Well-Balanced Schemes for Shallow Water Equations

Daniel Cortild

November 3rd, 2023

Abstract

The shallow water equations constitute a fundamental framework for modeling water flow over various surfaces, including rivers, lakes, and reservoirs, with extensive applications in oceanography. This report presents a meticulously designed well-balanced numerical scheme for solving the one-dimensional Saint-Venant equations. The approach begins with a finite-volume discretization of the spatial term, followed by the derivation of a scheme that accurately maintains the "lake at rest" steady-state condition, and subsequently extended to accommodate the dynamic state of "moving water". The obtained semi-discrete ODE is then solved via a strong stability preserving method. The numerical solutions are then analyzed, revealing notable observations, such as the first method's inability to uphold the latter steady-state.

Keywords Shallow Water Equations, Finite Volume Methods, Well-Balanced Schemes, Strong Stability Preserving Methods

1 Introduction

Dealing with the flow of water over a surface, such as in a river, a canal, a lake or other similar situations, often gives rise to the shallow water equations. These equations are widely used in oceanography and atmospheric sciences to model natural phenomena such as tsunamis or hurricanes. They are also applicable and used in coastal or hydraulic engineering purposes to design harbour areas, develop urban coastal areas, construct coastal protection systems, etc.

The classical shallow water equations are given by the Saint-Venant system [3]. In the one-dimensional case, whilst neglecting viscosity and bottom friction, the equations are given by

$$h_t + (hu)_x = 0, \tag{1a}$$

$$(hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x = -ghB_x, \tag{1b}$$

where $x \in \mathbb{R}$ is the spatial coordinate, $t \in \mathbb{R}_{\geq 0}$ is the time coordinate, $h = h(x,t) \in \mathbb{R}$ is the depth of the water, $u = u(x,t) \in \mathbb{R}$ is the velocity of the water (assumed constant over the entire vertical layer), $B = B(x) \in \mathbb{R}$ is the predefined time-independent bottom topography, and $g \in \mathbb{R}_{\geq 0}$ is the constant gravitational acceleration.

The system (1) is a nonlinear hyperbolic system, which admits complicated, generally non-smooth, solutions that may contain shock and rarefaction waves. It may even, in the case of a discontinuous bottom topography, contain contact discontinuities. Unless the initial data is very simplistic [1], no analytical solution to the system has been found, and one thus has to ressort to numerical solutions. Even in the case of a flat bottom topography and a smooth initial condition, the solutions may break down [7], meaning that the design of a numerical solution may be a non-trivial task.

Another challenge in the development of a numerical solution to system (1) is that the system represents a set of balance laws, and that a good numerical solution should respect this balance between the flux terms and the source terms. We thus want our scheme to be able to preserve initial data that corresponds to steady-state solutions. Such schemes are called *well-balanced*. Their advantage over non-balanced schemes becomes apparent when simulating steady-states or near-steady-states, which is often the case in practical applications.

This report is divided into multiple sections. In Section 2, we consider a finite-volume scheme of system (1) to obtain a semi-discrete ODE. We design a well-balanced scheme in Section 3 by discretizing the source term appropriately. In Section 4 we briefly explain the concept of a strong stability preserving method, and write out the method that we shall use. Numerical results of this scheme are presented in Section 5. Finally, a conclusion of the work is laid out in Section 6.

2 Semi-Discretization in Space

A general system of balance laws may be represented as

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = \mathbf{S}(\mathbf{U}), \tag{2}$$

where $\mathbf{U} = \mathbf{U}(x,t)$ is a vector of unknowns, $\mathbf{F}$ is a flux function, and $\mathbf{S}$ is a source term. The system (2) is *hyperbolic* if the Jacobian $\partial\mathbf{F}/\partial\mathbf{U}$ is diagonalizable with real distinct eigenvalues.

We note that the Saint-Venant system (1) is of the form of system (2) by writing $\mathbf{U} = (h,q)^T$ with $q \coloneqq hu$ representing the discharge, and

$$\mathbf{F}(h,q) = \left(q, \frac{q^2}{h} + \frac{gh^2}{2}\right)^T, \quad \text{and} \quad \mathbf{S}(h,q) = (0, -ghB_x)^T. \tag{3}$$

It is hyperbolic since the Jacobian of $\mathbf{F}$ is given by

$$\frac{\partial\mathbf{F}}{\partial\mathbf{U}}(h,q) = \begin{pmatrix} 0 & 1 \\ -q^2/h^2 + gh & 2q/h \end{pmatrix},$$

which has real eigenvalues $\lambda_{1,2} = q/h \pm \sqrt{gh}$.

2.1 Finite-Volume Scheme

We now build a finite-volume scheme for the hyperbolic system of balance laws (2). We consider a uniform mesh with grid points x_j and control volumes $C_j = [x_{j-1/2}, x_{j+1/2}]$ of size Δx, centered at $x_j = \frac{1}{2}(x_{j-1/2} + x_{j+1/2})$. We approximate the cell average of $\mathbf{U}$ over cell C_j by $\overline{\mathbf{U}}_j$, and the average of $\mathbf{S}$ over C_j by $\overline{\mathbf{S}}$, namely we have that

$$\overline{\mathbf{U}}_j(t) \approx \frac{1}{\Delta x}\int_{C_j} \mathbf{U}(x,t)dx \quad \text{and} \quad \overline{\mathbf{S}}_j(t) \approx \frac{1}{\Delta x}\int_{C_j} \mathbf{S}(\mathbf{U}(x,t))dx. \tag{4}$$

We shall aim to solve for $\overline{\mathbf{U}}$, which, under grid refinement, provides a suitable solution for $\mathbf{U}$. We denote the components of $\overline{\mathbf{S}}_j(t)$ by $\overline{\mathbf{S}}_j^{(i)}(t)$.

