+1/2} - \overline{u}_{i+1/2}[\![u]\!]_{i+1/2}\widetilde{F}^{(1)}_{i+1/2} + [\![u]\!]_{i+1/2}\widetilde{F}^{(2)}_{i+1/2} \\ = g[\![h]\!]_{i+1/2}\overline{h}_{i+1/2}\overline{u}_{i+1/2} + \frac{1}{2}g\overline{h^2}_{i+1/2}[\![u]\!]_{i+1/2}. \end{aligned}$$

By equating jumps in the same variable we get the set of equations

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

which have solution

$$\widetilde{F}_{i+1/2} = \begin{bmatrix} \overline{h}_{i+1/2}\overline{u}_{i+1/2} \\ \frac{1}{2}g\overline{h^2}_{i+1/2} + \overline{h}_{i+1/2}\overline{u}^2_{i+1/2} \end{bmatrix}. \tag{2.12}$$

The resulting finite volume scheme is then

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(\widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2}\right), \tag{2.13}$$

which we denote as the *EEC* (Explicit Energy Conserving) scheme.

THEOREM 2.4. *The EEC scheme* (2.13) *is consistent with* (1.5) *and is second-order accurate. Furthermore, it is energy preserving, i.e. it satisfies the energy quality* (2.3) *with numerical entropy flux*

$$\widetilde{Q}_{i+1/2} = g\overline{h}_{i+1/2}\frac{u_ih_{i+1} + u_{i+1}h_i}{2} + \frac{1}{2}\overline{h}_{i+1/2}\overline{u}_{i+1/2}u_iu_{i+1}, \tag{2.14}$$

which is consistent with Q.

PROOF. Consistency of the EEC scheme follows immediately from the definition. A simple truncation error analysis shows that the scheme is of second order. The scheme satisfies the energy preservation criterion (2.5) by construction, and so satisfies the discrete energy equality with $\widetilde{Q}$ as in (2.6), which in this case is (2.14). □

2.3.4. Numerical experiment. We test the energy preserving properties of the AEC, PEC and EEC schemes. To measure the change in energy over time, we will consider the relative change in energy over time,

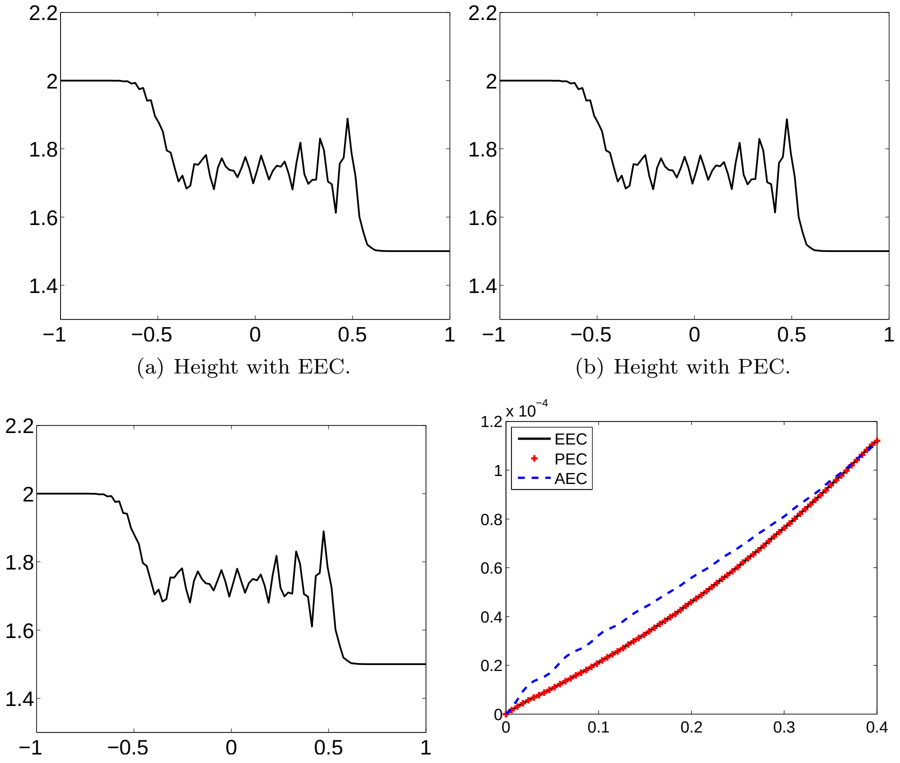
$$\frac{\|E(t) - E(0)\|_{L^1}}{\|E(0)\|_{L^1}},$$

where $E(t) = \sum_i E(U_i(t))$. By the discussion in the beginning of this chapter, $E(t)$ should be constant as long as the entropy flux through the boundary is zero. The energy flux Q is zero exactly when $u = 0$. Therefore we stop the simulation before any waves hit the boundary.

We consider a standard dam-break problem,

$$h(x,0) = \begin{cases} 2 & \text{if } x < 0 \\ 1.5 & \text{if } x > 0 \end{cases} \qquad u(x,0) \equiv 0. \tag{2.15}$$

The correct solution consists of a left-going rarefaction wave and a right-going shock. We compute with the energy preserving schemes on a mesh of 100 grid points up to time $t = 0.4$ using the second-order accurate Runge-Kutta (RK2) method (1.22) for temporal discretization. Figure 2.1(a)-(c) show the computed solutions. There is little discernible difference between the solutions, and all three schemes resolve the shock and the rarefaction wave correctly. However, in the wake of the shock there are unphysical oscillations with period of the order of Δx. This is to be expected from an energy preserving scheme. With the lack of any diffusive mechanism, nonlinear dispersive effects redistribute energy into higher wave numbers. As the highest wave number on a discrete mesh is $\frac{1}{\Delta x}$, the oscillations will have period precisely of order Δx. This is verified in Figure 2.2, where we repeat the experiment with the EEC scheme on a sequence of grid sizes. In all four solutions the period of the oscillations are of the order of Δx. Note that the amplitude of the oscillations in all experiment are of the order of the initial jump in h. To get rid of the oscillations, numerical dissipation must be added to allow a diffusion of energy.



(a) Height with EEC.

(b) Height with PEC.

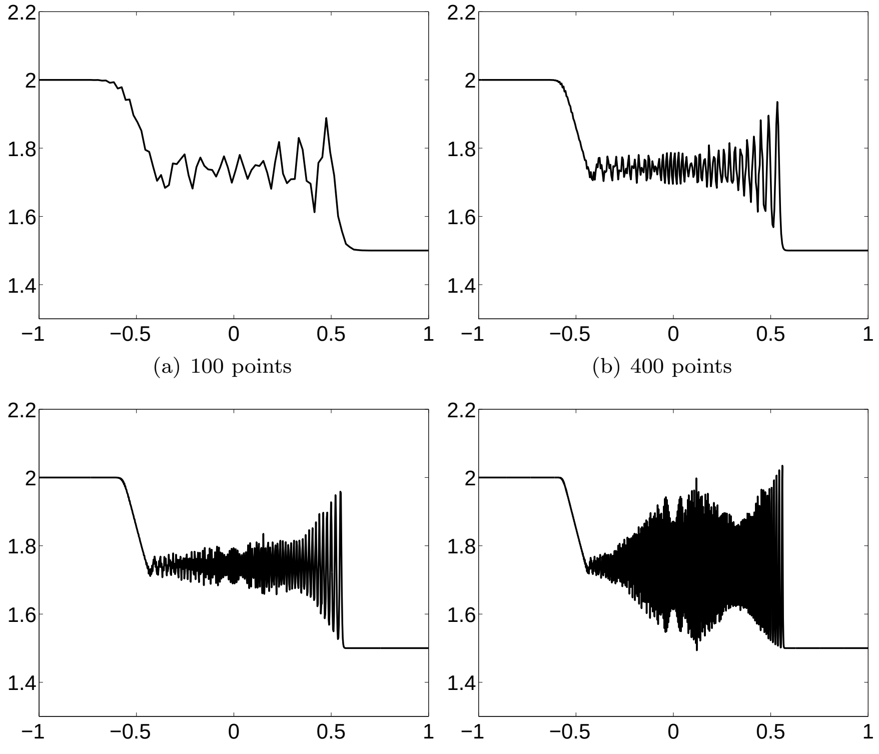
(c) Height with AEC.

(d) Relative energy vs. time.

FIGURE 2.1. Height h at $t = 0.4$ computed with the three energy preserving schemes.

In Figure 2.1(d) we see the relative energy of the solution as a function of time. There is a certain increase in energy over time, although we have proved that energy should be preserved. We claim that this increase is solely due to the error in the discretization (1.22) of the time integral. To demonstrate this, we repeat the experiment with the EEC scheme using a smaller CFL number; hence, Δt is lowered accordingly. In addition, we compute with the Runge-Kutta (RK3) method (1.23), which is more accurate than RK2. The effect can be seen in Figure 2.3. There is a drastic decrease in the amount of energy diffusion for both methods. Furthermore, the RK3 method is less diffusive than the RK2 method.

We remark that lowering the CFL number or changing the method of integration has no visible effect on the solution variables. This shows just how small the scales we are dealing with are.



(a) 100 points

(b) 400 points

(c) 800 points

(d) 1600 points

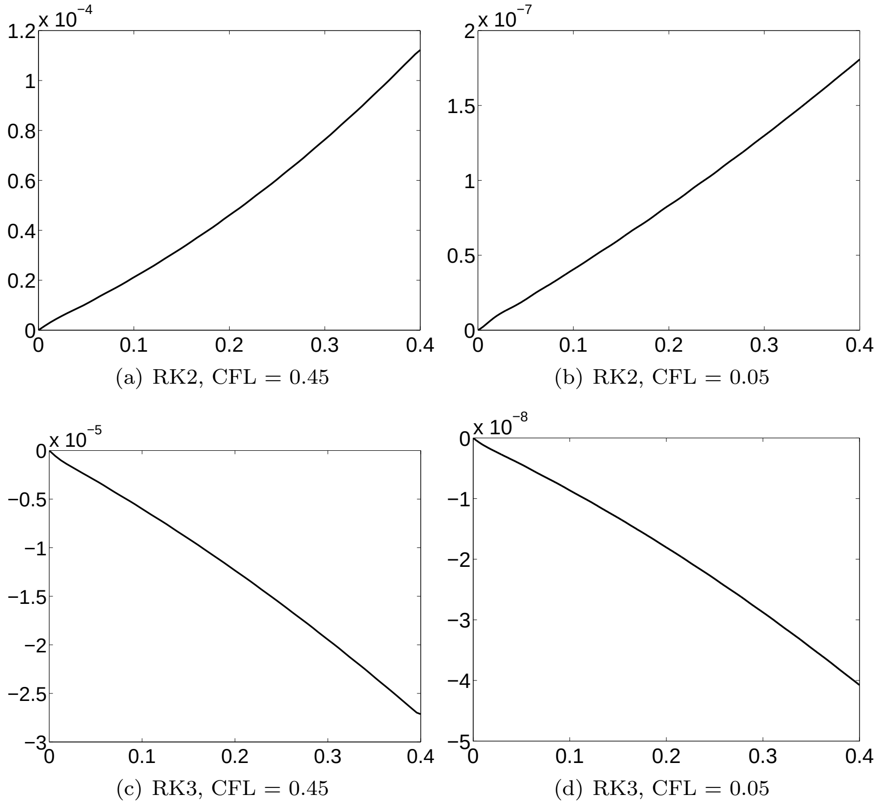
Figure 2.2. Height h at time $t = 0.4$ computed with the EEC scheme.

2.3.5. Numerical experiment: Computational cost. Next, we compare the computational efficiency of the schemes. As the error in energy decreases with the CFL number, it is interesting to see how the CFL number must be chosen to get a certain bound on this error. For each scheme we find the CFL number that ensures that the energy dissipation error is less than 10^{-3}, 10^{-4} and 10^{-5}. The runtime of each scheme is shown in Table 2.1. Clearly the EEC scheme is the fastest of the three, with runtimes almost three times lower than the PEC scheme. The AEC scheme is not far behind, but for some reason, we were unable to reduce errors to 10^{-5} with this scheme.

Energy error	10^{-3}	10^{-4}	10^{-5}
EEC	1	1.69	3.49
PEC	2.61	4.68	10.47
AEC	1.14	1.91	-

Table 2.1. Normalized runtimes for the three energy preserving schemes on the one-dimensional dam break problem with three different levels of error in energy.

In conclusion, we have designed a novel scheme that preserves the energy of the shallow water system exactly – modulo temporal discretization errors. The scheme is very simple to implement, and as it requires only a few floating point operations, is computationally very efficient. However, the presence of a shock resulted in the production of unphysical oscillations. This will be resolved in the next section.



(a) RK2, CFL = 0.45

(b) RK2, CFL = 0.05

(c) RK3, CFL = 0.45

(d) RK3, CFL = 0.05

FIGURE 2.3. Relative energy of solutions computed with the EEC scheme on 100 mesh points for two different RK schemes at two different CFL numbers.

2.4. Numerical diffusion

The EEC scheme was designed to preserve energy. However, energy should be *dissipated* at shocks. As we saw in the previous section, the scheme produces a cascade of energy into higher wave numbers, which results in oscillations in the trail of the shock. To amend this, numerical diffusion must be added to the scheme to obtain an energy *stable* scheme. To quantify this numerical diffusion, we write the numerical flux in the *viscous form* [**35**]

$$F_{i+1/2} = \frac{1}{2}\big(f(U_i) + f(U_{i+1})\big) - \frac{1}{2}P_{i+1/2}[\![V]\!]_{i+1/2}. \tag{2.16}$$

The construction of the viscosity coefficient matrix P is given in the following lemma.

LEMMA 2.5. *Assume that the numerical flux $F_{i+1/2} = F(U_i, U_{i+1})$ is Lipschitz in each parameter and is consistent with the conservation law. Then there exists a matrix $P_{i+1/2}$ such that $F_{i+1/2}$ can be written as* (2.16)*.*

PROOF. (2.16) can be rewritten as

$$\begin{aligned} P_{i+1/2}[\![V]\!]_{i+1/2} &= f_i + f_{i+1} - 2F_{i+1/2} \\ &= \big(F(U_{i+1}, U_{i+1}) - F(U_i, U_{i+1})\big) - \big(F(U_i, U_{i+1}) - F(U_i, U_i)\big) \end{aligned} \tag{2.17}$$

by the consistency of F. As previously remarked, the mapping from conserved variables U to entropy variables V is injective and differentiable, so we might as well consider F as a function of V. Letting $V(\xi) = V_i + \xi[\![V]\!]_{i+1/2}$ for $\xi \in [0,1]$, we have

$$\begin{aligned} F(V_i, V_{i+1}) - F(V_i, V_i) &= \int_0^1 \frac{\partial}{\partial \xi} F(V_i, V(\xi)) d\xi \\ &= \int_0^1 \partial_2 F(V_i, V(\xi)) d\xi [\![V]\!]_{i+1/2}, \end{aligned}$$

where $\partial_2 F$ is the partial derivative of F with respect to its second argument. Applying this to (2.17) then gives

$$P_{i+1/2} = \int_0^1 \partial_1 F(V(\xi), V_{i+1}) d\xi - \int_0^1 \partial_2 F(V_i, V(\xi)) d\xi.$$

□

The condition that will be imposed to obtain entropy stability will be one of comparison; an entropy stable scheme will be one that contains more diffusion than an entropy preserving scheme. To formalize this, let $\widetilde{F}$ be a numerical flux satisfying the requirement of Theorem 2.3, so that the corresponding scheme is entropy preserving, and let F be any other numerical flux consistent with (1.6). We denote their numerical viscosity matrices by $\widetilde{P}$ and P, respectively, and we let $D_{i+1/2} = P_{i+1/2} - \widetilde{P}_{i+1/2}$ be their difference.

THEOREM 2.6 (Tadmor [**35**]). *Assume that*

$$[\![V]\!]_{i+1/2}^\top D_{i+1/2} [\![V]\!]_{i+1/2} \geq 0 \tag{2.18}$$

for all V_i and V_{i+1}. Then the scheme with numerical flux F satisfies the entropy dissipation estimate

$$\begin{aligned} &\frac{d}{dt} E(U_i) + \frac{1}{\Delta x}\left(\widehat{Q}_{i+1/2} - \widehat{Q}_{i-1/2}\right) \\ &\quad = -\frac{1}{4\Delta x}\left([\![V]\!]_{i+1/2}^\top D_{i+1/2} [\![V]\!]_{i+1/2} + [\![V]\!]_{i-1/2}^\top D_{i-1/2} [\![V]\!]_{i-1/2}\right) \\ &\quad \leq 0, \end{aligned} \tag{2.19}$$

where

$$\widehat{Q}_{i+1/2} = \widetilde{Q}_{i+1/2} - \frac{1}{2} \overline{V}_{i+1/2}^\top D_{i+1/2} [\![V]\!]_{i+1/2}$$

is consistent with Q and $\widetilde{Q}$ is as in (2.6).

PROOF. As in the proof of theorem 2.6, we left-multiply (1.21) by $V_i^\top$ on both sides. Adding and subtracting $\widetilde{F}$ from F then gives

$$\begin{aligned} \frac{d}{dt} E(U_i) &= -\frac{1}{\Delta x}\left(V_i^\top \widetilde{F}_{i+1/2} - V_i^\top \widetilde{F}_{i-1/2}\right) \\ &\quad - \frac{1}{\Delta x}\left(V_i^\top (F_{i+1/2} - \widetilde{F}_{i+1/2}) - V_i^\top (F_{i-1/2} - \widetilde{F}_{i-1/2})\right) \\ &= -\frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\right) \\ &\quad + \frac{1}{2\Delta x}\left(V_i^\top D_{i+1/2} [\![V]\!]_{i+1/2} - V_i^\top D_{i-1/2} [\![V]\!]_{i-1/2}\right). \end{aligned}$$

The second part of the right-hand side can be rewritten as

$$\frac{1}{2\Delta x}\left(V_i^\top D_{i+1/2} [\![V]\!]_{i+1/2} - V_i^\top D_{i-1/2} [\![V]\!]_{i-1/2}\right)$$

$$
\begin{aligned}
&= \frac{1}{2\Delta x}\left(\overline{V}^{\top}_{i+1/2}D_{i+1/2}[\![V]\!]_{i+1/2} - \overline{V}^{\top}_{i-1/2}D_{i-1/2}[\![V]\!]_{i-1/2}\right)\\
&\quad - \frac{1}{4\Delta x}\left([\![V]\!]^{\top}_{i+1/2}D_{i+1/2}[\![V]\!]_{i+1/2} + [\![V]\!]^{\top}_{i-1/2}D_{i-1/2}[\![V]\!]_{i-1/2}\right)\\
&\le \frac{1}{2\Delta x}\left(\overline{V}^{\top}_{i+1/2}D_{i+1/2}[\![V]\!]_{i+1/2} - \overline{V}^{\top}_{i-1/2}D_{i-1/2}[\![V]\!]_{i-1/2}\right),
\end{aligned}
$$

as $[\![V]\!]^{\top}_{i\pm1/2}D_{i\pm1/2}[\![V]\!]_{i\pm1/2} \ge 0$. This gives the entropy dissipation estimate.