We construct our semi-discretization by integrating the system (2) over C_j and dividing by Δx to obtain

$$\frac{d}{dt}\overline{\mathbf{U}}_j(t) = -\frac{1}{\Delta x}\left[\boldsymbol{\mathcal{F}}_{j+1/2}(t) - \boldsymbol{\mathcal{F}}_{j-1/2}(t)\right] + \frac{1}{\Delta x}\int_{C_j} \mathbf{S}(\mathbf{U}(x,t))dx,$$

where the numerical flux $\boldsymbol{\mathcal{F}}$ is given by the Lax-Friedrichs flux, namely,

$$\boldsymbol{\mathcal{F}}_{j+1/2}(t) = \frac{1}{2}\left[\mathbf{F}\left(\mathbf{U}^+_{j+1/2}(t)\right) + \mathbf{F}\left(\mathbf{U}^-_{j+1/2}(t)\right) - \frac{\Delta x}{\Delta t}\cdot\left(\mathbf{U}^+_{j+1/2}(t) - \mathbf{U}^-_{j+1/2}(t)\right)\right], \tag{5}$$

where $\mathbf{U}^\pm_{j+1/2}$ are the reconstructions of the right and left values at the interfaces $x_{j+1/2}$. This reconstruction shall be further elaborated on in Section 2.2.

By further applying the approximation of $\mathbf{S}$ given in Equation (4), we obtain the semi-discrete system

$$\frac{d}{dt}\overline{\mathbf{U}}_j = -\frac{1}{\Delta x}\left[\boldsymbol{\mathcal{F}}_{j+1/2} - \boldsymbol{\mathcal{F}}_{j-1/2}\right] + \overline{\mathbf{S}}_j, \tag{6}$$

where the dependencies on time are omitted for readability. Once again, the components of $\boldsymbol{\mathcal{F}}_{j+1/2}$ are denoted by $\boldsymbol{\mathcal{F}}^{(i)}_{j+1/2}$.

The semi-discrete system may be implemented by numerically solving the ODE with an appropriate solver. Methods like explicit or implicit Euler are applicable, but methods of choice generally result to strong stability preserving Runge-Kutta or multi-stage methods [4]. This will be further explained in Section 4.

2.2 Reconstruction of Variables

The numerical flux in Equation (5) requires a reconstruction of $\mathbf{U}_j$ at the cell interfaces, denoted by $\mathbf{U}^{\pm}_{j\pm1/2}$. In the methods shown in Section 3, we shall explicitely reconstruct the variables, depending on the method at hand. In this section we briefly present how to reconstruct variables $k^{\pm}_{j\pm1/2}$ from a sequence of variables $(\overline{k}_j)$, using linear numerical derivatives.

The recontructed variables shall be given by

$$k^{-}_{j+1/2} = \overline{k}_j + \frac{\Delta x}{2}\left(\overline{k}_j\right)_x \quad \text{and} \quad k^{+}_{j+1/2} = \overline{k}_{j+1} - \frac{\Delta x}{2}\left(\overline{k}_{j+1}\right)_x,$$

which is a linear extrapolation. The numerical derivative $(\overline{k}_j)_x$ is defined as

$$\left(\overline{k}_j\right)_x = \text{minmod}\left(\theta \cdot \frac{\overline{k}_j - \overline{k}_{j-1}}{\Delta x}, \frac{\overline{k}_{j+1} - \overline{k}_{j-1}}{2}, \theta \cdot \frac{\overline{k}_{j+1} - \overline{k}_j}{\Delta x}\right),$$

in which $\theta \in [1, 2]$ is a parameter that lets us tune how much importance one-sided derivatives have. For the rest of this report, we select $\theta = 1$. We recall for complenetess that

$$\text{minmod}(z_1, \ldots, z_n) = \begin{cases} \min(z_1, \ldots, z_n) & \text{if } z_1, \ldots, z_n \geq 0 \\ \max(z_1, \ldots, z_n) & \text{if } z_1, \ldots, z_n \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We note that reconstructing the components of $\mathbf{U}$ would lead to a reconstruction $\mathbf{U}^{\pm}_{j+1/2}$. This is however not always desirable, as will be shown in Section 3.

3 Well-Balanced Scheme

To derive a numerical scheme for the semi-discrete system (6) applied to the Saint-Venant system (3), we must select a reconstruction of the variable $\mathbf{U}^{\pm}_{j\pm1/2}$ and an appropriate quadrature for the integral $\overline{\mathbf{S}}_j$ in Equation (4). Although this seems like an easy task, the simple choice, being to select $\mathbf{U}^{\pm}_{j\pm1/2} = (h^{\pm}_{j\pm1/2}, q^{\pm}_{j\pm1/2})$ and a standard quadrature, such as the midpoint rule, which would result in

$$\overline{\mathbf{S}}^{(2)}_j \approx -g\overline{h_j}B_x(x_j), \tag{7}$$

leads to a non-well-balanced scheme, incapable of preserving trivial steady-state solutions [5]. This will be shown in Section 5, by means of a numerical example.