$\widehat{Q}$ is consistent with Q since, if $U_i = U_{i+1}$, then $[\![V]\!]_{i+1/2} = 0$, and so $\widehat{Q}_{i+1/2} = \widetilde{Q}_{i+1/2} + 0 = Q_i$. □

Remark 2.7. A sufficient condition for (2.18) is that the symmetric part of the matrix $D_{i+1/2}$ is positive for all V_i, V_{i+1}. Since $D_{i+1/2}$ in general depends nonlinearly on V_i and V_{i+1}, this is only a necessary condition when we are in the scalar case, in which case (2.18) means that the number $D_{i+1/2}$ is nonnegative.

In general it is hard to directly design entropy stable diffusion operators. Our approach will be to modify the diffusion operator of the Roe flux (1.26) to obtain an entropy stable diffusion operator. The Roe flux is selected specifically for its simplicity and good accuracy, and other fluxes might indeed be chosen.

We will replace the flux average term $\frac{1}{2}(f(U_i)+f(U_{i+1}))$ in (1.26) by the energy preserving EEC flux (2.13), and the diffusion operator $R|\Lambda|R^{-1}[\![U]\!]_{i+1/2}$ by a term of the form $A_{i+1/2}[\![V]\!]_{i+1/2}$ for a positive matrix $A_{i+1/2}$. A similar approach has been considered by Roe in [**33**] for the Euler equations. The construction of the matrix $A_{i+1/2}$ relies on the following result, a special case of a more general result of Barth [**5**].

Lemma 2.8. *Let U, U_i and U_{i+1} be given states. We define the symmetric positive definite change-of-variables matrix at $U = [h, hu]^{\top}$ as*

$$U_V = \frac{1}{g}\begin{bmatrix} 1 & u \\ u & u^2 + gh \end{bmatrix}.$$

(i) Define the following scaled version of the eigenvector matrix R of $f'(U)$:

$$R = \frac{1}{\sqrt{2g}}\begin{bmatrix} 1 & 1 \\ u - \sqrt{gh} & u + \sqrt{gh} \end{bmatrix}.$$

Then we have

$$RR^{\top} = U_V.$$

(ii) We have

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

where

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

is U_V evaluated at the arithmetic averages

$$h = \overline{h}_{i+1/2}, \qquad \textit{and} \qquad u = \overline{u}_{i+1/2}. \tag{2.20}$$

Proof. The results follow simply by insertion. □

Now, given left and right states U_i and U_{i+1}, we evaluate the Roe matrix $R|\Lambda|R^{-1}$ at the state given in lemma 2.8 (ii), $U_{i+1/2} = [\overline{h}_{i+1/2},\ \overline{h}_{i+1/2}\overline{u}_{i+1/2}]^{\top}$. We are free to evaluate the matrix wherever we want, although by using states other than the Roe average we will lose the property of exact resolution of single shocks.

By inserting this state we get

$$R|\Lambda|R^{-1}[\![U]\!]_{i+1/2} = R|\Lambda|R^{-1}(U_V)_{i+1/2}[\![V]\!]_{i+1/2}$$

$$
\begin{aligned}
&= R|\Lambda|R^{-1}RR^\top [\![V]\!]_{i+1/2}\\
&= R|\Lambda|R^\top [\![V]\!]_{i+1/2}.
\end{aligned}
$$

We define the numerical flux

$$
F^{\mathrm{ERoe}}_{i+1/2} = \widetilde{F}_{i+1/2} - \frac{1}{2}R|\Lambda|R^\top [\![V]\!]_{i+1/2}, \tag{2.21}
$$

and we denote the corresponding scheme as the ERoe (entropy stable Roe) scheme. The following theorem is a result of theorem 2.6.

THEOREM 2.9. *The ERoe scheme is consistent with the shallow water system* (1.5)*, it is first-order accurate and is energy stable, i.e. it satisfies the energy dissipation estimate*

$$
\begin{aligned}
\frac{d}{dt}E(U_i) \ &+ \frac{1}{\Delta x}\left(\widehat{Q}_{i+1/2} - \widehat{Q}_{i-1/2}\right)\\
&= -\frac{1}{2\Delta x}\left([\![V]\!]^\top_{i+1/2}(R|\Lambda|R^\top)_{i+1/2}[\![V]\!]_{i+1/2} + [\![V]\!]^\top_{i-1/2}(R|\Lambda|R^\top)_{i-1/2}[\![V]\!]_{i-1/2}\right)\\
&\leq 0,
\end{aligned}
$$

where

$$
\widehat{Q}_{i+1/2} = \widetilde{Q}_{i+1/2} - \frac{1}{4}\overline{V}^\top_{i+1/2}R|\Lambda|R^\top [\![V]\!]_{i+1/2}
$$

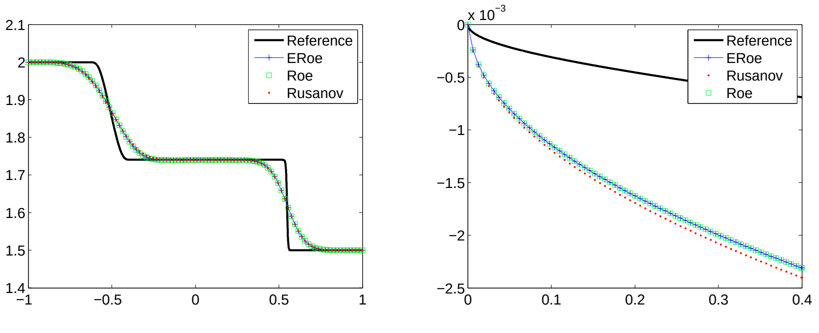
is consistent with Q. Here, $\widetilde{Q}$ is the energy flux (2.14) *of the EEC scheme.*

PROOF. Comparing the numerical viscosity matrices of the energy preserving EEC scheme and the ERoe scheme, we find that their difference is

$$
D_{i+1/2} = P_{i+1/2} - \widetilde{P}_{i+1/2} = R|\Lambda|R^\top,
$$

which is a positive symmetric matrix, an hence satisfies the stability criterion (2.18) for all V_i and V_{i+1}. The energy dissipation estimate then follows from theorem 2.6. □

2.4.1. Numerical experiments. We test the energy stability of the ERoe scheme in the dam break problem of the previous section. The results are computed and compared with the Rusanov scheme (1.25) and the Roe scheme (1.26), along with a reference solution computed with the Rusanov scheme on a fine mesh of 3200 grid points. Plots of height at $t = 0.4$ can be seen in Figure 2.4(a). While the shock and rarefaction wave are more smeared out than for the EEC scheme, we see that all unphysical oscillations have been cleared out. Thus, adding numerical diffusion to the EEC scheme has had the desired effect.



(a) Height at $t = 0.4$.

(b) Energy vs. time.

FIGURE 2.4. Solutions computed with the Roe, Rusanov and ERoe schemes for the one-dimensional dam break problem with 100 mesh points.

In Figure 2.4(b) we display the relative energy in the solution over time. The energy is clearly not preserved. Instead, a dissipation of energy is taking place – as it should, since there is a shock in the solution. Comparing the three schemes, the Rusanov scheme dissipates slightly more energy than the other two. It is well-known that the Rusanov scheme is more dissipative than many other schemes.

We see that all three schemes dissipate more energy than the reference solution does. We can conclude that, in this experiment, there is too much numerical diffusion in all three schemes. However, finding the correct amount of diffusion can be very difficult, and a smaller amount of diffusion may come at the cost of unwanted oscillations.

Numerical experiment: Large dam break. In the previous experiment, the solutions computed with the three schemes were almost indistinguishable. We now consider an experiment that emphasizes the differences between the Roe and the ERoe scheme.

The initial data is given by

$$h(x,0) = \begin{cases} 15 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} \qquad u(x,0) \equiv 0. \tag{2.22}$$

The correct solution consists of a left-going rarefaction wave and a right-going shock. We compute on a mesh of 100 grid points up to $t = 0.4$, using a CFL number of 0.45, as before. A reference solution is computed with the Rusanov scheme on a mesh of 3200 grid points. The result can be viewed in Figure 2.5.

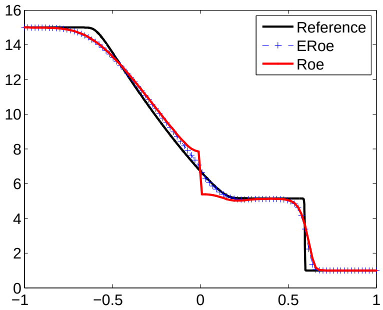


FIGURE 2.5. Height at time $t = 0.4$ computed with the Roe and ERoe schemes for the large dam break problem with 100 mesh points.

While the solutions are close away from the origin, there is a large discrepancy at $x = 0$. Here, the Roe scheme immediately produces an unphysical steady shock, whereas the ERoe scheme computes the rarefaction wave as it should. It is well-known that the Roe scheme is not entropy (energy) stable, and that it may produce steady shocks were there should be none (see e.g. [**36**]).

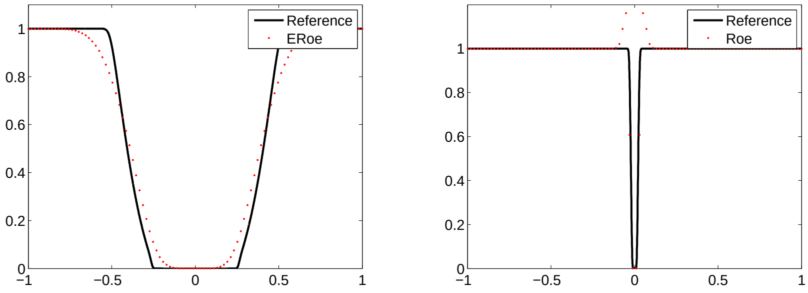
Numerical experiment: Near-zero heights. This experiment will illustrate how well the ERoe scheme preserves positivity. Recall that the height variable h is positive in exact solutions of the shallow water system, and a negative value of h would be meaningless. Hence, we should require that approximate solutions preserve the positivity of h.

The initial data is given by

$$h(x,0) \equiv 1, \qquad u(x,0) = \begin{cases} -u_0 & \text{if } x < 0 \\ u_0 & \text{if } x > 0 \end{cases} \tag{2.23}$$

for a $u_0 > 0$. Depending on how large u_0 is, the solution will create a gap in the water around the origin, with depths close to zero.

We let $u_0 = 4$, and we try to compute up to time $t = 0.1$ on the same mesh as before. The results are shown in Figure 2.6. The Roe scheme broke positivity at around $t = 0.0056$, and the simulation was halted. The ERoe scheme, on the other hand, computed the solution correctly, albeit somewhat more diffused than the reference solution.



(a) Height with ERoe at $t = 0.1$ (b) Height with Roe at $t = 0.0056$

FIGURE 2.6. Solutions computed with the Roe, and ERoe schemes for the expansion problem with 100 mesh points.

We do not claim that the ERoe scheme is positivity preserving. In fact, the scheme broke positivity when we increased u_0 to 8. Still, this and the previous experiment show that the ERoe scheme is more stable than the Roe scheme in the presence of large shocks and close-to-zero heights.

Numerical experiment: Computational efficiency. With the apparent advantages of the ERoe scheme over the Roe and Rusanov schemes, it is natural to ask whether these advantages come at the price of an increase in runtime. As a measure of accuracy we will see how close the computed height lies to the correct solution. We compute a reference solution of the dam break problem of Section 2.4.1 with the Rusanov scheme on a mesh of 3200 grid points, and for each of the three schemes, we find the mesh at which the relative error in height in the L^1 norm is less than 1, 0.5 and 0.1 per cent. The runtime of each scheme is presented in Table 2.2. Clearly, the ERoe scheme gives the best performance, and dissipates less energy than the other schemes. Surprisingly, the Rusanov scheme performs better than the Roe scheme, dissipating less energy at the same runtime.

Relative height error	1	0.5	0.1
Rusanov	1.05	8.24	203.41
Roe	1.15	8.43	208.29
ERoe	1	7.36	171.7

TABLE 2.2. Normalized run-times for the Rusanov, Roe and ERoe schemes on the one-dimensional dam break problem with three different levels of relative error in height

2.4.2. Increasing the order of accuracy. While the ERoe scheme is energy stable and has several attractive features, it is only first-order accurate. Consequently, sharp features such as shocks are diffused out, as seen in the dam break experiment in Section 2.4.1. In this section we apply a standard procedure to increase the order of accuracy of the scheme to obtain a second-order accurate scheme

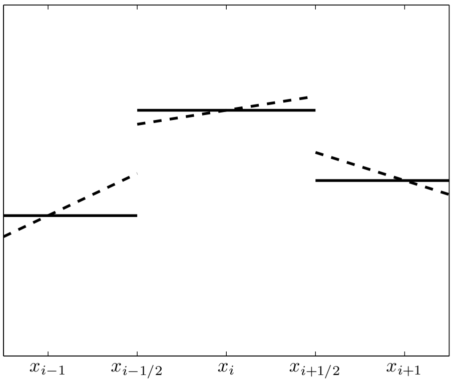


FIGURE 2.7. Original value U (solid line) and reconstructed values $\widetilde{U}$ (dashed line)

which we denote as the *ERoe2* scheme. This scheme will resolve shocks much more accurately, at the cost of a slight increase in runtime.

The method that we will employ relies on a reconstruction of the solution at each time step. Recall that in the finite volume framework, we solve for the average $U_i = \frac{1}{\Delta x}\int_{I_i} U(x,t)\,dx$ over a grid cell I_i. Thus, the approximate solution is constant in each grid cell. By applying a first-order interpolation or *reconstruction* of U in the neighboring grid cells, we obtain a linear (in each grid cell) function

$$\widetilde{U}(x) = U_i + \sigma_i(x - x_i) \qquad \text{for } x \in I_i,$$

where $\sigma_i \in \mathbb{R}^n$ is the slope. Given a scheme

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\big(F(U_i, U_{i+1}) - F(U_{i-1}, U_i)\big),$$

we replace U in the flux terms by the reconstructed variables:

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\Big(F\left(U_i^E, U_{i+1}^W\right) - F\left(U_{i-1}^E, U_i^W\right)\Big),$$

where

$$U_i^W = U_i + \sigma_i(x_{i-1/2} - x_i) = U_i - \frac{\Delta x}{2}\sigma_i$$

and

$$U_i^E = U_i + \sigma_i(x_{i+1/2} - x_i) = U_i + \frac{\Delta x}{2}\sigma_i$$

are the left and right states of the reconstructed variables in the grid cell I_i. See Figure 2.7 for an illustration. This procedure formally increases the order of accuracy of the scheme to two [**26**]. Higher order schemes may be obtained by using e.g. a quadratic or cubic reconstruction of U, but we will not consider this here.

There are several choices available for selecting the slope σ_i. Three candidates are the forward, backward and central differences

$$\frac{U_{i+1} - U_i}{\Delta x}, \qquad \frac{U_i - U_{i-1}}{\Delta x}, \qquad \frac{U_{i+1} - U_{i-1}}{2\Delta x}.$$

In each grid cell I_i and for each component of U_i, we will select the one of the three that gives the least oscillatory solution. In Figure 2.7, we see that the reconstructed value in grid cell I_i *increases* the gap at the cell interface $x_{i+1/2}$. This may lead to unwanted oscillations in the solution, particularly near shocks. To avoid oscillations, we use a so-called slope limiter in the construction of the slopes σ_i. We employ the minmod limiter

$$\sigma_i^{(k)} = \mathrm{mm}\left(\frac{U_{i+1}^{(k)} - U_i^{(k)}}{\Delta x}, \frac{U_i^{(k)} - U_{i-1}^{(k)}}{\Delta x}, \frac{U_{i+1}^{(k)} - U_{i-1}^{(k)}}{2\Delta x}\right) \tag{2.24}$$

where

$$\mathrm{mm}(x,y,z) := \begin{cases} \max\{|x|,|y|,|z|\} & \text{if } \mathrm{sign}\,(x) = \mathrm{sign}\,(y) = \mathrm{sign}\,(z) \\ 0 & \text{otherwise,} \end{cases} \tag{2.25}$$

and $\sigma_i^{(k)}$ is the kth component of σ_i. See LeVeque [**26**] for more information on slope limiters.

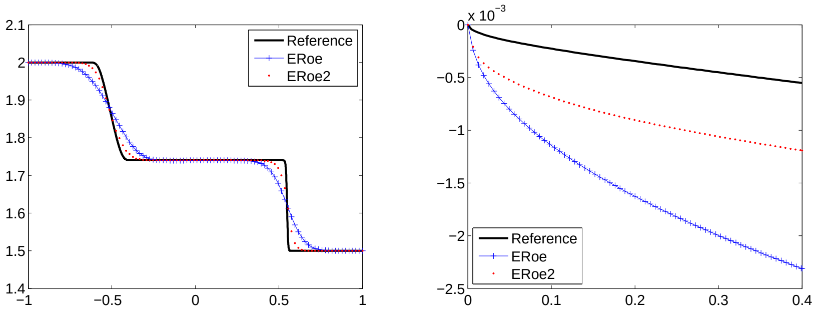
Instead of performing reconstruction in the conserved variables, we will reconstruct the energy variables. Given states $\{U_i\}$ at time t, we let $V_i = V(U_i)$, and perform a linear reconstruction of V, obtaining left and right states V_i^E and V_i^W. Let $U_i^E = U(V_i^E)$ and $U_i^W = U(V_i^W)$. The second-order accurate ERoe2 scheme then has flux

$$F_{i+1/2}^{\mathrm{ERoe2}} = \widetilde{F}(U_i, U_{i+1}) - \frac{1}{2}R|\Lambda|R^\top \left(V_{i+1}^W - V_i^E\right),$$

where R and Λ are evaluated at the average state

$$h = \frac{h_i^E + h_{i+1}^W}{2}, \qquad u = \frac{u_i^E + u_{i+1}^W}{2}$$

(compare to (2.20)). We do not use the reconstructed values in the EEC flux $\widetilde{F}$, as it is already second-order accurate.



(a) Height (b) Energy vs. time

FIGURE 2.8. Solutions computed with the first and second-order accurate versions of the ERoe scheme with 100 mesh points.

2.4.3. Numerical experiment. To witness the gain in accuracy with the ERoe2 scheme, we repeat the dam break problem of Section 2.4.1. As is clearly seen in Figure 2.8(a), the solution is now much less diffused out. Both the shock and the rarefaction wave are more accurately reproduced, and the shock spans only a few grid cells. Figure 2.8(b) displays the energy over time. The amount of energy diffusion is more than halved, and the profile lies much closer to the reference solution.

2.5. Two-dimensional energy preserving and stable schemes

Next, we generalize the schemes of the preceding sections to the two-dimensional shallow water system (1.1). As so much is similar to the one-dimensional case, the exposition will be brief and without proofs. First, we introduce the two-dimensional equivalent of an entropy pair.

DEFINITION 2.10. A triple (E, Q^x, Q^y) is an *entropy triple* for the conservation law (1.9) if $E : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and

$$(Q^x)'(U)^\top = E'(U)^\top f'(U) \qquad \text{and} \qquad (Q^y)'(U)^\top = E'(U)^\top g'(U)$$

for all $U \in \mathbb{R}^n$.

It is readily verified that the energy

$$E(U) = \frac{1}{2}\left(hu^2 + hv^2 + gh^2 + ghb\right) \tag{2.26}$$

of the shallow water system gives rise to an entropy triple (E, Q^x, Q^y) with

$$Q^x = \frac{1}{2}\left(hu^3 + huv^2\right) + guh^2 \qquad \text{and} \qquad Q^y(U) = \frac{1}{2}\left(hu^2v + hv^3\right) + gvh^2.$$

The energy variables are now

$$V = E'(U) = \begin{bmatrix} gh - \frac{u^2+v^2}{2} \\ u \\ v \end{bmatrix}, \tag{2.27}$$

and the energy potentials in each direction are

$$\psi^x = \frac{1}{2}guh^2 \qquad \text{and} \qquad \psi^y = \frac{1}{2}gvh^2.$$

Theorem 2.3, which gives a condition for schemes to be entropy preserving, is easily generalized to the two-dimensional case. By the same argument as for the one-dimensional EEC scheme, we find that the two-dimensional EEC scheme has numerical fluxes

$$\widetilde{F}_{i+1/2,j} = \left[\begin{array}{l} \overline{h}_{i+1/2,j}\ \overline{u}_{i+1/2,j} \\ \overline{h}_{i+1/2,j}(\overline{u}_{i+1/2,j})^2 + \frac{g}{2}(\overline{h^2})_{i+1/2,j} \\ \overline{h}_{i+1/2,j}\ \overline{u}_{i+1/2,j}\ \overline{v}_{i+1/2,j} \end{array}\right], \tag{2.28a}$$

and

$$\widetilde{G}_{i,j+1/2} = \left[\begin{array}{l} \overline{h}_{i,j+1/2}\ \overline{v}_{i,j+1/2} \\ \overline{h}_{i,j+1/2}\ \overline{u}_{i,j+1/2}\ \overline{v}_{i,j+1/2} \\ \overline{h}_{i,j+1/2}(\overline{v}_{i,j+1/2})^2 + \frac{g}{2}(\overline{h^2})_{i,j+1/2} \end{array}\right]. \tag{2.28b}$$

LEMMA 2.11. *The two-dimensional EEC scheme is consistent with the two-dimensional shallow water equations and is energy preserving, i.e. it satisfies the energy equality*

$$\frac{d}{dt}E(U_i) + \frac{1}{\Delta x}\left(\widetilde{Q}^x_{i+1/2,j} - \widetilde{Q}^x_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(\widetilde{Q}^y_{i,j+1/2} - \widetilde{Q}^y_{i,j-1/2}\right) = 0$$

where

$$\widetilde{Q}^x = \overline{V}^\top_{i+1/2,j}\widetilde{F}_{i+1/2,j} - [\![\psi^x]\!]_{i+1/2,j} \tag{2.29a}$$

and

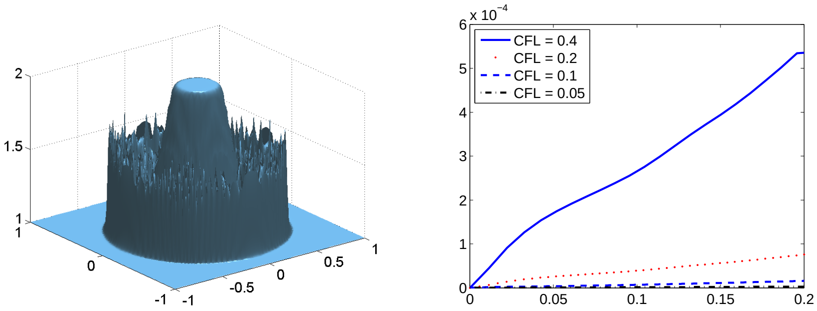
$$\widetilde{Q}^y = \overline{V}^\top_{i,j+1/2}\widetilde{G}_{i,j+1/2} - [\![\psi^y]\!]_{i,j+1/2}. \tag{2.29b}$$

Note that the two-dimensional EEC scheme reduces to its one-dimensional counterpart whenever there is no variation in the y-direction.

2.5.1. Numerical experiments. We test the EEC scheme on a two-dimensional equivalent of a dam break problem. The initial data is given by

$$h(x,y,0) = \begin{cases} 2 & \text{if } \sqrt{x^2+y^2} < 0.5 \\ 1 & \text{otherwise} \end{cases} \qquad u(x,y,0) \equiv 0. \tag{2.30}$$

The exact solution is an expanding shock wave with an inwards-going rarefaction wave. We compute on a uniform mesh of 100 × 100 grid points in the domain $(x,y) \in [0,1] \times [0,1]$ up to time $t = 0.2$. We use $g = 1$ as before. As seen in Figure 2.9(a), the shock and the rarefaction wave are computed correctly, but there are unphysical oscillations in the wake of the shock. As before, these are caused by a lack of numerical diffusion, and will be dampened when adding a diffusion operator in the next section.



(a) Height at $t = 0.2$.

(b) Energy for four different CFL numbers.

FIGURE 2.9. Solution of the cylindrical dam break problem computed on a uniform 100 × 100 mesh with the EEC scheme.

Figure 2.9(b) shows the relative change in energy over time for four different choices of CFL numbers. Clearly, the increase in energy is dampened with smaller time steps, and in the limit $\Delta t \to 0$, the scheme preserves energy exactly.

Numerical experiment: Computational cost. Next, we perform a test of the computational efficiency of the EEC and PEC schemes, similar to the one-dimensional dam break problem. We run the problem (2.30) on a 50 × 50 mesh and we find the CFL numbers that give a relative energy error of 10^{-3}, 10^{-4} and 10^{-5}. The runtimes are displayed in Table 2.3. Clearly, the EEC scheme gives the smallest energy error for the same runtime, with differences in runtime of the order of 5 or 6.

Relative energy error	10^{-3}	10^{-4}	10^{-5}
EEC	1	2.18	5.07
PEC	4.79	11.5	30.47

TABLE 2.3. Normalized run-times for the two energy preserving schemes on the two-dimensional cylindrical dam-break problem with three different levels of error in energy.

2.6. Two-dimensional ERoe

The generalization of the ERoe scheme to the two-dimensional shallow water system is straight-forward. If R^x, Λ^x, R^y and Λ^y are the eigenvector and eigenvalue matrices of $f'(U)$ and $g'(U)$, respectively, then the numerical fluxes F^{ERoe} and G^{ERoe} will be of the form

$$F^{\mathrm{ERoe}}_{i+1/2,j} = \widetilde{F}_{i+1/2,j} - \frac{1}{2} R^x |\Lambda^x| (R^x)^\top [\![V]\!]_{i+1/2,j}$$

and

$$G^{\mathrm{ERoe}}_{i,j+1/2} = \widetilde{G}_{i,j+1/2} - \frac{1}{2} R^y |\Lambda^y| (R^y)^\top [\![V]\!]_{i,j+1/2}.$$

To obtain the specific states at which the eigenmatrices are evaluated, we apply the following lemma, a straight-forward generalization of lemma 2.8.

LEMMA 2.12. *Let U, $U_{i,j}$, $U_{i+1,j}$ and $U_{i,j+1}$ be given states. We define the symmetric positive definite change-of-variables matrix at $U = [h, hu, hv]^\top$ as*

$$U_V = \partial_V U = \frac{1}{g} \begin{bmatrix} 1 & u & v \\ u & u^2 + gh & uv \\ v & uv & v^2 + gh \end{bmatrix}.$$

(i) *Define the following scaled version of the eigenvector matrices R^x and R^y of $f'(U)$ and $g'(U)$, respectively:*

$$R^x = \frac{1}{\sqrt{2g}} \begin{bmatrix} 1 & 0 & 1 \\ u - \sqrt{gh} & 0 & u + \sqrt{gh} \\ v & \sqrt{2gh} & v \end{bmatrix},$$
(2.31)
$$R^y = \frac{1}{\sqrt{2g}} \begin{bmatrix} 1 & 0 & 1 \\ u & -\sqrt{2gh} & u \\ v - \sqrt{gh} & 0 & v + \sqrt{gh} \end{bmatrix}.$$

Then we have

$$R^x (R^x)^\top = R^y (R^y)^\top = U_V.$$

(ii) *We have*

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

and

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

where $(U_V)_{i+1/2,j}$ and $(U_V)_{i,j+1/2}$ are U_V, evaluated at the arithmetic averages

(2.32a) $$h = \overline{h}_{i+1/2,j}, \qquad u = \overline{u}_{i+1/2,j}, \qquad v = \overline{v}_{i+1/2,j}$$

and

(2.32b) $$h = \overline{h}_{i,j+1/2}, \qquad u = \overline{u}_{i,j+1/2}, \qquad v = \overline{v}_{i,j+1/2}$$

respectively.

By the same technique as before, we end up with eigenvector matrices R^x and R^y (2.31), evaluated at the average states (2.32). We have the following energy stability result.

PROPOSITION 2.13. *The two-dimensional ERoe scheme is consistent with the shallow water system* (1.1). *Furthermore, it is energy stable, i.e. it satisfies the energy dissipation estimate*

$$\frac{d}{dt} E(U_{i,j}) + \frac{1}{\Delta x} \left(\widehat{Q}^x_{i+1/2,j} - \widehat{Q}^x_{i-1/2,j} \right) + \frac{1}{\Delta y} \left(\widehat{Q}^y_{i,j+1/2} - \widehat{Q}^y_{i,j-1/2} \right) \leq 0,$$

where

$$\widehat{Q}^x_{i+1/2,j} = \widetilde{Q}^x_{i+1/2,j} - \frac{1}{2} \overline{V}^\top_{i+1/2,j} R^x |\Lambda^x| (R^x)^\top [\![V]\!]_{i+1/2,j}$$

and

$$\widehat{Q}^y_{i,j+1/2} = \widetilde{Q}^y_{i,j+1/2} - \frac{1}{2} \overline{V}^\top_{i,j+1/2} R^y |\Lambda^y| (R^y)^\top [\![V]\!]_{i,j+1/2}$$

are consistent with Q^x *and* Q^y*, respectively. Here,* $\widetilde{Q}^x$ *and* $\widetilde{Q}^y$ *are the energy fluxes* (2.29) *of the two-dimensional EEC scheme.*

2.6.1. Two-dimensional ERoe2. We perform a two-dimensional reconstruction in the energy variables to obtain a two-dimensional, second-order accurate version of the ERoe scheme. A standard piecewise linear reconstruction gives reconstructed values

$$V(x,y) = V_{i,j} + \sigma_{i,j}(x - x_i) + \gamma_{i,j}(y - y_j) \qquad \text{for } (x,y) \in I_{i,j},$$

with $\sigma_{i,j}$ and $\gamma_{i,j}$ determined by a slope limiter, similar to the one-dimensional ERoe2 scheme. The numerical fluxes of the two-dimensional ERoe2 scheme are then

$$F^{\mathrm{ERoe2}}_{i+1/2,j} = \widetilde{F}(U_{i,j}, U_{i+1,j}) - \frac{1}{2} R^x |\Lambda^x| (R^x)^\top \left(V^W_{i+1,j} - V^E_{i,j}\right)$$

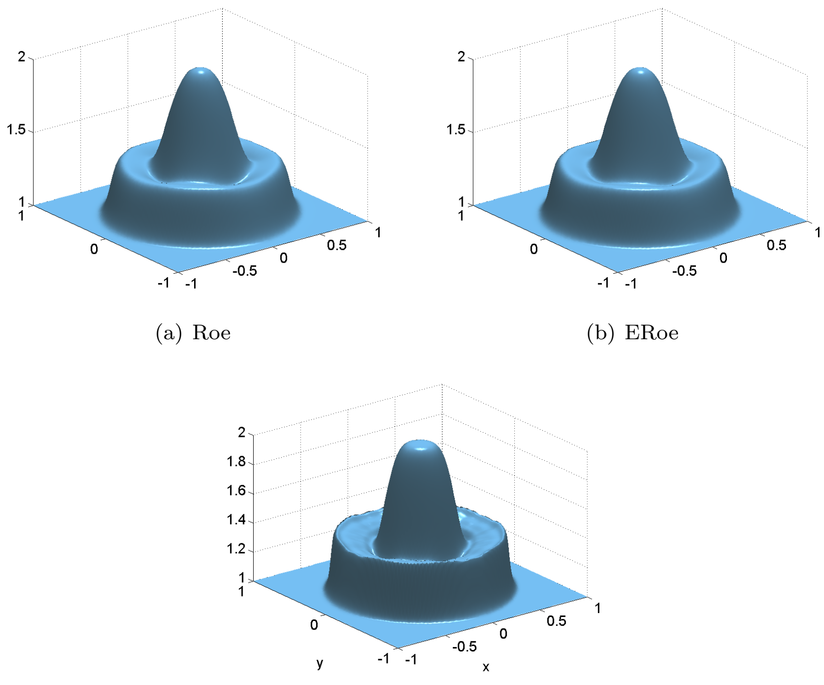
and

$$G^{\mathrm{ERoe2}}_{i,j+1/2} = \widetilde{G}(U_{i,j}, U_{i,j+1}) - \frac{1}{2} R^y |\Lambda^y| (R^y)^\top \left(V^S_{i,j+1} - V^N_{i,j}\right),$$

and $\widetilde{F}$ and $\widetilde{G}$ are as in (2.28).

THEOREM 2.14. *The two-dimensional ERoe2 scheme is second-order accurate and is consistent with* (1.1).

2.6.2. Numerical experiments. We repeat the cylindrical dam break problem of Section 2.5.1. As shown in Figure 2.10, the oscillations have been cleared out. The shock and rarefaction wave are slightly more diffused out. The ERoe2 scheme resolves the expanding shock more accurately than the Roe and ERoe schemes.



(a) Roe

(b) ERoe

(c) ERoe2

FIGURE 2.10. Approximate heights for the cylindrical dam break problem at time $t = 0.2$ computed on a 100×100 mesh with the Roe, ERoe and ERoe2 schemes.

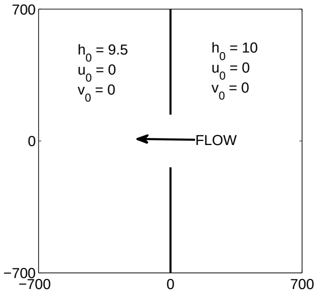


FIGURE 2.11. Setup of the physical dam break.

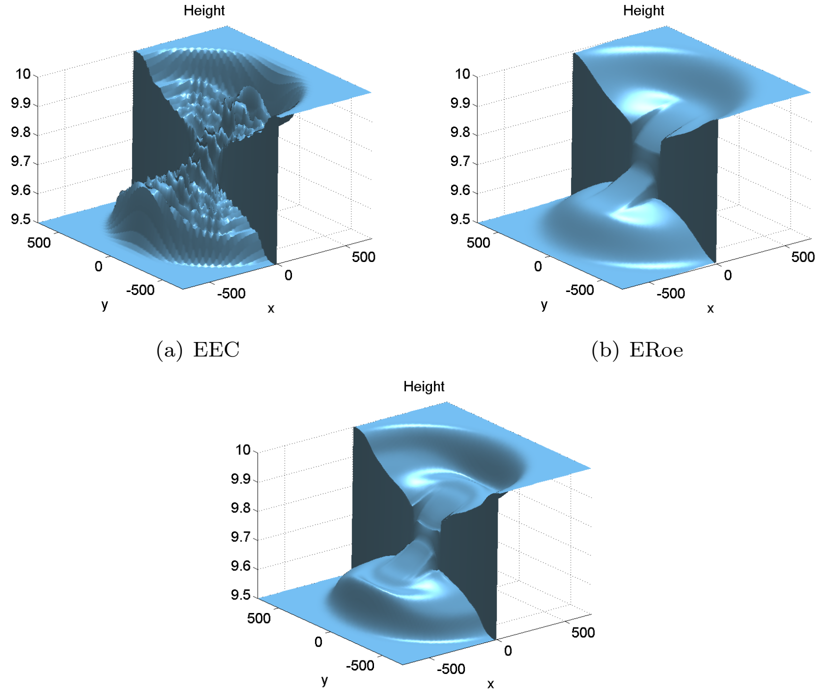
Numerical experiment: Physical dam break. We test the two-dimensional EEC, ERoe and ERoe2 schemes in a slightly more realistic problem that was first studied in [**10**] and was also considered in [**8, 38**]. We compute in a basin of 1400 × 1400 meters with a dam along the y-axis. At $t = 0$, a 280 meter long section of the dam fails, and water flows through the breach. The initial data is given by

$$h(x, y, 0) = \begin{cases} 10 & \text{if } x > 0 \\ 9.5 & \text{if } x < 0 \end{cases} \qquad \text{and} \qquad u = v = 0.$$

The setup is displayed in Figure 2.11. Along the boundary of the domain we impose Neumann boundary conditions, and along the dam we set a reflective boundary. The acceleration due to gravity is set to $g = 9.812$. We compute on a mesh of 100×100 grid points up to $t = 50$, and show the solution computed by the EEC, ERoe and ERoe2 schemes in Figure 2.12. The EEC scheme solves the resulting shock and rarefaction waves quite accurately, but with unphysical oscillations due to a lack of energy dissipation. The diffusion operator of the ERoe scheme corrects these, but at the cost of a less accurate resolution of the shock wave. The second-order accurate ERoe2 scheme gives a sharper resolution of the shock and the rarefaction waves.