A good scheme should be able to preserve steady-state solutions of the system. We call such a scheme a *well-balanced* scheme. These steady-states may be computed as

$$q \equiv \hat{q} = \text{constant}, \quad \text{and} \quad E \coloneqq \frac{u^2}{2} + g(h + B) \equiv \hat{E} = \text{constant}. \tag{8}$$

Two types of equilibria arise from the above equations. The first type is attained at $\hat{q} = 0$, in which case we consider the water surface at rest (with a zero velocity). This will be considered in Section 3.2. If $\hat{q}$ is not necessarily assumed zero, we achieve a new family of equilibria in which the surface might not be at rest. These will be considered in Section 3.3. Both methods require a reconstruction of the bottom topography, that is discussed first in Section 3.1.

3.1 Reconstruction of the Bottom Topography

We first replace the original bottom topography by a continuous piecewise linear approximation of it, which coincides with the original topography at the points $x_{j+1/2}$. More precisely, we call $\tilde{B}$ the approximated bottom topography, which is defined as, assuming $x \in C_j$

$$\tilde{B}(x) = B_{j-1/2} + (B_{j+1/2} - B_{j-1/2}) \cdot \frac{x - x_{j-1/2}}{\Delta x},$$

where, for $B^{\pm}(x_{j+1/2})$ representing the values at the right and left of the interface at $x_{j+1/2}$,

$$B_{j+1/2} \coloneqq \frac{B^{+}(x_{j+1/2}) + B^{-}(x_{j+1/2})}{2}.$$

Note that when B is continuous at $x_{j+1/2}$, the above reduces to $B_{j+1/2} = B(x_{j+1/2})$. The linearization is shown in Figure 1. It is important to realise that the definition is consistent at the cell interfaces, and that we can assure the continuity of the reconstruction, even in case of a discontinuous bottom topogaphy.

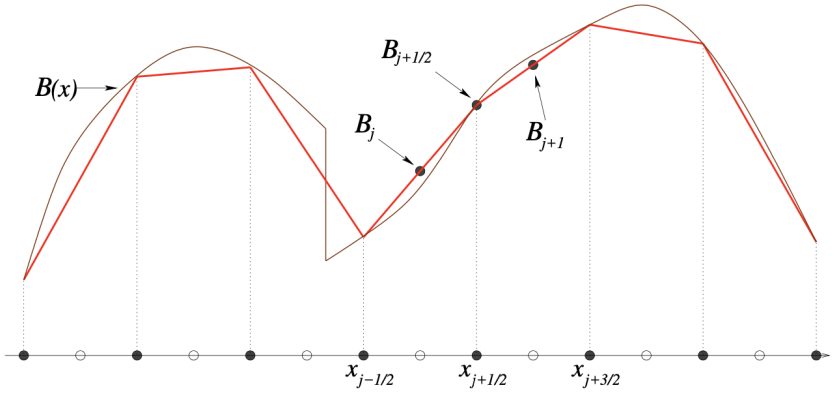

Figure 1: Bottom topography B and its linear approximation $\tilde{B}$. Figure from [6].

Since $\tilde{B}$ is linear over a cell C_j, we note that $\tilde{B}(x_j)$ coincides with its cell average over C_j and the average of $\tilde{B}$ evaluated at the endpoints of C_j, namely

$$B_j \coloneqq \tilde{B}(x_j) = \frac{1}{\Delta x}\int_{C_j} \tilde{B}(x)dx = \frac{B_{j-1/2} + B_{j+1/2}}{2}.$$

However, contrary to what notation suggests, it does not necessarily hold that $B_j = B(x_j)$.

3.2 Preserving Still-Water Equilibrium

As a special case of the steady-states presented in Equation (8), when $u \equiv 0$, and thus $q \equiv 0$, we observe the "lake at rest" steady-state, given by

$$q \equiv 0, \quad \text{and} \quad w \coloneqq h + B \equiv \hat{w} = \text{constant}, \tag{9}$$

which corresponds to still water and a flat water surface, although the bottom topography may not be flat.

We shall design a scheme which is capable of preserving this equilibrium. In order to do so, we follow the ideas proposed in [6].

We recall that we have $\overline{\mathbf{U}}_j = (\overline{h}_j, \overline{q}_j)$ available. We define $\overline{w}_j = B_j + \overline{h}_j$, and reconstruct $w^{\pm}_{j\pm1/2}$ and $q^{\pm}_{j\pm1/2}$, as in Section 2.2. We then define $h^{\pm}_{j\pm1/2} = w^{\pm}_{j\pm1/2} - B_{j\pm1/2}$, thus also defining $\mathbf{U}^{\pm}_{j\pm1/2}$. The motivation behind reconstructing w and q rather than h and q is that the variables expected to be constant in the equilibrium are exactly w and q. Indeed, if h were to be reconstructed instead of w, since $\overline{h}_j$ is not constant over all cells in an equilibrium situation, it may happen that $h^{+}_{j\pm1/2} \neq h^{-}_{j\pm1/2}$, and hence that $w^{+}_{j+1/2} \neq w^{-}_{j-1/2}$, which is undesirable. If however w is reconstructed, we know that $w^{-}_{j+1/2} = w^{+}_{j+1/2}$, since $\overline{w}_j$ is constant over all cells, hence producing a much more attractive result.