2.7. Conclusion

We have designed novel energy preserving and energy stable finite volume schemes for the shallow water system in both one and two space dimensions. The energy preserving scheme is easy to implement, has a low computational cost and is second order accurate. The energy stable ERoe scheme clears out the oscillations of the EEC scheme and is more stable and accurate than comparable schemes, but is only first-order accurate. A reconstruction in the energy variables gives a second-order accurate method.



(a) EEC

(b) ERoe

(c) ERoe2

FIGURE 2.12. Height computed with the EEC, ERoe and ERoe2 schemes in the physical dam break problem.

CHAPTER 3

Well-balanced schemes

3.1. Introduction

Until now we have only considered the case of a flat bottom topography. This is unrealistic in practice; when computing flows in rivers, near shores or across ridges, bottom topography can have a large influence on the flow. The usual way of introducing a variable topography to the shallow water system (1.1) is to add a source term to the equations, obtaining (1.2). Figure 3.1 illustrates the one-dimensional setup.

Equation (1.2) is a *balance law* [**26**] – a conservation law with an additional source term:

$$U_t + f(U)_x + g(U)_y = -B(U, \mathbf{x}) \qquad \text{for } \mathbf{x} = (x, y) \in \mathbb{R}^2. \tag{3.1}$$

A crucial point of study for balance laws is that of steady states. A *steady state* is a solution of (3.1) that is constant in time. Inserting the ansatz $U_t \equiv 0$ in (3.1), we see that for such a state, the flux terms must exactly balance with the source term. Conversely, if these terms balance, then the solution is a steady state.

In the case of the full two-dimensional shallow water system with bottom topography, there is no simple criteria that ensures that a solution is a steady state. However, one very simple steady state is available, the so-called *lake at rest*

$$h + b \equiv \text{constant}, \qquad u = v = 0. \tag{3.2}$$

This represents a solution that is completely at equilibrium. The lake at rest appears in several contexts. For instance, the world's oceans are at large at rest, with relatively small perturbations on top of that.

When simulating flows with a variable topography, a crucial question is whether the numerical method will preserve steady states. It is well-known that most standard schemes such as the Roe and Rusanov schemes do not preserve even the lake

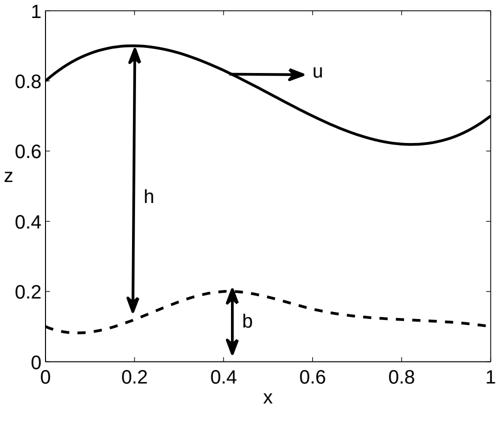


FIGURE 3.1. Height (solid line) and bottom (dashed line) in shallow water with a variable bottom topography.

at rest. Given initial data satisfying (3.2), these schemes will produce incorrect waves and oscillations that are of the order of discretization error. On the other hand, if a scheme *does* compute the lake at rest steady state correctly, then we may expect that slight perturbations of that state will also be correctly computed, and does not contain unphysical oscillations. Thus, the preservation of the lake at rest is essential for long-time simulations of near-steady flows.

The decisive step in designing finite volume schemes for the balance law (1.2) is the discretization of the source term. The general scheme is of the form

$$\frac{d}{dt}U_{i,j} = -\frac{1}{\Delta x}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) - \frac{1}{\Delta y}\left(G_{i,j+1/2} - G_{i,j-1/2}\right) - B_{i,j}, \tag{3.3}$$

where $B_{i,j}$ is a discrete version of the source term. In the presence of a steady state, the right-hand side should vanish, thus producing a solution that is constant in time.

Definition 3.1. A scheme of the form (3.3) is *well-balanced* if it preserves the lake at rest. In other words, given initial data

$$h_{i,j} + b_{i,j} \equiv \text{constant}, \qquad u_{i,j} = v_{i,j} = 0 \qquad \text{for all } i, j, \tag{3.4}$$

the solution computed by the scheme should satisfy

$$\frac{d}{dt}U_{i,j} = 0 \qquad \text{for all } i, j.$$

Remark 3.2. The lake at rest is far from being the only steady state of interest for (1.2). For instance, in the one-dimensional case, a general steady state is identified by

$$hu \equiv \text{constant}, \qquad g(h+b) - \frac{u^2}{2} \equiv \text{constant}.$$

We refer to [**12**] for well-balanced schemes for this steady state.

3.1.1. Energy preservation and stability. The theory of entropies for balance laws is similar to that of conservation laws. An entropy triple (E, Q^x, Q^y) for (3.1) consists of a convex $E : \mathbb{R}^n \times \mathbb{R}^2 \to \mathbb{R}$ such that

$$E(U,\mathbf{x})_t + Q^x(U,\mathbf{x})_x + Q^y(U,\mathbf{x})_y = 0 \qquad \text{for all } \mathbf{x} = (x,y) \in \mathbb{R}^2$$

whenever U is a smooth solution of (3.1). Similar to before, this entropy equality is replaced by an inequality for weak solutions.

The relevant entropy for (1.2) is the *total energy*

$$E(U,\mathbf{x}) = \frac{1}{2}\left(hu^2 + hv^2 + gh^2 + ghb(\mathbf{x})\right), \tag{3.5}$$

which is the sum of potential and kinetic energy. We obtain an entropy triple (E, Q^x, Q^y) by letting

$$Q^x(U,\mathbf{x}) = \frac{1}{2}\left(hu^3 + huv^2\right) + guh^2 + ghub(\mathbf{x})$$

and

$$Q^y(U,\mathbf{x}) = \frac{1}{2}\left(hu^2v + hv^3\right) + gvh^2 + ghvb(\mathbf{x}).$$

Note that these reduce to the energy (2.26) of the shallow water system with flat topography whenever $b \equiv 0$. As before, we define the entropy variables $V(U,\mathbf{x}) = E_U(U,\mathbf{x})$, which for the energy (3.5) is

$$V = \begin{bmatrix} g(h+b(\mathbf{x})) - \frac{u^2+v^2}{2} \\ u \\ v \end{bmatrix}, \tag{3.6}$$

(compare to (2.27)).

We study finite volume schemes for (1.2) that satisfy discrete entropy equalities and -inequalities for the energy (3.5). As for the homogeneous equation (1.1), such schemes cannot converge towards entropy violating solutions. What is more, energy preservation and -stability will imply stability for the approximations by bounding the total energy

$$\int E(U(x,y,t))\;dxdy$$

uniformly.

We will first consider the one-dimensional version of (1.2),

$$\begin{aligned} &h_t + (hu)_x = 0 \\ &(hu)_t + \left(\frac{1}{2}gh^2 + hu^2\right)_x = -ghb_x. \end{aligned} \tag{3.7}$$

In Sections 3.2 through 3.4, we modify the energy preserving and energy stable schemes developed in the previous chapter to obtain well-balanced schemes. The schemes are generalized to two dimensions in Section 3.5.

3.2. A well-balanced energy preserving scheme

The main challenge when designing schemes for (3.7) lies in the discretization of the source term. We will employ the discretization

$$B_i = \frac{g}{2}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}\frac{b_{i+1}-b_i}{\Delta x} + \overline{h}_{i-1/2}\frac{b_i-b_{i-1}}{\Delta x} \end{bmatrix} \tag{3.8}$$

(we denote $b_i = b(x_i)$). The EEC scheme with bottom topography is then

$$\frac{d}{dt}U_i = -\frac{1}{\Delta x}\left(\widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2}\right) - B_i, \tag{3.9}$$

where $\widetilde{F}_{i+1/2}$ is given by (2.13). We refer to this scheme as the EEC scheme in the remainder, as it reduces to the original EEC scheme in the case $b_i \equiv$ constant.

THEOREM 3.3. *The EEC scheme has the following properties.*

(i) It is consistent with (3.7) *and is second-order accurate.*

(ii) It is energy preserving, i.e. it satisfies the discrete entropy equality

$$\frac{d}{dt}E + \frac{1}{\Delta x}\left(\widetilde{H}_{i+1/2} - \widetilde{H}_{i-1/2}\right) = 0,$$

where

$$\widetilde{H}_{i+1/2} = \widetilde{Q}_{i+1/2} + g\overline{h}_{i+1/2}\frac{b_i u_{i+1} + b_{i+1}u_i}{2}$$

and $\widetilde{Q}$ is as in (2.14)*. $\widetilde{H}$ is consistent with Q.*

(iii) It is well-balanced.

PROOF. Consistency in this case simply means that $\widetilde{F}(U,U) = f(U)$, which is satisfied as the EEC scheme without bottom topography is consistent. By a simple Taylor expansion, the discretization (3.8) is of second order.

For (ii), we left-multiply (3.9) by $V_i^\top$ and get

$$\frac{d}{dt}E(U_i) = -\frac{1}{\Delta x}\left(\begin{bmatrix} gh_i - u_i^2/2 \\ u_i \end{bmatrix} + \begin{bmatrix} gb_i \\ 0 \end{bmatrix}\right)^\top \left(\widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2}\right) - V_i^\top B_i$$

(directly from the proof of Theorem 2.3)

$$\begin{aligned} &= -\frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\right) - \frac{1}{\Delta x}\begin{bmatrix} gb_i \\ 0 \end{bmatrix}^\top \left(\widetilde{F}_{i+1/2} - \widetilde{F}_{i-1/2}\right) - V_i^\top B_i \\ &= -\frac{1}{\Delta x}\left(\widetilde{Q}_{i+1/2} - \widetilde{Q}_{i-1/2}\right) - \frac{1}{\Delta x}gb_i\left(\overline{h}_{i+1/2}\overline{u}_{i+1/2} - \overline{h}_{i-1/2}\overline{u}_{i-1/2}\right) \end{aligned}$$

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

(after some cancellations)

$$\begin{aligned}
&= -\frac{1}{\Delta x}\Bigg(\left(\widetilde{Q}_{i+1/2} + g\overline{h}_{i+1/2}\frac{b_i u_{i+1} + b_{i+1}u_i}{2} \right) \\
&\qquad - \left(\widetilde{Q}_{i-1/2} + g\overline{h}_{i-1/2}\frac{b_{i-1} u_i + b_i u_{i-1}}{2} \right) \Bigg) \\
&= -\frac{1}{\Delta x}\left(\widetilde{H}_{i+1/2} - \widetilde{H}_{i-1/2} \right).
\end{aligned}$$

Lastly, assume that the lake at rest conditions $u_i \equiv 0$ and $(h+b)_i \equiv$ constant are satisfied. Then the first term on the right-hand side of (3.9) is

$$-\frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h^2}_{i+1/2} - \overline{h^2}_{i-1/2} \end{bmatrix}$$

Using the second identity in (1.28), we get that

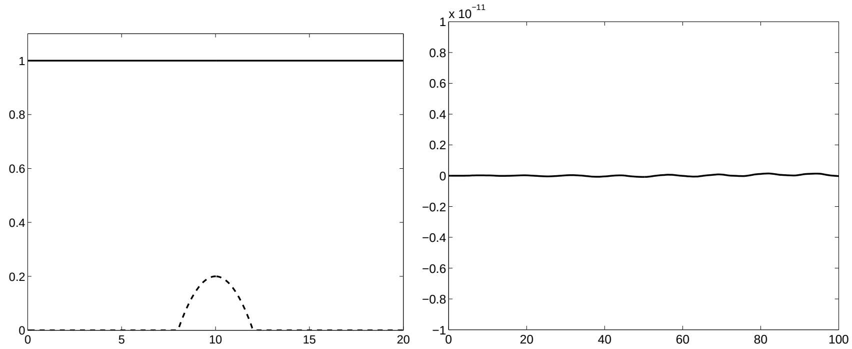
$$\begin{aligned}
\frac{d}{dt}U_i &= -\frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![h]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![h]\!]_{i-1/2} \end{bmatrix} \\
&\quad - \frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![b]\!]_{i-1/2} \end{bmatrix} \\
&= -\frac{g}{2\Delta x}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2}[\![h+b]\!]_{i+1/2} + \overline{h}_{i-1/2}[\![h+b]\!]_{i-1/2} \end{bmatrix} \\
&= 0,
\end{aligned}$$

since $[\![h+b]\!]_{i\pm 1/2} = 0$. □

3.2.1. Numerical experiments. We test the EEC scheme on a problem that has been featured in [**4, 18, 25**], among others. The bottom topography is given by

$$b(x) = \begin{cases} \frac{4-(x-10)^2}{20} & \text{if } |x-10| < 2 \\ 0 & \text{else,} \end{cases} \tag{3.10}$$

and we impose the lake at rest initial condition $h_i + b_i \equiv 1$, $u_i \equiv 0$. We compute on a mesh of 200 grid points up to time $t = 100$, and we use $g = 9.812$ as the gravitational constant. The result is displayed in Figure 3.2. There is no visible



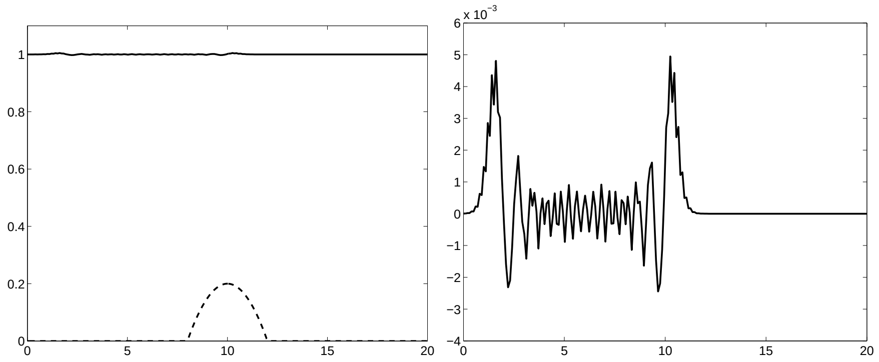
(a) Water level $h+b$ (solid line) and bottom topography (dotted line)

(b) Relative change in energy over time

FIGURE 3.2. Lake at rest at $t = 100$ using 200 grid points

change in the solution over time. Indeed, the relative change in energy is of the order of only 10^{-12}, as seen in Figure 3.2(b). This tiny error is solely due to floating point and temporal discretization errors.

Numerical experiment: Perturbed lake at rest. With the success of the EEC scheme in computing the lake at rest problem, we can expect that small perturbations on top of this steady state will also be computed correctly. We perturb the initial height by $+0.01$ in the region $|x-6|<1/4$. The perturbation should break up into two smaller waves, one left- and one right-going. As seen in Figure 3.3, this is exactly what happens. However, in between the waves there are unphysical oscillations that are due to a lack of numerical viscosity.



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

(b) Deviation from the steady state.

FIGURE 3.3. Lake at rest with perturbation at $t=1.5$.

3.3. A well-balanced energy stable scheme

Recall that the ERoe scheme has numerical flux

$$F^{\mathrm{ERoe}}_{i+1/2} = \widetilde{F}_{i+1/2} - \frac{1}{2}R|\Lambda|R^{\top}[\![V]\!]_{i+1/2}, \tag{3.11}$$

where R and Λ are evaluated at the average state

$$h=\overline{h}_{i+1/2}, \qquad u=\overline{u}_{i+1/2},$$

and V is the vector of energy variables (2.11). To generalize this scheme to the shallow water system with bottom topography, we replace the energy variables V with the energy variables for (3.7),

$$V=\begin{bmatrix} g(h+b(\mathbf{x}))-\frac{u^2}{2} \\ u \end{bmatrix}. \tag{3.12}$$

We refer to the resulting scheme as the ERoe scheme, as it reduces to (2.21) whenever $b\equiv$ constant. We have the following result for ERoe.

THEOREM 3.4. *The ERoe scheme satisfies the following properties.*

(i) It is consistent with (3.7) *and is first-order accurate.*

(ii) It satisfies the discrete energy dissipation estimate

$$\begin{aligned} &\frac{d}{dt}E_i + \frac{1}{\Delta x}\left(\widehat{H}_{i+1/2}-\widehat{H}_{i-1/2}\right) \\ &\quad = -\frac{1}{4\Delta x}[\![V]\!]^{\top}_{i+1/2}R_{i+1/2}|\Lambda_{i+1/2}|R^{\top}_{i+1/2}[\![V]\!]_{i+1/2} \\ &\qquad -\frac{1}{4\Delta x}[\![V]\!]^{\top}_{i-1/2}R_{i-1/2}|\Lambda_{i-1/2}|R^{\top}_{i-1/2}[\![V]\!]_{i-1/2} \\ &\quad \leq 0, \end{aligned} \tag{3.13}$$

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

$$\begin{aligned}\widehat{H}_{i+1/2} &= \overline{V}^{\top}_{i+1/2}F_{i+1/2} - \overline{\Psi}_{i+1/2} - \frac{g}{4}\overline{h}_{i+1/2}[\![u]\!]_{i+1/2}[\![b]\!]_{i+1/2} \\ &\quad + \frac{1}{2}\overline{V}^{\top}_{i+1/2}R_{i+1/2}|\Lambda_{i+1/2}|R^{\top}_{i+1/2}[\![V_{i+1/2}]\!].\end{aligned} \tag{3.14}$$

(iii) It is well-balanced.

Proof. The proof of (i) is straightforward. The proof of (3.13) follows the proof of the corresponding result in Theorem 2.9 and we omit the details here. To prove (iii), we assume that the data satisfies $u_i \equiv 0$ and $h_i + b_i \equiv$ constant. Then we have

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

Consequently, by the definition of the energy variables,

$$[\![V]\!]_{i+1/2} \equiv 0 \qquad \text{for all } i.$$

Hence the diffusion operator in (3.11) drops out, and the scheme reduces to the EEC scheme. Thus, by Theorem 3.3 (iii), we have

$$\frac{d}{dt}h_i \equiv 0 \qquad \text{and} \qquad \frac{d}{dt}(h_iu_i) \equiv 0 \qquad \text{for all } i.$$

□

3.4. A well-balanced, second-order accurate scheme

We continue with the second-order accurate ERoe2 scheme. Recall from Chapter 2 that this scheme uses a linear reconstruction of the energy variables to obtain left and right states V_i^W and V_i^E in each grid cell. The resulting flux is then

$$F^{\text{ERoe2}}_{i+1/2} = \widetilde{F}(U_i, U_{i+1}) - \frac{1}{2}R|\Lambda|R^{\top}\left(V^W_{i+1} - V^E_i\right). \tag{3.15}$$

The motivation for performing reconstruction in the energy variables rather than the conserved variables has been postponed until now: By reconstructing in the energy variables (3.12), the ERoe2 scheme is well-balanced.

Theorem 3.5. *The ERoe2 scheme* (3.15) *is consistent with* (3.7)*, second-order accurate and well-balanced.*

Proof. Consistency is trivial. Using a Taylor expansion proves the second-order accurate local truncation error.

To prove that the ERoe2 scheme preserves the lake at rest, we observe that when the data satisfies $u_i \equiv 0$ and $h_i + b_i \equiv$ constant, we have

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

whence $[\![V]\!]_{i+1/2} \equiv 0$. Therefore, by the non-oscillatory property of the slope limiter, we obtain $\sigma_i \equiv 0$, so

$$V_i^E = V_i^W = V_i = \text{constant} \qquad \text{for all } i.$$

As a consequence, the term $\left(V^W_{i+1} - V^E_i\right)$ in the diffusion operator in (3.15) vanishes, and we follow the argument in the proof of (iii) in Theorem 3.3 to conclude that

$$\frac{d}{dt}h_i \equiv 0 \qquad \text{and} \qquad \frac{d}{dt}(h_iu_i) \equiv 0 \qquad \text{for all } i.$$

Hence, the discrete lake at rest is preserved. □

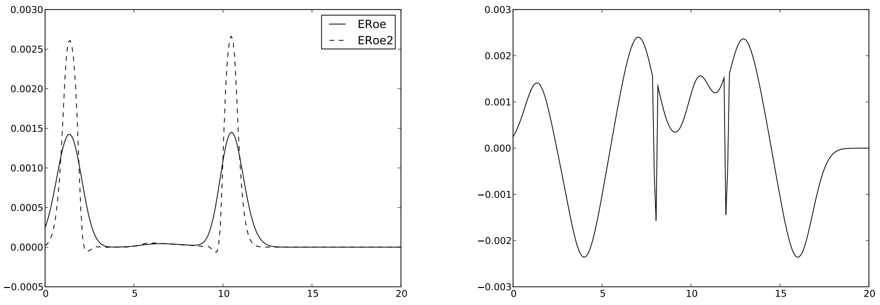
N	Roe	EEC	ERoe	ERoe2
50	2.76e-2	6.27e-14	1.92e-18	3.17e-16
100	7.60e-3	1.62e-13	2.14e-18	4.48e-17
200	2.02e-3	6.74e-13	3.35e-18	2.34e-16
400	5.15e-4	1.76e-12	2.22e-17	1.04e-15

TABLE 3.1. The L^1 error in height for the lake at rest with different schemes on a sequence of meshes at time $t = 10$.

3.4.1. Numerical experiments.

3.4.1. Numerical experiments. We repeat the lake at rest problem of Section 3.2.1 and compare the Roe, EEC, ERoe and ERoe2 schemes on a sequence of meshes. As we have proved that the latter three schemes are well-balanced, they should preserve the steady state exactly. We display the difference at $t = 10$ from the initial data in the L^1 norm in Table 3.1. The error in the well-balanced schemes are of the order of machine precision, as expected. The Roe scheme has errors of the order of truncation error, due to its lack of well-balancing.

Numerical experiment: Perturbed lake at rest. In the perturbed lake at rest problem, the EEC scheme produced an unacceptable amount of oscillations. These were caused by a lack of numerical diffusion. We repeat the experiment with the ERoe and ERoe2 schemes to see the effect that the well-balanced diffusion operators have on the EEC scheme. The result is displayed in Figure 3.4(a). Comparing to Figure 3.3(a), we see that all oscillations are gone. The second-order accurate ERoe2 scheme resolves the two waves much more sharply than the ERoe scheme. Due to the well-balancing of both schemes, the small-scale perturbation is resolved correctly without any disturbance from spurious waves. Contrast this with the unbalanced Roe scheme (Figure 3.4(b)), for which unphysical waves ruin the solution.



(a) ERoe and ERoe2

(b) Roe

FIGURE 3.4. Deviation from the lake at rest steady state at $t = 1.5$ on a mesh of 200 mesh points.

3.5. Generalization to two spatial dimensions

Next we generalize the ideas of the previous sections to obtain well-balanced, energy preserving and energy stable schemes for the full two-dimensional shallow water equations (1.2). The two-dimensional counterparts of the EEC, ERoe and ERoe2 schemes are very similar to the one-dimensional schemes, and so the discussion will be brief.

The natural generalization of the source term discretization (3.8) is

$$B_{i,j} = \frac{g}{2}\begin{bmatrix} 0 \\ \overline{h}_{i+1/2,j}\frac{b_{i+1,j}-b_{i,j}}{\Delta x} + \overline{h}_{i-1/2,j}\frac{b_{i,j}-b_{i-1,j}}{\Delta x} \\ \overline{h}_{i,j+1/2}\frac{b_{i,j+1}-b_{i,j}}{\Delta y} + \overline{h}_{i,j-1/2}\frac{b_{i,j}-b_{i,j-1}}{\Delta y}\end{bmatrix}. \tag{3.16}$$

Nothing is altered in the flux of the two-dimensional EEC scheme. In the ERoe and ERoe2 schemes we replace the vector of energy variables V by (3.6). The schemes are then of the form (3.3), with either the EEC, ERoe or ERoe2 numerical fluxes in place of F and G. Two-dimensional counterparts of theorems 3.3, 3.4 and 3.5 are easily obtained.

THEOREM 3.6. *The two-dimensional EEC scheme has the following properties.*

(i) It is consistent with (1.2) *and is second-order accurate.*

(ii) It is energy preserving, i.e. it satisfies the discrete entropy equality

$$\frac{d}{dt}E + \frac{1}{\Delta x}\left(\widetilde{H}^x_{i+1/2,j} - \widetilde{H}^x_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(\widetilde{H}^y_{i,j+1/2} - \widetilde{H}^y_{i,j-1/2}\right) = 0,$$

where

$$\widetilde{H}^x_{i+1/2,j} = \widetilde{Q}^x_{i+1/2,j} + g\overline{h}_{i+1/2,j}\frac{b_{i,j}u_{i+1,j} + b_{i+1,j}u_{i,j}}{2}$$

and

$$\widetilde{H}^y_{i,j+1/2} = \widetilde{Q}^y_{i,j+1/2} + g\overline{h}_{i,j+1/2}\frac{b_{i,j}u_{i,j+1} + b_{i,j+1}u_{i,j}}{2},$$

and $\widetilde{Q}^x$ and $\widetilde{Q}^y$ are as in (2.29). *$\widetilde{H}^x$ and $\widetilde{H}^y$ are consistent with Q^x and Q^y, respectively.*

(iii) It is well-balanced.

THEOREM 3.7. *The two-dimensional ERoe scheme satisfies the following properties.*

(i) It is consistent with (1.2) *and is first-order accurate.*

(ii) It satisfies the discrete energy dissipation estimate

$$\frac{d}{dt}E + \frac{1}{\Delta x}\left(\widehat{H}^x_{i+1/2,j} - \widehat{H}^x_{i-1/2,j}\right) + \frac{1}{\Delta y}\left(\widehat{H}^y_{i,j+1/2} - \widehat{H}^y_{i,j-1/2}\right) \leq 0,$$

where $\widehat{H}^x$ and $\widehat{H}^y$ are consistent with Q^x and Q^y, respectively

(iii) It is well-balanced.

THEOREM 3.8. *The two-dimensional ERoe2 scheme is consistent with* (1.2)*, second-order accurate and well-balanced.*

3.5.1. Numerical experiments. We test the well-balancing of the three schemes in a two-dimensional lake at rest problem that has been featured in [**25, 30**]. The bottom topography is given by

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

and we use the lake at rest initial data

$$h_{i,j} + b_{i,j} \equiv 1, \qquad u_{i,j} \equiv v_{i,j} \equiv 0 \qquad \text{for all } i,j.$$

This setup is illustrated in Figure 3.5. We set $g = 9.812$ and compute in the domain $(x,y) \in [0,2]\times[0,1]$, using standard Neumann boundary conditions. We compute with the Roe, EEC, ERoe and ERoe2 schemes on a sequence of meshes up to time $t = 1$. The deviation from the steady state is displayed in Table 3.2. Clearly the three well-balanced schemes preserve the initial data to machine precision, whereas the Roe scheme produces spurious waves in the solution.

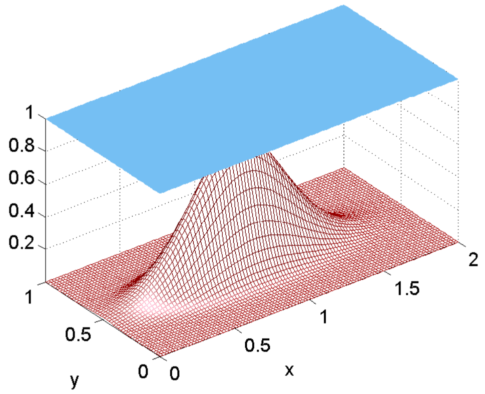


FIGURE 3.5. Water level and bottom topography for the two-dimensional lake at rest.

N	Roe	EEC	ERoe	ERoe2
50	$1.71 \cdot 10^{-1}$	$2.30 \cdot 10^{-15}$	$2.95 \cdot 10^{-15}$	$3.53 \cdot 10^{-15}$
100	$8.73 \cdot 10^{-2}$	$3.50 \cdot 10^{-14}$	$3.48 \cdot 10^{-15}$	$5.76 \cdot 10^{-15}$
200	$5.81 \cdot 10^{-2}$	$2.06 \cdot 10^{-11}$	$3.95 \cdot 10^{-15}$	$4.70 \cdot 10^{-15}$

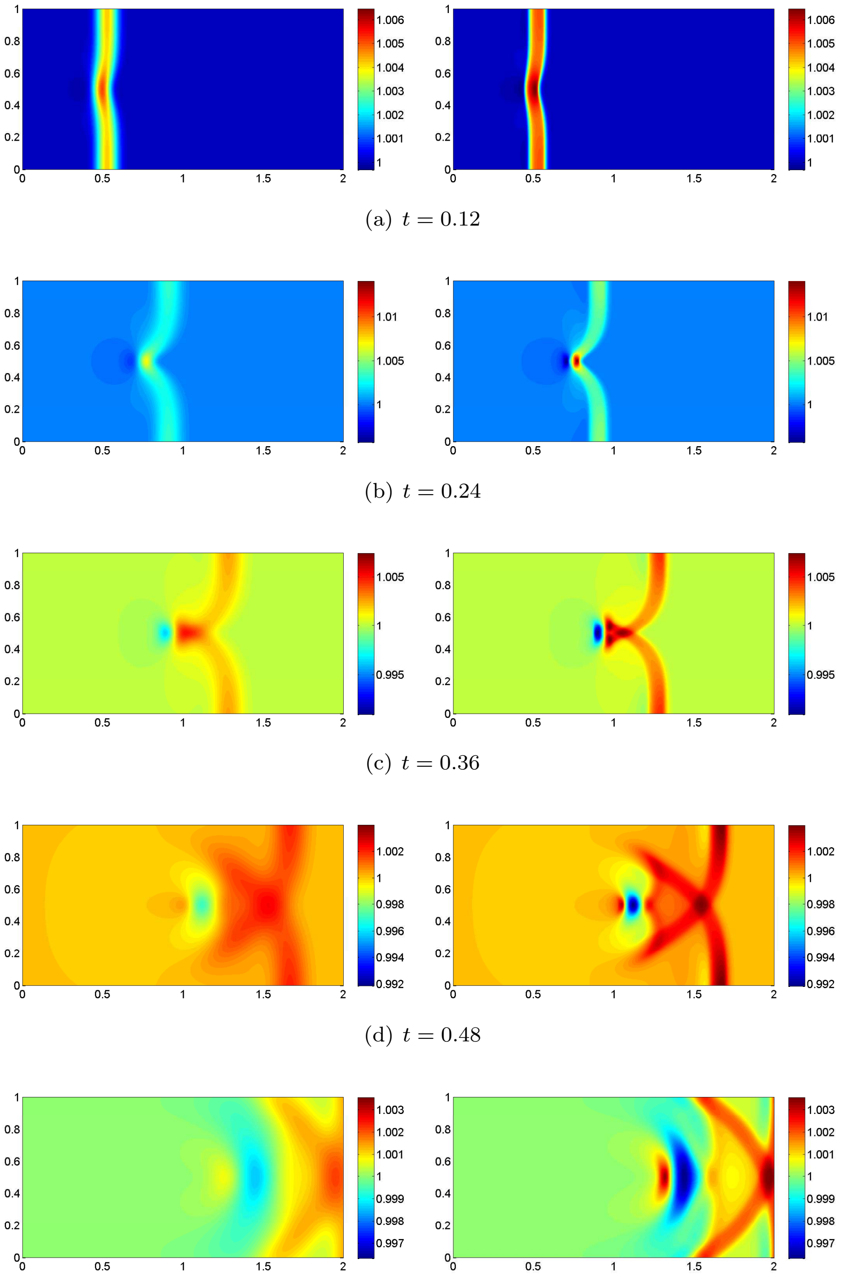
TABLE 3.2. The L^1 error in height for the two-dimensional lake at rest with different schemes on a sequence of $2N \times N$ meshes at time $t = 1$.

Numerical experiment: Perturbed lake at rest. Lastly, we compute a perturbed version of the lake at rest. We add $+0.01$ to the initial height in the region $x \in [0.1, 0.2]$. This perturbation should break up into two smaller waves moving in either direction. The left-going wave hits the boundary at approximately $t = 0.03$; this will test how well the schemes handle the boundary condition. The right-going wave will go over the bump in the bottom topography, creating a complex wave pattern.

Figure 3.6 shows the computed height with the ERoe and ERoe2 schemes on a mesh of 600×300 mesh points. At $t = 0.12$, the left-going wave has cleanly left the domain without any trace of bounce-back waves. In the remaining snapshots, the ERoe2 scheme clearly gives the sharpest resolution of the flow. The results coincide well with the results of [**25**] and [**30**].

3.6. Conclusion

We have designed novel well-balanced, energy preserving and energy stable finite volume schemes for the shallow water equations with bottom topography in one and two space dimensions. The only difference from their homogeneous counterparts is a simple discretization of the source term, and so they inherit many of the properties of the schemes of Chapter 2, including simplicity and accuracy.



(a) $t = 0.12$

(b) $t = 0.24$

(c) $t = 0.36$

(d) $t = 0.48$

(e) $t = 0.60$

FIGURE 3.6. A simulation of the two-dimensional lake at rest with perturbation using the ERoe and ERoe2 scheme with 600×300 mesh points. Left column: ERoe, right column: ERoe2. (Note that the scales in the figures at each point in time are different.)

CHAPTER 4

Vorticity preservation

4.1. Introduction

When numerically approximating physical models, order of convergence is important. However, just as important is the path of convergence. Approximate solutions may lie near the correct solution in some L^p norm, but still may lie in an entirely physically incorrect part of the phase space. Thus, it is desirable that the scheme converges along a path over which the properties of the approximations resemble the true solution. This is especially important in simulations of non-deterministic systems such as weather systems. In short-range simulations, grid size and order of convergence are the key factors in the accurate resolution of the flow field. However, in simulations over longer periods of time it is the overall statistical properties of the flow that are important [**2**]. We have already considered consistency of numerical methods with respect to the correct preservation/dissipation of energy. In this chapter we consider the correct resolution of vorticity in the flow.

Recall that the essence of the finite volume method is a partition of the computational domain into polygonal control volumes. By approximating the solution of the differential equation by an average over each control volume, the focus is shifted towards computing the correct flux through the edges of the control volume. This decoupling into normal directions may lead to a poor resolution of tangential flows or forces. As a result, or perhaps as a cause, the vorticity, which satisfies its own evolution equation, is not resolved correctly.

Of the more recent approaches to resolving vortical flows are the projection methods [**20**]. These were originally developed by Chorin [**9**] and Bell, Colella and Glaz [**6**] to impose the divergence constraint in the incompressible Navier-Stokes equations, but have also been studied by Brackbill and Barnes [**7**] and Toth [**39**] in the context of the MHD equations. The projection method introduces a correction term to a finite volume scheme to control the unphysical generation of vorticity. At each time step we solve simultaneously for the flow field and the vorticity. By applying a projection of the flow field in the form of an elliptic operator, we obtain – to within a discretization error – a solution with a correct vorticity field.

Denoting the vorticity of (1.1) by $\omega = v_x - u_y$, one may readily calculate that smooth solutions of the shallow water equations satisfy

$$\omega_t + (u\omega)_x + (v\omega)_y = 0. \tag{4.1}$$

Standard finite volume schemes might not respect this relation, which might lead to an accumulation of errors over time [**20**]. However, the relation (4.1) is hard to impose numerically, so instead we aim to control the *pseudo-vorticity* of the shallow water equations. The pseudo-vorticity of a flow field is defined as the curl of the momentum, in contrast to vorticity, which is the curl of velocity. As momentum, and not velocity, is a conserved variable, it will be easier to control errors in pseudo-vorticity, rather than vorticity. For the rest of this chapter we refer to pseudo-vorticity simply as *vorticity.* We denote vorticity by $\Omega = (m_2)_x - (m_1)_y$, where $m_1 = hu$ and $m_2 = hv$ are the momentum in the $x-$ and $y-$directions, respectively. Then, the change of vorticity over time is

$$\Omega_t = (m_2)_{xt} - (m_1)_{yt} = ((m_2)_t)_x - ((m_1)_t)_y,$$

and inserting $(m_1)_t$ and $(m_2)_t$ from (1.1) and rearranging, we find that the vorticity of the shallow water equations satisfies the evolution equation

$$\begin{aligned}\Omega_t + (u\Omega)_x + (v\Omega)_y + \big(m_2(u_x+v_y)\big)_x - \big(m_1(u_x+v_y)\big)_y \\ + \left(\frac{u^2+v^2}{2}h_y\right)_x - \left(\frac{u^2+v^2}{2}h_x\right)_y = 0.\end{aligned} \tag{4.2}$$

Before presenting the projection method for the shallow water equations, we will consider a simpler, linearized version of the equations, the system wave equation. This system exhibits a particularly simple evolution equation for vorticity. Thus, the exposition will be as transparent as possible.