Note that, at the "lake at rest" equilibrium, given by Equations (9), we have $q \equiv 0$ and $h = \hat{w} - B$. In a discrete setting, we have that $\overline{q}_j \equiv 0$ and $\overline{w}_j \equiv \hat{w}$, which implies that

$$q^{\pm}_{j+1/2} = 0 \quad \text{and} \quad w^{\pm}_{j+1/2} = \hat{w}.$$

In specific, in the equilibrium state, it holds that $h^+_{j+1/2} = h^-_{j+1/2}$, and thus that

$$\mathbf{F}(\mathbf{U}^\pm_{j+1/2}) = \left(0, \frac{g}{2} \cdot (\hat{w} - B_{j+1/2})^2\right)^T.$$

Thus the numerical flux (5) is given by

$$\boldsymbol{\mathcal{F}}^{(1)}_{j+1/2} = 0 \quad \text{and} \quad \boldsymbol{\mathcal{F}}^{(2)}_{j+1/2} = \frac{g}{2} \cdot (\hat{w} - B_{j+1/2})^2.$$

The semi-discrete system (6) then reduces to

$$\frac{d}{dt}\overline{h}_j = 0 \quad \text{and} \quad \frac{d}{dt}\overline{q}_j = g \cdot (\hat{w} - B_j) \cdot \frac{B_{j+1/2} - B_{j-1/2}}{\Delta x} + \overline{\mathbf{S}}^{(2)}_j.$$

It is clear that the right-hand side will vanish in the equilibrium state if we evaluate the source term using

$$\overline{\mathbf{S}}^{(2)}_j \approx -\frac{1}{\Delta x}\int_{C_j} ghB_x dx \approx -g \cdot \overline{h}_j \cdot \frac{B_{j+1/2} - B_{j-1/2}}{\Delta x}. \tag{10}$$

We note that the above is similar to the midpoint quadrature in Equation (7), by replacing B_x by its finite difference approximation.

3.3 Preserving Moving-Water Equilibrium

The preservation of the moving-water steady-state is more complicated than the still-water equilibrium, and has only been achieved recently. We follow the ideas laid out by Cheng and Kurganov [2].

We recall that we compute $\overline{\mathbf{U}}_j = \left(\overline{h}_j, \overline{q}_j\right)^T$ at each time level, meaning we have access to $\overline{h}_j$ and $\overline{q}_j$. We define

$$u_j = \frac{\overline{q}_j}{\overline{h}_j} \quad \text{and} \quad E_j = \frac{u_j^2}{2} + g \cdot \left(\overline{h}_j + B_j\right).$$

We now reconstruct the values $E^\pm_{j\pm1/2}$ and $q^\pm_{j\pm1/2}$ according to Section 2.2. The motivation behind reconstructing E and q rather than h and q is similar to the one in the previous case.

3.3.1 Reconstruction of Water Height

We now reconstruct the values of $h^\pm_{j+1/2}$ as the solution of the cubic

$$\phi(h) = \frac{\left(q^\pm_{j+1/2}\right)^2}{2h^2} + g \cdot (h + B_{j+1/2}) - E^\pm_{j+1/2} = 0,$$

which then also yields the reconstructed $\mathbf{U}^\pm_{j\pm1/2}$. The dependency of ϕ on j and on $\pm$ is understood but omitted for readability. This cubic has only one relevant solution, which can be obtained by numerically solving the equation via Newton's method with an appropriate initial guess. The importance of the initial guess arises from Newton's method only converging locally. On the other hand, this allows us to predict the root it will converge to. For each point, we approximate the Froude number $\mathrm{Fr} = |u|/\sqrt{gh}$ using the known values, and select an initial value such that the algorithm converges to the supersonic or the subsonic solution, depending on the sign of $\mathrm{Fr} - 1$.

Firstly we notice that if $q^\pm_{j+1/2} = 0$, the equation $\phi(h) = 0$ is trivially solved by

$$h^\pm_{j+1/2} = \frac{E^\pm_{j+1/2}}{g} - B_{j+1/2}.$$

We now concern ourselves with the case $q^\pm_{j+1/2} \neq 0$. Before describing the initial guess, we point out that ϕ has a negative root, and hence at most 2 positive roots. Of course, the negative root is not relevant, as we would like

to enforce positive water heights. Since we want the simulation to be physically relevant, we assume the equation has at least one positive root. Over $\mathbb{R}_{>0}$, ϕ is convex and attains its minimum at

$$h_0 = \sqrt[3]{\left(q^{\pm}_{j+1/2}\right)^2/g} > 0,$$

which corresponds to the sonic point, namely the point where $\mathrm{Fr} = 1$. Once again, the depence of h_0 on j and on $\pm$ is omitted for readability, and will be made clear when needed. As such, whenever ϕ possesses distinct positive roots, it must possess one in the interval $(0, h_0)$, corresponding to the supersonic point (where $\mathrm{Fr} \geq 1$) and one in the interval $(h_0, +\infty)$, corresponding to the subsonic point (where $\mathrm{Fr} \leq 1$). Since we assume ϕ to have a root over $\mathbb{R}_{>0}$, we know that $\phi(h_0) \leq 0$. We note that if $\mathrm{Fr} = 1$, the unique solution is $h = h_0$.