If the solution of a conservation law (or any other differential equation) consists of relatively small perturbations on top of a steady state, then the flow is mainly governed by linear effects; see LeVeque [**26**]. By *linearizing* the equations around the underlying steady state, we obtain a much simpler, linear conservation law that is still a good approximation of the original problem. Thus, assume that the solution of (1.9) is of the form $U = U^0 + \widetilde{U}$, where U^0 is a constant, steady state and $\widetilde{U}(\mathbf{x},t) = \left[\widetilde{h}, \widetilde{m}_1, \widetilde{m}_2\right]^\top$ is small. Inserting U into (1.9) and removing terms of order $\|\widetilde{U}\|^2$ gives the conservation law

$$\widetilde{U}_t + f'(U^0)\widetilde{U}_x + g'(U^0)\widetilde{U}_y = 0.$$

Assuming that the velocity of the background flow is zero, we have

$$f'(U^0) = \begin{bmatrix} 0 & 1 & 0 \\ gh^0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad g'(U^0) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ gh^0 & 0 & 0 \end{bmatrix}.$$

Denote the wave speed in the fluid by $c = \sqrt{gh^0}$, and let $\zeta = c\widetilde{h}$, $m_1 = \widetilde{m}_1$ and $m_2 = \widetilde{m}_2$. Then $[\zeta, m_1, m_2]^\top$ solves the (equivalent) conservation law

$$\begin{bmatrix} \zeta \\ m_1 \\ m_2 \end{bmatrix}_t + \begin{bmatrix} 0 & c & 0 \\ c & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \zeta \\ m_1 \\ m_2 \end{bmatrix}_x + \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & 0 \\ c & 0 & 0 \end{bmatrix} \begin{bmatrix} \zeta \\ m_1 \\ m_2 \end{bmatrix}_y = 0. \tag{4.3}$$

This is the system wave equation. The connection between this and the wave equation

$$u_{tt} = c^2(u_{xx} + u_{yy})$$

is seen by letting $\zeta = u_t$, $m_1 = -cu_x$ and $m_2 = -cu_y$; this gives precisely (4.3).

The Jacobians of the flux functions of (4.3) both have eigenvalues

$$\lambda_1 = -c, \qquad \lambda_2 = 0, \qquad \lambda_3 = c,$$

so the equation is hyperbolic. In the remainder, we drop mention of the background flow U^0 and consider (4.3) as a conservation law in itself with solution $U = [\zeta, m_1, m_2]^\top$.

The system wave equation has a very simple expression for vorticity. Writing $\Omega = (m_2)_x - (m_1)_y$ as before, we have

$$\Omega_t = -\big((c\zeta_y)_x - (c\zeta_x)_y\big) = 0,$$

and so Ω is constant in time.

4.1.1. Vorticity of the Rusanov scheme. As the wave equation is hyperbolic, all the theory of finite volume schemes from Section 1.3 apply to this system. We will use the Roe and the Rusanov schemes. Recall that the Rusanov scheme uses a (rather coarse) approximation of the local wave speed, the maximum of the neighboring eigenvalues. For the wave equation the maximum eigenvalue is always c, and so the Rusanov flux in each direction is

$$F^{\mathrm{Rus}}_{i+1/2,j} = \frac{1}{2}\big(f(U_{i,j}) + f(U_{i+1,j})\big) - \frac{c}{2}(U_{i+1,j} - U_{i,j})$$

and

$$G^{\mathrm{Rus}}_{i,j+1/2} = \frac{1}{2}\big(g(U_{i,j}) + g(U_{i,j+1})\big) - \frac{c}{2}(U_{i,j+1} - U_{i,j}).$$

We investigate the vorticity preserving properties of the Rusanov scheme. For discretizing vorticity we use the second-order approximation

$$\Omega_{i,j} = D_x(m_2)_{i,j} - D_y(m_1)_{i,j}, \tag{4.4}$$

where D_x and D_y are the central differences

$$D_x U_{i,j} = \frac{1}{\Delta x}\mu_x \delta_x U_{i,j} = \frac{U_{i+1,j} - U_{i-1,j}}{2\Delta x} \tag{4.5a}$$

and

$$D_y U_{i,j} = \frac{1}{\Delta y}\mu_y \delta_y U_{i,j} = \frac{U_{i,j+1} - U_{i,j-1}}{2\Delta y}, \tag{4.5b}$$

and $\delta_x, \delta_y, \mu_x$ and μ_y are the difference and average operators

$$\delta_x u_{i+1/2,j} = u_{i+1,j} - u_{i,j}, \qquad \delta_y u_{i,j+1/2} = u_{i,j+1} - u_{i,j},$$
$$\mu_x u_{i+1/2,j} = \frac{u_{i,j} + u_{i+1,j}}{2}, \qquad \mu_y u_{i,j+1/2} = \frac{u_{i,j} + u_{i,j+1}}{2}.$$

Central differences are used to avoid the need to stagger variables. With this discretization of vorticity, we have the following dissipation estimate for the Rusanov scheme.

LEMMA 4.1. *Solutions of* (4.3) *computed by the Rusanov scheme satisfy*

$$\frac{d}{dt}\Omega_{i,j} = \frac{c}{2}\left(\frac{1}{\Delta x}\delta_x^2 \Omega_{i,j} + \frac{1}{\Delta y}\delta_y^2 \Omega_{i,j}\right) \tag{4.6}$$

in the interior of the domain. In particular, if the initial vorticity is zero, then it stays zero at all times. Moreover, if $\Omega_{i,j} = 0$ *at the boundary, then the energy of* Ω *is non-increasing,*

$$\begin{aligned}\frac{d}{dt}\left(\Delta x \Delta y \sum_{i,j} \Omega_{i,j}^2\right) &= -c\sum_{i,j}\Big(\Delta y\,(\Omega_{i+1,j} - \Omega_{i,j})^2 + \Delta x\,(\Omega_{i,j+1} - \Omega_{i,j})^2\Big) \\ &\le 0,\end{aligned}$$

with equality only when $\Omega \equiv 0$.

PROOF. Inserting the definition of F^{Rus} and G^{Rus} into the finite volume scheme (1.21) and rearranging, we find that

$$\begin{aligned}\frac{d}{dt}(m_1)_{i,j} &= \frac{c}{2}\left(\frac{1}{\Delta x}\delta_x\Big(\delta_x(m_1)_{i,j} - 2\mu_x\zeta_{i,j}\Big) + \frac{1}{\Delta y}\delta_y^2(m_1)_{i,j}\right), \\ \frac{d}{dt}(m_2)_{i,j} &= \frac{c}{2}\left(\frac{1}{\Delta y}\delta_y\Big(\delta_y(m_2)_{i,j} - 2\mu_y\zeta_{i,j}\Big) + \frac{1}{\Delta x}\delta_x^2(m_2)_{i,j}\right).\end{aligned}$$

Inserting this into $\frac{d}{dt}\Omega_{i,j} = \frac{d}{dt}\left(D_x(m_2)_{i,j} - D_y(m_1)_{i,j}\right)$ and using the fact that δ and μ commute, we get (4.6). If $\Omega_{i,j}(0) \equiv 0$, then (4.6) reads $\frac{d}{dt}\Omega = 0$, so we get $\Omega \equiv 0$.

For the energy dissipation estimate, we remark that for a one-dimensional grid function u_i we have

$$\sum_{i=1}^{N} u_i \delta^2 u_i = -\sum_{i=1}^{N-1} (u_{i+1} - u_i)^2 + u_N(u_{N+1} - u_N) - u_1(u_1 - u_0). \tag{4.7}$$

This implies that

$$\begin{aligned}\frac{d}{dt}\left(\Delta x \Delta y \sum_{i,j} \Omega_{i,j}^2\right) &= 2\Delta x \Delta y \sum_{i,j} \Omega_{i,j} \frac{d}{dt}\Omega_{i,j} \\ &= c\Delta x \Delta y \sum_{i,j} \Omega_{i,j} \left(\frac{1}{\Delta x}\delta_x^2 \Omega_{i,j} + \frac{1}{\Delta y}\delta_y^2 \Omega_{i,j}\right)\end{aligned}$$

(by (4.7); the boundary terms drop out)

$$= -c \sum_{i,j} \left(\Delta y \left(\Omega_{i+1,j} - \Omega_{i,j}\right)^2 + \Delta x \left(\Omega_{i,j+1} - \Omega_{i,j}\right)^2\right).$$

□

The vorticity of the wave equation should be constant in time, and by the above lemma, the Rusanov scheme respects this whenever the initial vorticity is zero. However, if it is nonzero at the initial time, then the scheme will incorrectly dissipate vorticity. What is more, the energy dissipation estimate relies on having zero vorticity at the boundary. If this is not the case, then in many cases there will be an excessive amount of vorticity production at the border. A prime example is when using Neumann boundary conditions, simply copying the conserved variables onto the boundary. As we will see later, the Rusanov scheme will produce a large amount of vorticity along the boundary. In the case of a periodic boundary, the boundary terms in (4.7) vanish and vorticity will be dissipated. In either case, vorticity is not kept constant, as it should be. The projection method, which we describe next, will correct this.

4.2. Vorticity projection for the wave equation

We motivate the discrete projection method by first considering the continuous problem. Let $\Omega = \Omega(\mathbf{x}, t)$ be the vorticity at the initial time step. Given a solution $U^n = [h^n, m_1^n, m_2^n]^\top$ at time t_n, we compute a candidate solution $\widetilde{U}^{n+1} = \left[\widetilde{h}, \widetilde{m}_1, \widetilde{m}_2\right]^\top$ at the next time step t_{n+1} with any method available – not necessarily numerical. We project the momentum field $\widetilde{m} = (\widetilde{m}_1, \widetilde{m}_2)$ onto the space of functions with curl equal to Ω as follows. Let ψ be the solution of the Poisson equation

$$-\Delta\psi = \widetilde{\Omega} - \Omega,$$

where Δ is the Laplace operator and $\widetilde{\Omega} = (\widetilde{m}_2)_x - (\widetilde{m}_1)_y$ is the vorticity of $\widetilde{U}$. Define $U^{n+1} = [h, m_1, m_2]^\top$ as

$$\zeta = \widetilde{\zeta}, \qquad m_1 = \widetilde{m}_1 - \psi_y, \qquad m_2 = \widetilde{m}_2 + \psi_x.$$

Observe that this makes ψ the stream function of the field $\widetilde{m} - m$. In other words, streamlines of $\widetilde{m} - m$ coincide with level curves of ψ. The vorticity of U^{n+1} is then

$$(m_2)_x - (m_1)_y = \widetilde{\Omega} + \Delta\psi = \Omega,$$

and so U^{n+1} is an approximate solution at the next time step with correct vorticity. This process is repeated at each time step.

At the boundary of the computational domain we set $U^{n+1} = \widetilde{U}$. As ψ is the stream function of $\widetilde{m} - m$, this makes the boundary of the domain a level curve of

ψ. Hence, by adding a constant if necessary, we may assume that ψ vanishes at the boundary.

4.2.1. Discrete projection method. We discretize the projection method so that the discretized vorticity (4.4) will be preserved exactly. As $\Omega_{i,j}$ should be constant in time, we need only make sure that the vorticity is equal to $\Omega_{i,j} = \Omega_{i,j}(0)$ at all times. Given a solution at time t_n, the solution at time t_{n+1} is computed as follows.

- Compute a candidate solution $\widetilde{U}_{i,j}^{n+1}$ at time t_{n+1} with any consistent finite volume scheme. Let $\widetilde{\Omega}_{i,j}^{n+1}$ be its vorticity, computed with the formula (4.4).
- Find the solution $\psi = (\psi_{i,j})$ of the discrete Poisson problem

$$\begin{cases} -\left(D_x^2 + D_y^2\right)\psi_{i,j} = \widetilde{\Omega}_{i,j}^{n+1} - \Omega_{i,j} & \text{in the interior,} \\ \psi_{i,j} = 0 & \text{at the boundary,} \end{cases} \tag{4.8}$$

where D_x and D_y are the discrete derivatives (4.5).

- Define the solution at the next time step $U_{i,j}^{n+1} = [\zeta_{i,j}, (m_1)_{i,j}, (m_2)_{i,j}]^\top$ as

$$\zeta_{i,j} = \widetilde{\zeta}_{i,j}, \qquad (m_1)_{i,j} = (\widetilde{m}_1)_{i,j} - D_y\psi_{i,j}, \qquad (m_2)_{i,j} = (\widetilde{m}_2)_{i,j} + D_x\psi_{i,j}.$$

PROPOSITION 4.2. *The solution $U = (\zeta, m_1, m_2)$ computed by the projection method satisfies*

$$D_x(m_2)_{i,j} - D_y(m_1)_{i,j} = \Omega_{i,j}.$$

Hence, the solution has the correct amount of vorticity.

PROOF. Simply inserting the definition of m_1 and m_2 gives

$$\begin{aligned} D_x(m_2)_{i,j} - D_y(m_1)_{i,j} &= D_x(\widetilde{m}_2)_{i,j} - D_y(\widetilde{m}_1)_{i,j} + \left(D_x^2 + D_y^2\right)\psi_{i,j} \\ &= \Omega_{i,j}. \end{aligned}$$

□

We denote the projection method using the Rusanov scheme in the prediction step as the *VPRus* scheme, and similarly for *VPRoe*.

Recall from Section 2.2 that the energy $E = \frac{1}{2}U^\top U$ is an entropy for conservation laws with linear and symmetric fluxes. As the wave equation (4.3) is precisely this, we have an energy for the wave equation. Next, we show the stability result that in the case of zero initial vorticity, the projection step of the VP methods does not introduce additional energy to the solution.

PROPOSITION 4.3. *Let $E = \frac{1}{2}\left(\zeta^2 + m_1^2 + m_2^2\right)$ be the energy of a solution computed by the projection method, and let $\widetilde{E} = \frac{1}{2}\left(\widetilde{\zeta}^2 + \widetilde{m}_1^2 + \widetilde{m}_2^2\right)$ be the energy of the solution obtained in the prediction step. Then*

$$\sum_{i,j} E_{i,j} \leq \sum_{i,j} \widetilde{E}_{i,j} - 2\sum_{i,j} \Omega_{i,j}\psi_{i,j}.$$

In particular, if the initial vorticity is zero, then

$$\sum_{i,j} E_{i,j} \leq \sum_{i,j} \widetilde{E}_{i,j}.$$

PROOF. We use summation by parts extensively. As $\psi = 0$ at the boundary, all boundary terms will drop out. For notational convenience, we leave out variable indexes. Then

$$\sum_{i,j} E = \sum_{i,j} \left(\zeta^2 + (\widetilde{m}_1 - D_y\psi)^2 + (\widetilde{m}_2 + D_x\psi)^2\right)$$

$$=\sum_{i,j}\widetilde{E}-2\sum_{i,j}(\widetilde{m}_1 D_y\psi-\widetilde{m}_2 D_x\psi)+\sum_{i,j}\left((D_x\psi)^2+(D_y\psi)^2\right).$$

The second term is

$$\begin{aligned}-2\sum_{i,j}(\widetilde{m}_1 D_y\psi-\widetilde{m}_2 D_x\psi)&=2\sum_{i,j}\psi(D_y\widetilde{m}_1-D_x\widetilde{m}_2)\\&=-2\sum_{i,j}\psi\left(\Omega-(D_x^2+D_y^2)\psi\right)\\&=-2\sum_{i,j}\Omega\psi-2\sum_{i,j}\left((D_x\psi)^2+(D_y\psi)^2\right).\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{i,j}E&=\sum\widetilde{E}-2\sum_{i,j}\Omega\psi-\sum_{i,j}\left((D_x\psi)^2+(D_y\psi)^2\right)\\&\le\sum_{i,j}\widetilde{E}-2\sum_{i,j}\Omega\psi.\end{aligned}$$

If $\Omega_{i,j}\equiv 0$, then the last term drops out. □

4.2.2. Solving for the stream function. The discrete Poisson equation (4.8) must be solved at every time step, so a fast solver is essential to the computational efficiency of the scheme. Assume we are computing on an $N \times M$ uniform Cartesian grid. If we write $C_{i,j}=\widehat{\Omega}_{i,j}^{n+1}-\Omega_{i,j}$, then the problem can be rewritten as

$$\frac{1}{4\Delta x^2}T_N\Psi+\frac{1}{4\Delta y^2}\Psi T_M=C, \tag{4.9}$$

where $\Psi=(\psi_{i,j})_{i,j}$ and T_N is the $N\times N$ symmetric, positive definite matrix

$$T_N=\begin{bmatrix}2&0&-1&&&&&\\0&2&0&-1&&&&\\-1&0&2&0&-1&&&\\&\ddots&&\ddots&&\ddots&&\\&&-1&0&2&0&-1\\&&&-1&0&2&0\\&&&&-1&0&2\end{bmatrix}. \tag{4.10}$$

We solve this equation with a method from [**28**] that requires $O(N^3)$ floating point operations. Let R_N be the matrix of eigenvectors and $D_N=\operatorname{diag}(\lambda_N^1,\dots,\lambda_N^N)$ the matrix of eigenvalues of T_N. R_N may be chosen such that $R_N^2=I_N$, the identity matrix in $\mathbb{R}^{N\times N}$. Multiplying (4.9) by R_N on the left and R_M on the right, we find that

$$\frac{1}{4\Delta x^2}D_N X+\frac{1}{4\Delta y^2}XD_M=R_N C R_M,$$

where $X=R_N\Psi R_M$. The (i,j) entry of the left-hand side of this equation is

$$\frac{\lambda_N^i}{4\Delta x^2}X_{i,j}+\frac{\lambda_M^j}{4\Delta y^2}X_{i,j}.$$

Hence, $X=(R_N C R_M)./S$, where

$$S_{i,j}=\frac{\lambda_N^i}{4\Delta x^2}+\frac{\lambda_M^j}{4\Delta y^2}$$

and ./ denotes component-wise division. The solution of (4.9) is therefore

$$\Psi=R_N X R_M.$$

4.2.3. Periodic boundary conditions. The above discussion did not rely on the specifics of the boundary condition on U. In the special case of a periodic boundary,

$$U_{0,j} = U_{N,j}, \qquad U_{-1,j} = U_{N-1,j}$$

etc., we can improve the method by applying the same condition on ψ. This will result in the matrix equation

$$\frac{1}{4\Delta x^2} P_N \Psi + \frac{1}{4\Delta y^2} \Psi P_M = C, \tag{4.11}$$

with

$$P_N = \begin{bmatrix} 2 & 0 & -1 & 0 & \dots & -1 & 0 \\ 0 & 2 & 0 & -1 & \dots & 0 & -1 \\ -1 & 0 & \ddots & & & \vdots & \vdots \\ 0 & -1 & & \ddots & & -1 & 0 \\ \vdots & \vdots & & & \ddots & 0 & -1 \\ -1 & 0 & \dots & -1 & 0 & 2 & 0 \\ 0 & -1 & \dots & 0 & -1 & 0 & 2 \end{bmatrix}. \tag{4.12}$$

We were unable to find the general expression of the eigenvectors and eigenvalues of P_N, so instead of using the method described in the previous section, we will employ the conjugate gradient method [**28**]. The conjugate gradient method searches for the solution of a matrix equation $Az = b$ along orthogonal search paths $s_k \in \mathbb{R}^{NM}$, and it finds the exact solution after at most NM iterations. As this number grows large quadratically, we will use the method as an iterative method, halting the process when $||s_k||_{\ell^2}$ is less than some $\varepsilon > 0$. Given a right-hand side b, we set $\varepsilon = \alpha||b||_{\ell^2}$ for an $\alpha > 0$, so that the allowed error in the solution is proportional to b. We chose $\alpha = 10^{-5}$ in this thesis.