To ensure the convergence of Newton's method towards the appropriate root when $\mathrm{Fr} \neq 1$, we shall select an initial guess $h_\star$ in the appropriate interval, and make sure $\phi(h_\star) > \varepsilon$ for some $\varepsilon > 0$, which is required for convergence of Newton's method. Knowing that $\phi(h_0) < 0$, we shall start at $h_\star = h_0$, and multiply this initial guess by a constant, either larger or smaller than 1 depending on the approximative Froude number, until we reach an initial guess satisfying $\phi(h_\star) > \varepsilon$.

Formally, we define the approximative Froude numbers using a cell approximation of q and an upwind approximation of h

$$\mathrm{Fr}^-_{j+1/2} = \frac{\left|\overline{q}_j\right|}{\sqrt{g \cdot \overline{h}_j}} \quad \text{and} \quad \mathrm{Fr}^+_{j+1/2} = \frac{\left|\overline{q}_{j+1}\right|}{\sqrt{g \cdot \overline{h}_j}}.$$

and note that if $\mathrm{Fr}^{\pm}_{j+1/2} = 1$, we can assume we can solve the equation $\phi(h) = 0$ anaytically, and set $h^{\pm}_{j+1/2} = h^{\pm}_0$. Otherwise we set

$$h^{\pm}_* = \begin{cases} \lambda_1^m \cdot h^{\pm}_0 & \text{if } \mathrm{Fr}^{\pm}_{j+1/2} > 1, \\ \lambda_2^m \cdot h^{\pm}_0 & \text{if } \mathrm{Fr}^{\pm}_{j+1/2} < 1. \end{cases}$$

The value m is selected as the smallest nonnegative integer such that $\phi(h_*) > \varepsilon$. The parameters λ_1, λ_2 must satisfy $0 < \lambda_1 < 1 < \lambda_2$. For our experiments we select $\varepsilon = 10^{-4}$, $\lambda_1 = 0.9$ and $\lambda_2 = 1.1$. Solving the equation $\phi(h) = 0$ using Newton's method with initial guess h_* will then converge to the appropriate solution $h^{\pm}_{j+1/2}$.

We note that for Froude numbers close to 1 but not exactly 1, Newton's method might not converge. In that case we also set $h^{\pm}_{j+1/2} = h^{\pm}_0$.

3.3.2 Discretization of the Source Term

In the general moving-water equilibrium given by Equation (8), we have $E_j \equiv \hat{E}$ and $\overline{q}_j \equiv \hat{q}$. In specific, the initial guesses for computing $h^{\pm}_{j+1/2}$, as explained in the previous subsection, are identical, which implies that $h_{j+1/2} \coloneqq h^+_{j+1/2} = h^-_{j+1/2}$. The numerical flux (5) is now given by

$$\boldsymbol{\mathcal{F}}^{(1)}_{j+1/2} = \hat{q} \quad \text{and} \quad \boldsymbol{\mathcal{F}}^{(2)}_{j+1/2} = \frac{\hat{q}^2}{h_{j+1/2}} + \frac{g \cdot h^2_{j+1/2}}{2}.$$

We may thus write the semi-discrete system (6) as

$$\frac{d}{dt}\overline{h}_j = 0 \quad \text{and} \quad \frac{d}{dt}\overline{q}_j = -\frac{g}{2\Delta x}\left(h^2_{j+1/2} - h^2_{j-1/2}\right) - \frac{\hat{q}^2}{\Delta x}\left(\frac{1}{h_{j+1/2}} - \frac{1}{h_{j-1/2}}\right) + \overline{\mathbf{S}}^{(2)}_j. \tag{11}$$

We would like to remove the dependency on $\hat{q}$ in the above equation. This may be done by noting that $E_{j-1/2} = \hat{E} = E_{j+1/2}$, namely that

$$0 = E_{j+1/2} - E_{j-1/2} = \frac{\hat{q}^2}{2}\left(\frac{1}{h^2_{j+1/2}} - \frac{1}{h^2_{j-1/2}}\right) + g \cdot (h_{j+1/2} - h_{j-1/2}) + g \cdot (B_{j+1/2} - B_{j-1/2}).$$

Replacing the factor $g \cdot (h_{j+1/2} - h_{j-1/2})$ in Equation (11) by its value given above yields

$$\frac{d}{dt}\overline{q}_j = \frac{g}{2\Delta x}\left(B_{j+1/2} - B_{j-1/2}\right)\left(h_{j+1/2} + h_{j-1/2}\right) + \frac{\hat{q}^2}{4\Delta x}\left(\frac{1}{h_{j+1/2}^2} - \frac{1}{h_{j-1/2}^2}\right)\left(h_{j+1/2} + h_{j-1/2}\right) - \frac{\hat{q}^2}{\Delta x}\left(\frac{1}{h_{j+1/2}} - \frac{1}{h_{j-1/2}}\right) + \overline{\mathbf{S}}_j^{(2)}.$$