The matrix-matrix equation (4.11) can be rewritten as a matrix-vector equation $Az = b$, with

$$z = \mathrm{vec}(\Psi), \quad b = \mathrm{vec}(C) \quad \text{and} \quad A = \frac{1}{4\Delta x^2} P_N \otimes I_M + \frac{1}{4\Delta y^2} I_N \otimes P_M,$$

$\otimes$ denoting the Kroenecker product and vec(Ψ) the column-first vectorized version of Ψ [**28**]. Contrary to T_N, the matrix P_N is only positive semidefinite; its kernel is spanned by the vectors

$$r_N^1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}, \quad r_N^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \in \mathbb{R}^N$$

if N is even, and by

$$r_N^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \in \mathbb{R}^N$$

if N is odd. Hence, the matrix A will be positive semidefinite with a kernel spanned by the vector(s)

$$r_N^k \otimes r_M^l.$$

As a consequence, the equation $Az = b$ does not have a solution whenever b has a nonzero component in $\ker A$. However, the equation $Az = \widehat{b}$, where

$$\widehat{b} = \mathrm{proj}_{(\ker A)^\perp}(b) \ \ (= \mathrm{proj}_{\mathrm{Im} A}(b)),$$

does have a unique solution in $(\ker A)^{\perp}$. What is more, the conjugate gradient method is well-defined and converges for this modified equation. Therefore, we propose using the solution Ψ of this modified equation.

Note that this is a sort of preconditioning of the problem: We have modified the ill-conditioned problem $Az = b$ by multiplying on both sides by the projection matrix of $(\ker A)^{\perp}$, thus obtaining $Az = \widehat{b}$.

4.2.4. Numerical experiments. In this section we test the VP schemes for the wave equation (4.3) in two numerical experiments. The main emphasis will be on how well the schemes preserve vorticity. Also considered is the runtime of the VP schemes, compared to the runtime of the underlying predictor scheme. Furthermore, we compare the accuracy in the conserved variables (ζ, m_1, m_2). In both experiments we use $c = 1$ and a CFL number of 0.45.

Numerical experiment: Periodic waves. The first experiment features a periodic boundary and a nonzero initial vorticity. The periodic boundary condition will rid us of any unwanted production of vorticity along the boundary. The initial conditions are given by

$$\zeta \equiv 0, \qquad m_1 = m_2 = \cos(\pi(x+y)) - \cos(\pi(x-y)).$$

It is readily checked that

$$\zeta(x,y,t) = \sqrt{2}\sin(\pi(x+y))\sin(\sqrt{2}\pi t),$$
$$m_1(x,y,t) = m_2(x,y,t) = \cos(\pi(x+y))\cos(\sqrt{2}\pi t) - \cos(\pi(x-y))$$

is the solution of this initial value problem. The corresponding vorticity is

$$\Omega(x,y,t) = (m_2)_x - (m_1)_y = 2\pi\sin(\pi(x-y)).$$

As expected, the expression for Ω is constant in time. The initial data is plotted in Figure 4.2(a).

We compute for $(x,y) \in [-2,2] \times [-2,2]$ up to time $t = 2$. The solution at the final time step computed with the Rusanov, VPRus, Roe and VPRoe schemes are plotted in Figure 4.2, along with the exact solution. While ζ is left untouched by the VP schemes, the vorticity field is resolved much more sharply than in the predictor schemes. This is verified in Table 4.1 and 4.2, where we show relative errors

$$\frac{\|\Omega - \Omega\text{exact}\|_{L^2}}{\|\Omega\text{exact}\|_{L^2}}$$

for Ω and ζ on a sequence of meshes. The VP schemes preserve vorticity up to machine precision, while the Rusanov and Roe schemes have errors in vorticity of the order of discretization error. The L^2 dissipation estimate of the Rusanov scheme is verified in Figure 4.1. The VPRus (and VPRoe, not shown) scheme preserves vorticity exactly, while the Roe and Rusanov dissipate it excessively.

Momentum is shown in the third column of Figure 4.2. Clearly, the projection methods have a positive effect on the accuracy, preventing too much diffusion in m. Indeed, as shown in Table 4.3, the error in momentum is about $20 - 40$ per cent lower than in the prediction solvers.

The conjugate gradient method used in the elliptic solver converges rapidly, with only 5 to 10 iterations needed to get below the error threshold. Thus, the overhead is low, and the VP schemes takes only about 1.5 times more time to run than the predictor schemes.

Numerical experiment: Expanding wave. This experiment features a smooth solution with an open (Neumann) boundary condition The initial data is given by

$$\zeta = c\exp\left(-15(x^2+y^2)\right), \qquad m_1 = m_2 = 0. \tag{4.13}$$

As the initial vorticity is zero, it should stay zero at all later times. The exact solution is plotted in Figure 4.3.

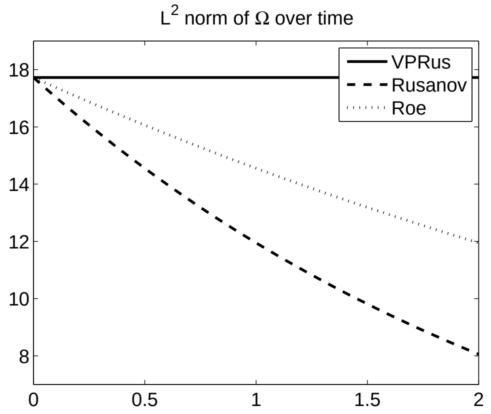


FIGURE 4.1. L^2 norm of vorticity over time on a mesh of 160×160 grid points.

	Rusanov	VPRus	Roe	VPRoe
40	$8.45 \cdot 10^{-1}$	$2.97 \cdot 10^{-14}$	$6.14 \cdot 10^{-1}$	$2.73 \cdot 10^{-14}$
80	$6.24 \cdot 10^{-1}$	$1.69 \cdot 10^{-13}$	$3.87 \cdot 10^{-1}$	$2.60 \cdot 10^{-13}$
160	$3.89 \cdot 10^{-1}$	$2.30 \cdot 10^{-13}$	$2.18 \cdot 10^{-1}$	$3.44 \cdot 10^{-14}$
320	$2.19 \cdot 10^{-1}$	$1.59 \cdot 10^{-14}$	$1.16 \cdot 10^{-1}$	$5.16 \cdot 10^{-15}$

TABLE 4.1. Relative error in Ω.

	Rusanov	VPRus	Roe	VPRoe
40	$8.43 \cdot 10^{-1}$	$8.43 \cdot 10^{-1}$	$7.36 \cdot 10^{-1}$	$7.36 \cdot 10^{-1}$
80	$6.14 \cdot 10^{-1}$	$6.14 \cdot 10^{-1}$	$5.03 \cdot 10^{-1}$	$5.03 \cdot 10^{-1}$
160	$3.84 \cdot 10^{-1}$	$3.84 \cdot 10^{-1}$	$3.02 \cdot 10^{-1}$	$3.02 \cdot 10^{-1}$
320	$2.17 \cdot 10^{-1}$	$2.17 \cdot 10^{-1}$	$1.67 \cdot 10^{-1}$	$1.67 \cdot 10^{-1}$

TABLE 4.2. Relative error in ζ.

	Rusanov	VPRus	Roe	VPRoe
40	$8.64 \cdot 10^{-1}$	$5.67 \cdot 10^{-1}$	$7.03 \cdot 10^{-1}$	$5.20 \cdot 10^{-1}$
80	$6.29 \cdot 10^{-1}$	$4.12 \cdot 10^{-1}$	$4.59 \cdot 10^{-1}$	$3.51 \cdot 10^{-1}$
160	$3.90 \cdot 10^{-1}$	$2.55 \cdot 10^{-1}$	$2.65 \cdot 10^{-1}$	$2.07 \cdot 10^{-1}$
320	$2.19 \cdot 10^{-1}$	$1.43 \cdot 10^{-1}$	$1.43 \cdot 10^{-1}$	$1.13 \cdot 10^{-1}$

TABLE 4.3. Relative error in m, the momentum.

We solve for $(x, y) \in [-2, 2] \times [-2, 2]$ up to time $t = 2$. The spatial domain was discretized with $N = M = 50, 100, 150$ and 200 grid points in each direction. Vorticity at the final time step is shown in Figure 4.4. The Rusanov scheme preserves the initial vorticity exactly in the middle of the domain, but there is a large amount of vorticity production along the border. This vorticity propagates into the domain at a speed of one grid cell per time step. As the ratio $\frac{\Delta t}{\Delta x}$ is kept constant, this speed is invariant with respect to grid size. This is clear in Table 4.4, where we show vorticity errors for the four schemes. The error for the Rusanov scheme is about $3 \cdot 10^{-2}$, irrespective of grid size. The VPRus scheme clears out these errors, and only noise of the order of machine precision is left (note the scaling of the figures).

The vorticity of the Roe scheme is shown in Figure 4.4(c). The figure shows that the initial constant vorticity is not at all preserved. Again, the projection method clears out these errors completely.

The projection method gives no gain in accuracy for the conserved variables in this experiment. As the runtime of the method lies between 2 and 4 times that of the predictor solver, there is little use in vorticity projection when the main interest is in an accurate solution of conserved variables. However, the success of the method in clearing out vorticity errors gives a motivation for applying it to the shallow water equations, where nonlinearity creates a close interconnection between flow in the $x-$ and $y-$directions.

	Rusanov	VPRus	Roe	VPRoe
50	$2.43 \cdot 10^{-2}$	$3.64 \cdot 10^{-16}$	$8.10 \cdot 10^{-2}$	$2.61 \cdot 10^{-16}$
100	$2.83 \cdot 10^{-2}$	$7.96 \cdot 10^{-16}$	$5.83 \cdot 10^{-2}$	$6.71 \cdot 10^{-16}$
150	$2.88 \cdot 10^{-2}$	$1.54 \cdot 10^{-15}$	$4.68 \cdot 10^{-2}$	$1.18 \cdot 10^{-15}$
200	$2.84 \cdot 10^{-2}$	$1.75 \cdot 10^{-15}$	$3.94 \cdot 10^{-2}$	$1.54 \cdot 10^{-15}$

TABLE 4.4. $\|\Omega\|_{L^1}$ in the expanding wave problem.

4.3. Vorticity projection for the shallow water system

Next we extend the projection method to the shallow water system (1.1). Instead of having constant vorticity, solutions of this system satisfy the more complex evolution equation (4.2). This complicates the method somewhat.

Given a solution U^n at time t_n, we compute a candidate solution $\widetilde{U}^{n+1}$ at time t_{n+1} with any consistent finite volume scheme. Let Ω^n and $\widetilde{\Omega}^{n+1}$ be their respective vorticity. Now, solve the equation (4.2) with Ω^n as initial data by using any standard finite volume scheme. This gives the correct vorticity Ω^{n+1} at time t_{n+1}. Let ψ be the solution of

$$-\Delta\psi = \widetilde{\Omega}^{n+1} - \Omega^{n+1}.$$

Then, letting

$$h = \widetilde{h}, \qquad m_1 = \widetilde{m}_1 - \psi_y, \qquad m_2 = \widetilde{m}_2 + \psi_x,$$

we get a solution $U^{n+1} = [h, m_1, m_2]^\top$ that satisfies the correct equation for vorticity.

To obtain Ω^{n+1}, we employ the Nessyahu-Tadmor (NT) scheme [**22, 29**]. This is a second-order version of the Lax-Friedrichs scheme that solves for $\Omega^{n+1}_{i+1/2,j+1/2}$, an approximation of the staggered cell average

$$\frac{1}{\Delta x \Delta y}\int_{J_{i,j}} \Omega(x,y,t)\ dxdy \qquad \text{for } J_{i,j} = [x_i, x_{i+1}) \times [y_j, y_{j+1}).$$

Write the vorticity flux as $f(\Omega, U) = u\Omega + dm_2 + sh_y$ and $g(\Omega, U) = v\Omega - dm_1 - sh_x$, where $d = u_x + v_y$ and $s = \frac{u^2+v^2}{2}$ (compare with (4.2)). The expression for $\Omega^{n+1}_{i+1/2,j+1/2}$ is

$$\begin{aligned}
\Omega^{n+1}_{i+1/2,j+1/2} &= \frac{1}{\Delta x \Delta y}\int_{J_{i,j}} \Omega^n(x,y)\ dxdy \\
&\quad - \frac{1}{\Delta x \Delta y}\int_{t_n}^{t_{n+1}}\int_{J_{i,j}} f(\Omega,U)_x + g(\Omega,U)_y\ dxdydt \\
&\approx \frac{1}{4}\left(\Omega^n_{i,j} + \Omega^n_{i+1,j} + \Omega^n_{i,j+1} + \Omega^n_{i+1,j+1}\right) \\
&\quad - \frac{\Delta t}{\Delta x \Delta y}\int_{J_{i,j}} f\left(\Omega^{n+1/2}, U^{n+1/2}\right)_x + g\left(\Omega^{n+1/2}, U^{n+1/2}\right)_y\ dxdy,
\end{aligned}$$

where we have used the quadrature rule for the time integral. $\Omega^{n+1/2}$ and $U^{n+1/2}$ are approximations of Ω and U at time $t_{n+1/2}$. We select $U^{n+1/2} = \frac{1}{2}(U^n + \widetilde{U}^{n+1})$ and $\Omega^{n+1/2} = \Omega^n - \frac{\Delta t}{2}(f_x^n + g_y^n)$, where f_x^n and g_y^n are the gradients of the flux at time t_n. To go from staggered values $\Omega^{n+1}_{i+1/2,j+1/2}$ to mesh values $\Omega^{n+1}_{i,j}$, we apply a piecewise linear reconstruction of $\Omega^{n+1}_{i+1/2,j+1/2}$

$$\Omega^{n+1}(x,y) = \Omega^{n+1}_{i+1/2,j+1/2} + \sigma_i(x - x_{i+1/2}) + \gamma_j(y - y_{j+1/2}) \qquad \text{for } (x,y) \in J_{i,j}$$

and average over $I_{i,j}$ to get

$$\Omega^{n+1}_{i,j} = \frac{1}{\Delta x \Delta y}\int_{I_{i,j}} \Omega^{n+1}(x,y)\; dxdy.$$

All derivatives used in this method are obtained using the *maxmod* slope limiter

$$\Omega_{i,j} = \mathrm{mm}\left(2\frac{\Omega_{i,j} - \Omega_{i-1,j}}{\Delta x},\ 2\frac{\Omega_{i+1,j} - \Omega_{i,j}}{\Delta x},\ \frac{\Omega_{i+1,j} - \Omega_{i-1,j}}{2\Delta x}\right),$$

with mm as in (2.25). This slope limiter gives a better resolution of discontinuities than the minmod limiter (2.24); see [**29**].

4.3.1. Numerical experiment: Vorticity advection.

We test the projection method on a problem where the exact solution is known. It is readily checked that

$$\begin{aligned} h(x,y,t) &= 1 - \frac{c_1^2}{4c_2 g}e^{2f} \\ u(x,y,t) &= M\cos(\alpha) + c_1(y - y_0 - Mt\sin(\alpha))e^f \\ v(x,y,t) &= M\sin(\alpha) - c_1(x - x_0 - Mt\cos(\alpha))e^f \end{aligned}$$

where

$$f = f(x,y,t) = -c_2\left((x - x_0 - Mt\cos(\alpha))^2 + (y - y_0 - Mt\sin(\alpha))^2\right),$$

gives a smooth solution $U = [h, hu, hv]^\top$ of (1.1) for any choice of constants M, c_1, c_2, α, x_0 and y_0. The solution consists of a vortex traveling at a constant velocity M in a direction specified by the angle α. We let $M = 1/2$, $g = 1$, $(c_1, c_2) = (-0.04, 0.02)$ and $(x_0, y_0) = (-20, -10)$. To test the schemes' ability of resolving flows that are not aligned with the computational grid, we let $\alpha = \frac{\pi}{6}$. We compute for $(x,y) \in [-50, 50] \times [-50, 50]$ up to time $t = 100$. The exact solution is shown in Figure 4.5.

The computed solutions are plotted in Figure 4.7. The Rusanov scheme dissipates the solution by a large amount, and the vortex is barely visible in the plot. The VPRus scheme, on the other hand, preserves both the magnitude and the symmetry of the vorticity well, and the solution resembles the exact solution closely. The Roe scheme solves the vortex advection problem poorly; the solution looks malformed and unsymmetric. The projection method corrects this to a good extent.