Looking at the last two terms not including $\overline{\mathbf{S}}_j^{(2)}$, we can rewrite them as

$$\begin{aligned}
&\frac{\hat{q}^2}{4\Delta x}\left(\frac{1}{h_{j+1/2}^2} - \frac{1}{h_{j-1/2}^2}\right)\left(h_{j+1/2} + h_{j-1/2}\right) - \frac{\hat{q}^2}{\Delta x}\left(\frac{1}{h_{j+1/2}} - \frac{1}{h_{j-1/2}}\right) \\
&\quad = -\frac{\hat{q}^2}{4\Delta x}\frac{h_{j+1/2} - h_{j-1/2}}{h_{j+1/2}\cdot h_{j-1/2}} \cdot \left(\frac{(h_{j+1/2} + h_{j-1/2})^2}{h_{j+1/2}\cdot h_{j-1/2}} - 4\right) \\
&\quad = -\frac{\hat{q}^2}{4\Delta x} \cdot \left(\frac{1}{h_{j+1/2}} - \frac{1}{h_{j-1/2}}\right)^2 \cdot (h_{j+1/2} - h_{j-1/2}),
\end{aligned}$$

which leads to the following semi-discretization

$$\frac{d}{dt}\overline{q}_j = \frac{g}{2\Delta x}\left(B_{j+1/2} - B_{j-1/2}\right)\left(h_{j+1/2} + h_{j-1/2}\right) - \frac{\hat{q}^2}{4\Delta x} \cdot \left(\frac{1}{h_{j+1/2}} - \frac{1}{h_{j-1/2}}\right)^2 \cdot (h_{j+1/2} - h_{j-1/2}) + \overline{\mathbf{S}}_j^{(2)}.$$

It is clear that if we use the quadrature

$$-\frac{1}{\Delta x}\int_{C_j} ghB_x dx \approx \overline{\mathbf{S}}_j^{(2)} \approx -g\cdot\frac{h_{j+1/2}^- + h_{j-1/2}^+}{2}\cdot\frac{B_{j+1/2} - B_{j-1/2}}{\Delta x} + \frac{\left(u_{j+1/2}^- - u_{j-1/2}^+\right)^2}{4\Delta x}\cdot(h_{j+1/2}^- - h_{j-1/2}^+), \quad (12)$$

the right-hand side vanishes in the moving-water steady-state, thus achieving a moving-water equilibrium preserving scheme.

4 Strong Stabiliy Preserving Schemes

This section focuses on strong stability preserving schemes. Without referring explicitely to it, the content is inspired by the work of Gottlieb [4].

Although we now have suitable discretizations for the source term that generate semi-discrete ODEs in the form of Equation (6), this ODE still needs to be solved. In general, in order to avoid nonlinear solves, we would like to apply an explicit method. These however often compromise on stability. We will develop a method that mimics the simple stability criterion of the first-order forward Euler method, but achieves a higher order accuracy.

We consider a generic time-dependent ODE given by

$$u_t = f(u), \quad (13)$$

where $u = u(t)$ is a vector of unknowns and f models the system.

The simplest example of a explicit method is given by the forward Euler method. Applied to Equation (13), it reads

$$u^{n+1} = u^n + \Delta t f(u^n), \quad (14)$$

where u^n is a numerical approximation of $u(t^n)$.

We will make use of a *strong stability preserving* (SSP) method. We assume that the forward Euler method is strongly stable, namely that for all $0 \leq \Delta t \leq \Delta t_{FE}$ and u, it holds that

$$\|u + \Delta t F(u)\| \leq \|u\|. \quad (15)$$

Under that assumption, a method is called SSP if, for u^n the approximation of $u(t^n)$ obtained through the method, it holds that

$$\|u^{n+1}\| \leq \|u^n\|,$$

provided that $\Delta t \leq C \cdot \Delta t_{FE}$. The constant C is called the SSP-coefficient. We note that $\|\cdot\|$ is assumed a norm, semi-norm or convex functional.

We will consider explicit SSP Runge-Kutta methods. In Section 4.1, we prove a general result about SSP methods. In Section 4.2, we develop and prove the optimality of a specific second-order SSP scheme, namely SSPRK(2,2).

4.1 General SSP Runge-Kutta Schemes

We consider a general explicit Runge-Kutta scheme, which can be written in Shu-Osher form as

$$\begin{aligned} u^{(0)} &= u^n, \\ u^{(i)} &= \sum_{j=0}^{i-1} \left(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} F(u^{(j)}) \right), \quad i \in [1, m] \\ u^{n+1} &= u^{(m)}, \end{aligned} \tag{16}$$

where we require $\sum_{j=0}^{i-1} \alpha_{ij} = 1$ for all $i \in [1, m]$ for consistency.

The following theorem allows us to devise SSP schemes under some simple conditions.

Theorem 4.1. *Consider a time-dependent ODE given by $u_t = f(u)$, where u is the unknown. We suppose that the forward Euler method is stable under a certain norm, semi-norm or convex functional $\|\cdot\|$, meaning that Equation (15) holds, under the condition that $0 \leq \Delta t \leq \Delta t_{FE}$. If $\alpha_{ij}, \beta_{ij} \geq 0$, then the Runge-Kutta scheme (16) is SSP, with SSP-coefficient $C = \min_{i,j} \frac{\alpha_{ij}}{\beta_{ij}}$.*

Proof. Rewrite $u^{(i)}$ as

$$\begin{aligned} \|u^{(i)}\| &= \left\| \sum_{j=0}^{i-1} \left(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} F(u^{(j)}) \right) \right\| = \left\| \sum_{j=0}^{i-1} \alpha_{ij} \left(u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} F(u^{(j)}) \right) \right\| \\ &\leq \sum_{j=0}^{i-1} \alpha_{ij} \left\| u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} F(u^{(j)}) \right\| \leq \sum_{j=0}^{i-1} \alpha_{ij} \|u^n\| = \|u^n\|, \end{aligned}$$

where the first inequality holds since $\sum_{j=0}^{i-1} \alpha_{ij} = 1$ and $\|\cdot\|$ is convex, and the second inequality holds by assumption of the forward Euler being strongly stable. We note that the second inequality conditions $\Delta t \frac{\beta_{ij}}{\alpha_{ij}} \leq \Delta t_{FE}$, which holds under the condition $\Delta t \leq C \Delta t_{FE}$ with the given C. In specific, the above holds for $i = m$, yielding $\|u^{n+1}\| \leq \|u^n\|$, as wanted. □

4.2 Second-Order Optimal SSP Runge-Kutta method

Using the theorem from the previous section, we now devise an optimal second-order two-stage explicit SSP Runge-Kutta method. We call a method optimal within a subset of methods if it achieves that largest possible SSP-coefficient.

Theorem 4.2. *The optimal two-stage second-order explicit SSP Runge-Kutta methods with $\alpha_{ij}, \beta_{ij} \geq 0$ is given by*

$$\begin{aligned} u^{(1)} &= u^n + \Delta t f(u^n), \\ u^{n+1} &= \frac{1}{2} u^n + \frac{1}{2} u^{(1)} + \frac{1}{2} \Delta t f(u^{(1)}). \end{aligned}$$

We refer to this method as SSPRK(2,2).

Proof. We can write any two-stage explicit Runge-Kutta method in its Shu-Osher form as

$$u^{(1)} = \alpha_{10}u^n + \Delta t\beta_{10}f(u^n)$$
$$u^{n+1} = \alpha_{20}u^n + \alpha_{21}u^{(1)} + \beta_{20}\Delta t f(u^n) + \beta_{21}\Delta t f(u^{(1)}).$$

Consistency enforces that $\alpha_{10} = 1$ and $\alpha_{20} + \alpha_{21} = 1$. We observe the method can be written in a Butcher tableau as

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \beta_{10} & \beta_{10} & 0 \\ \hline & \beta_{10}\alpha_{21} + \beta_{20} & \beta_{21} \end{array}.$$

In order to achieve second-order accuracy, we require that

$$\beta_{10}\alpha_{21} + \beta_{20} + \beta_{21} = 1 \quad \text{and} \quad \beta_{21}\beta_{10} = \frac{1}{2}$$

Note that all the coefficients are uniquely determined by α_{21} and β_{10}, by writing

$$\alpha_{20} = 1 - \alpha_{21}, \quad \beta_{20} = 1 - \frac{1}{2\beta_{10}} - \alpha_{21}\beta_{10}, \quad \text{and} \quad \beta_{10} = \frac{1}{2\beta_{21}}.$$

It is clear that the scheme SSPRK(2,2) has a SSP-coefficient of $C = 1$. Assume there exist coefficients such that $C > 1$. Since $\alpha_{10} = 1$, we require $\beta_{10} < 1$. As such, $\alpha_{21} > \beta_{21} = \frac{1}{2\beta_{10}}$, meaning that also $\alpha_{21}\beta_{10} > \frac{1}{2}$. As such we can say that

$$\beta_{20} = 1 - \frac{1}{2\beta_{10}} - \alpha_{21}\beta_{10} < 1 - \frac{1}{2} - \frac{1}{2} < 0,$$

which contradicts the assumption that $\alpha_{ij}, \beta_{ij} \geq 0$. □

5 Numerical Results

We implement 3 different methods, based on the discretizations of the source term as given in Equations (7), (10) and (12), called A, B and C respectively. We set a gravitational constant of $g = 1$, and consider 100 spatial points over the spatial domain $[0, 1]$.

For the numerical solver of the semi-discrete system (6), we shall make use of SSPRK(2,2) devised in Section 4.2, which is stable under the condition

$$\Delta t \leq \Delta t_{FE},$$

where Δt_{FE} satisfies a CFL condition [7] given by

$$\lambda_{\max}\frac{\Delta t_{FE}}{\Delta x} \leq 1.$$

Here $\lambda_{\max}$ corresponds to the maximal eigenvalue of the Jacobian $\partial F/\partial U$ given in Section 2, and is evaluated at each iteration. We select

$$\Delta t = \frac{\Delta x}{2\lambda_{\max}}$$

to be on the safe side.

5.1 Still-Water Equilibrium

We first select a flat bottom topography given by $B(x) = 1$, with derivative $B_x(x) = 0$, and fix $(B+h)(x, 0) = 4$. We also set $u(x, 0) = 0$, since the water is supposed to be at rest. The results are depicted in Figure 2, and we observe that the equilibrium is preserved for each method.