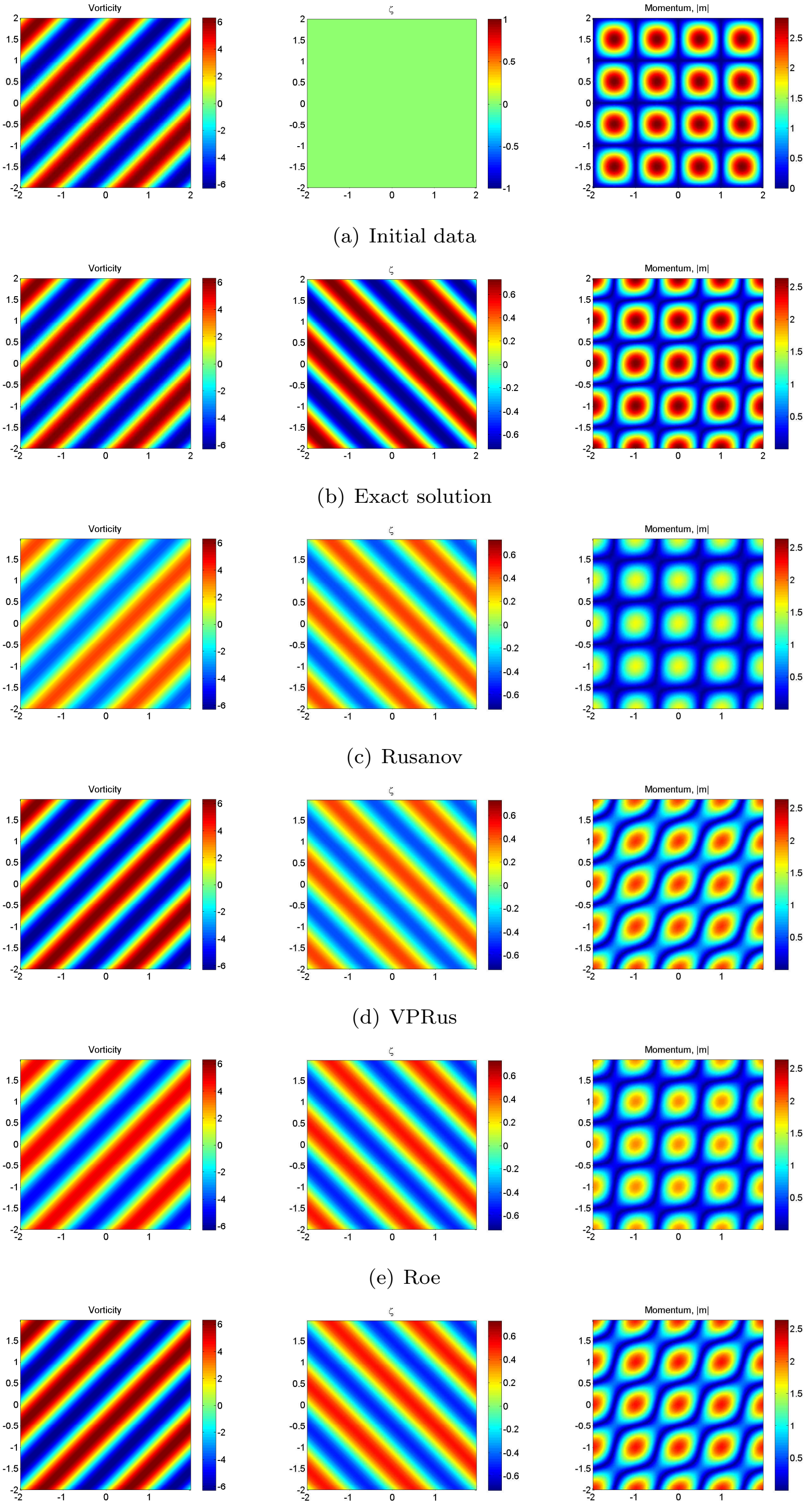
The VP schemes take about twice as long to run as the prediction solver on the same grid, and as such, a comparison of error versus grid size between the schemes would be unfair. Instead, we compute error versus runtime over a sequence of meshes, from 50 × 50 to 250 × 250 grid points. These are plotted in Figure 4.6. Clearly, the VPRus scheme gives the best error per runtime ratio of the schemes considered. The VPRoe scheme also performs well, but it seems like the non-symmetry of the Roe scheme pollutes the solution of the height variable and makes the solution look unsymmetric.

As a final remark, we note that using second-order versions of the Rusanov and Roe schemes gives similar results: Vorticity projection greatly enhances the accuracy of both schemes.

4.4. Conclusion

The method of vorticity projection was designed to reduce or eliminate errors in variables other than the conserved variables. We have demonstrated that with an efficient implementation of the method, the resolution of the conserved variables may actually be better, at the same computational cost, than when using more traditional schemes. At the same time, a significant improvement in the resolution of vorticity is achieved.

The efficiency and accuracy of the method for solving the equation for vorticity is vital to the projection method. In the case of the wave equation, this was an easy task since vorticity is constant in time. For the shallow water equations, we restricted ourselves to a second-order scheme for vorticity. Using other schemes than the NT scheme may give a more computationally efficient scheme.



(a) Initial data

(b) Exact solution

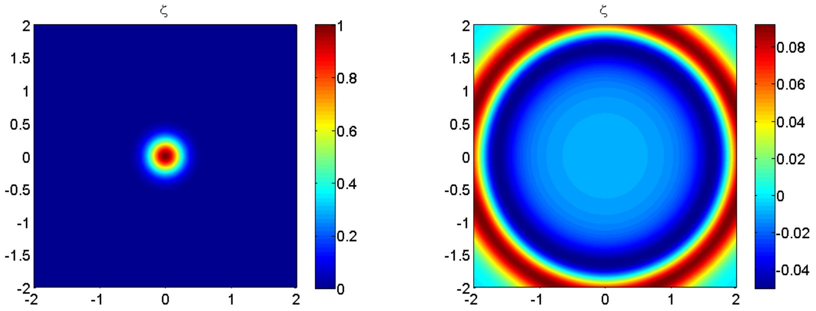
(c) Rusanov

(d) VPRus

(e) Roe

(f) VPRoe

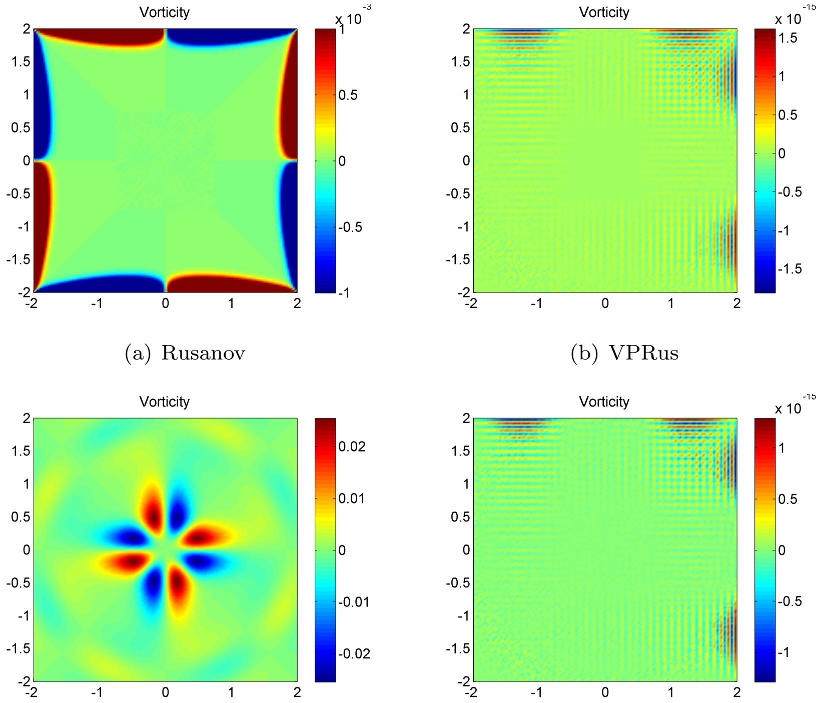
FIGURE 4.2. Solutions at $t = 2$ computed by the four schemes on a mesh of 160×160 grid points.



(a) $t = 0$

(b) $t = 2$

FIGURE 4.3. Exact solution of the expanding wave problem at the initial and final time steps.



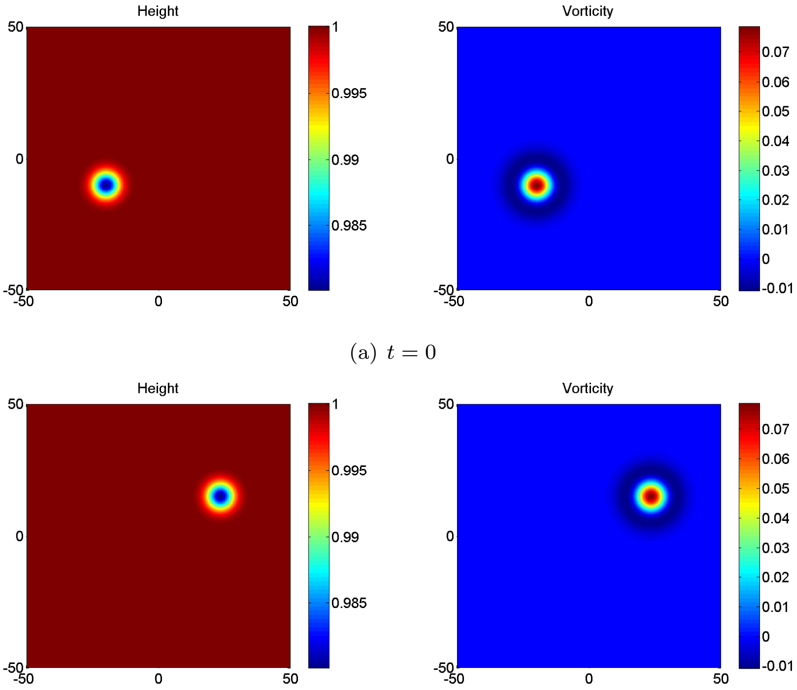
(a) Rusanov

(b) VPRus

(c) Roe

(d) VPRoe

FIGURE 4.4. Vorticity at $t = 2$ for the four schemes, computed on a mesh of 150×150 grid points. Note the scaling of each figure.



(a) $t = 0$

(b) $t = 100$

FIGURE 4.5. Exact solution of the vorticity advection problem at $t = 0$ and $t = 100$.

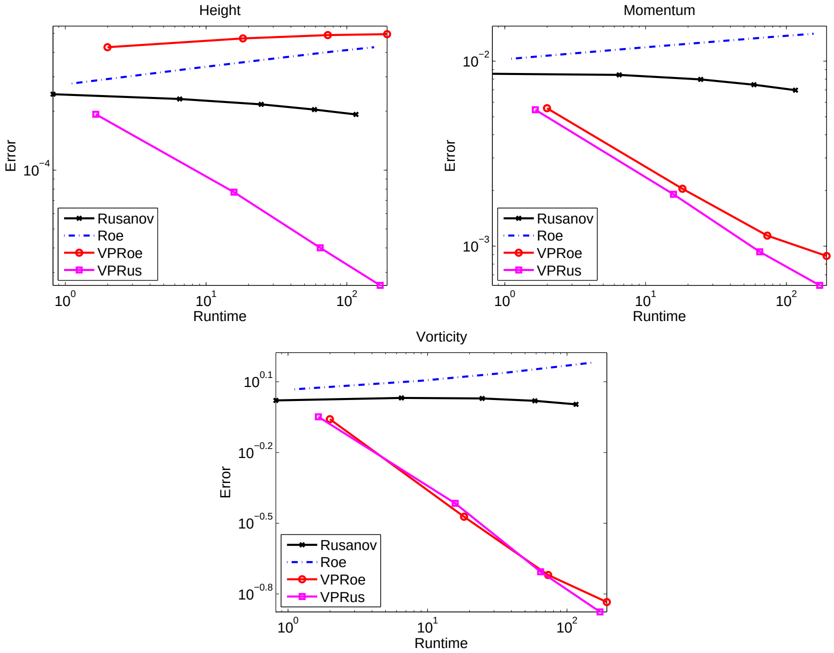
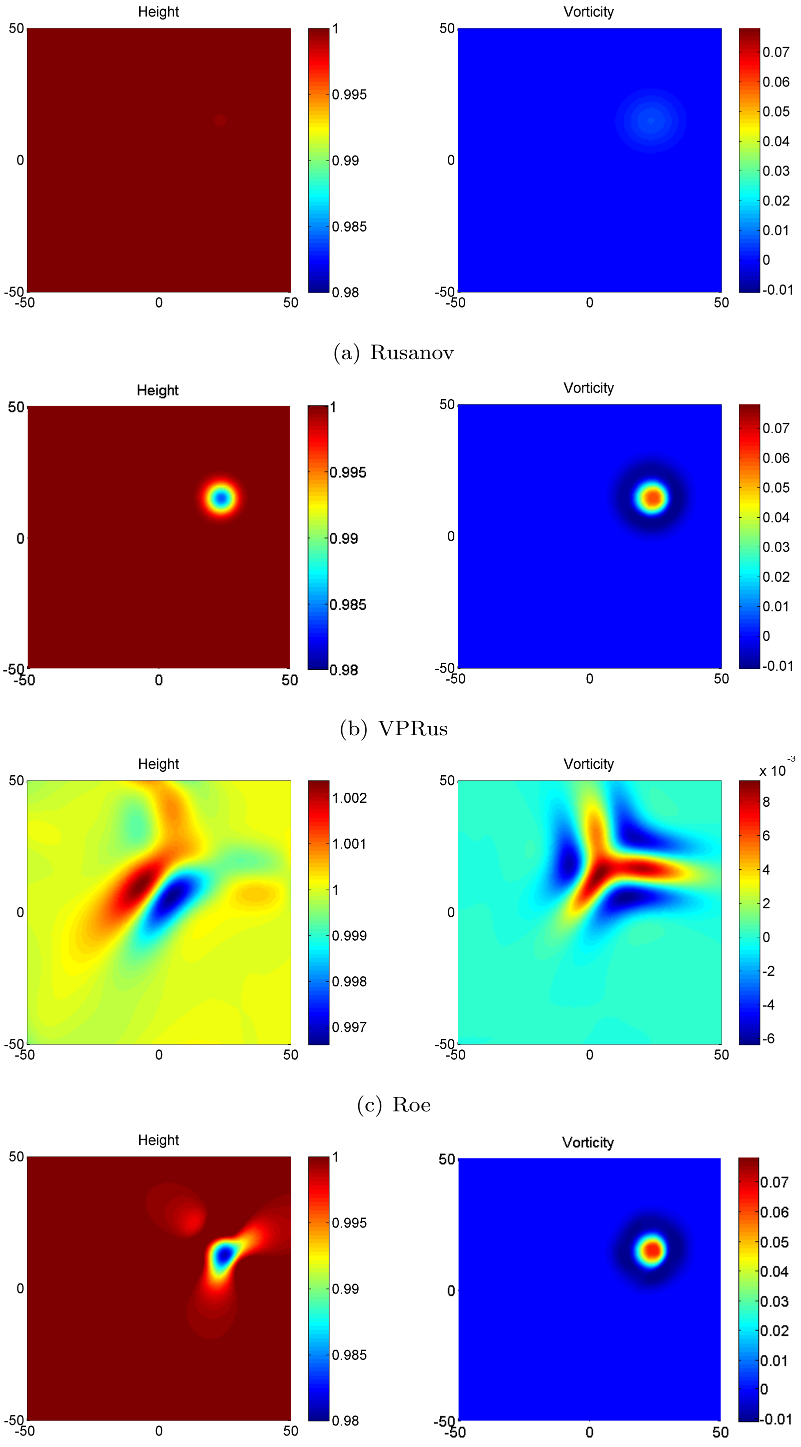


FIGURE 4.6. Runtime (x-axis) versus relative L^1 errors in height, momentum and vorticity (y-axis).



(a) Rusanov

(b) VPRus

(c) Roe

(d) VPRoe

FIGURE 4.7. Height and vorticity at $t = 100$ computed by the four schemes on a mesh of 200×200 grid points.

Bibliography

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

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

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

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

[5] T. J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. *An introduction to recent developments in theory and numerics for conservation laws*, Springer Verlag (1997).

[6] J. B. Bell, P. Colella and H. M. Glaz. A second-order projection method for the incompressible Navier-Stokes equations. *J. Comput. Phys* 85 (1989), pp. 257-283.

[7] J. U. Brackbill and D. C. Barnes. The effect of nonzero div B on the numerical solution of the magneto-hydrodynamic equations. *J. Comput. Phys.* 136 (1980), pp. 426-430.

[8] A. Chertock and A. Kurganov. On a hybrid finite-volume-particle method. *M2AN Math. Model. Num. Anal.,* 38, (2004), pp. 1971-1991.

[9] A. J. Chorin. Numerical Solution of the Navier-Stokes Equations. *Mathematics of Computation*, 22 (1968), pp. 745-762.

[10] R. J. Fennema and M. H. Chaudhry. Explicit methods for the 2d-transient free surface flows. *J. Hydraul. Eng. ASCE.,* 116 (1990), pp. 1013-1034.

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

[12] U. Fjordholm, S. Mishra and E. Tadmor. Well-balanced energy preserving and energy stable schemes for the shallow water equations with topography. *In preparation.*

[13] U. Fjordholm and S. Mishra. Vorticity preservation for the shallow water equations. *In preparation.*

[14] E. Godlewski and P.-A. Raviart. Hyperbolic systems of conservation laws. *Mathematiques & Applications*, no. 3/4, Ellipses, Paris (1991).

[15] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equiations of hydrodynamics. *Math. Sb.* 47 (1959), pp. 271-306.

[16] L. Gosse and A.-Y. LeRoux. A well-balanced scheme designed for inhomogeneous scalar conservation laws. *C. R. Acad. Sci Paris Sr I. Math.*, 323 (1996), pp. 543-546.

[17] S. Gottlieb, C. W. Shu and E. Tadmor. High order time discretizations with strong stability properties. *SIAM. Review,* 43 (2001), pp. 89 - 112.

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

[19] H. Holden and N. H. Risebro. Front tracking for hyperbolic conservation laws. *Springer Verlag*, Berlin, 2nd. printing (2007).

[20] F. Ismail and P. L. Roe. Towards a vorticity preserving second order finite volume scheme solving the Euler equations. 17th AIAA Computational Fluid Dynamics Conference

[21] R. Jeltsch and M. Torrilhon. On curl-preserving finite volume discretizations for shallow water equations. *Bit Numerical Mathematics* 46(1) (2006), pp. 35-53)

[22] Jiang,, G.-S. and Levy,, D. and Lin,, C.-T. and Osher,, S. and Tadmor,, E. High-Resolution Non-oscillatory Central Schemes with Non-staggered Grids for Hyperbolic Conservation Laws *SIAM J. Numer. Anal.*, 35 (1998), pp. 2147-2168.

[23] S. N. Kruzkov. First order quasi-linear equations in several independent variables. *Math. USSR Sb.*, 10 (1970), pp. 217-243.

[24] Peter D. Lax. Hyperbolic systems of conservation laws and the mathematical theory of shock waves. *Philadelphia: Society for Industrial and Applied Mathematics* (1973).

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

[26] Randall J. LeVeque. Finite volume methods for hyperbolic problems. *Cambridge University Press,* Cambridge (2002).

[27] D. Levy and E. Tadmor. Non-oscillatory central schemes for the incompressible 2-D Euler equations. *Mathematical Research Letters* 4(3) (1997), pp. 321-340.

[28] T. Lyche. Lecture Notes for Inf-Mat 4350. Available for download at http://www.uio.no/studier/emner/matnat/ifi/INF-MAT4350/h08/undervisningsmateriale/book2008.pdf.

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

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

[31] S. Osher. Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.* 21 (1984), pp. 217-235

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

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

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

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

[36] E. Tadmor. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numerica* 12 (2003), pp. 451-512.

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

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

[39] G. Toth. The div $B = 0$ constraint in shock capturing magneto-hydrodynamic codes. *J. Comp. Phys.* 161 (2000), pp. 605-652.