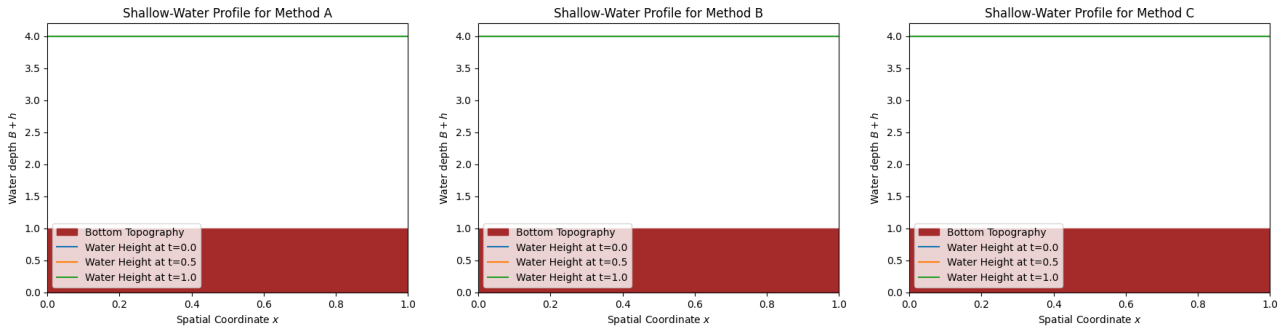

Figure 2: "Lake at rest" equilibrium with flat bottom topography.

As mentionned earlier, the first method should not be able to preserve the simple "lake at rest" equilibrium. We select a slightly inclined bottom topography, given by $B(x) = x$ with derivative $B_x(x) = 1$, and again fix $(B + h)(x, 0) = 4$ and $u(x, 0) = 0$. The results for methods A, B and C are shown in Figure 3. One observes that, as predicted, the still-water equilibrium is not maintained in method A, but is in the two others.

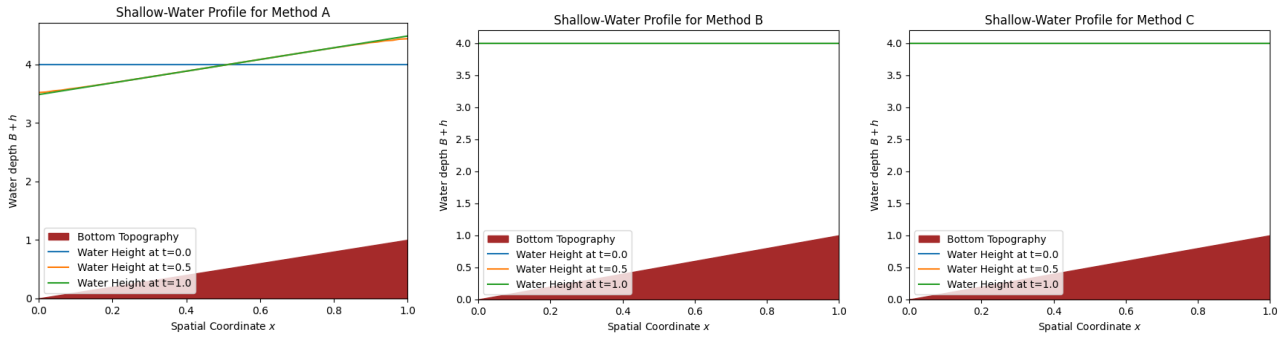

Figure 3: "Lake at rest" equilibrium with inclined bottom topography.

5.2 Moving-Water Equilibrium

We now move onto a more complicated equilibrium, namely the "moving water" equilibrium. We first consider a flat water height and a flat bottom, with constant non-zero velocity. In Figure 4, we picture the results of a bottom topography given by $B(x) = 1$, with derivative $B_x(x) = 0$, a constant initial water height of $h(x, 0) = 3$, and a constant initial velocity given by $u(x, 0) = 1$. All methods are capable of preserving this equilibrium state.

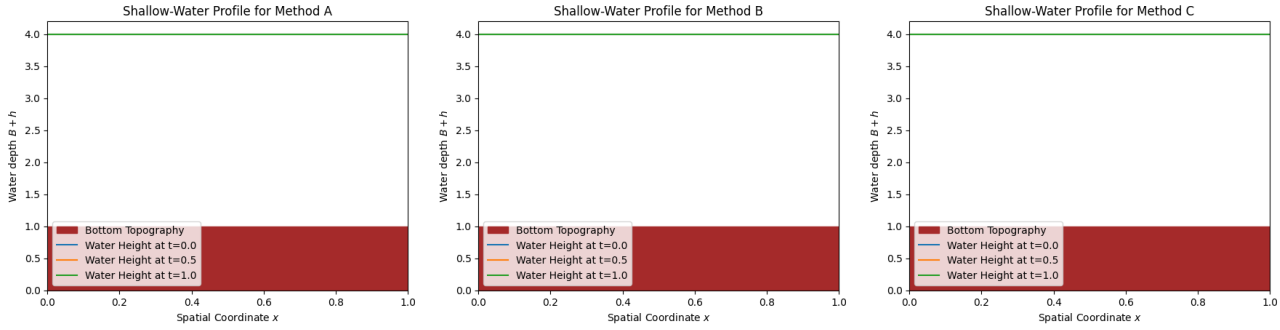

Figure 4: "Moving water" equilibrium with flat bottom topography and constant initial profile.

Next we consider a non-flat total water height. Due to its non-linear nature, we do not posses an analytical expression for the steady-state, and thus have to evolve into it. Figure 5 represents the evolution of a smoothened

step function, taking values 1 on the left and $1+\sqrt{3}$ on the right, where the width of the change is 0.2, and the velocity is selected such that the discharge hu is constantly 2 over the entire domain. The bottom is flat given by $B(x) = 1$, with derivative $B_x(x) = 0$. We observe that methods A and B converge to a shape in the style of a standing wave, but it continuously shifts to the left. Method C avoids this problem.

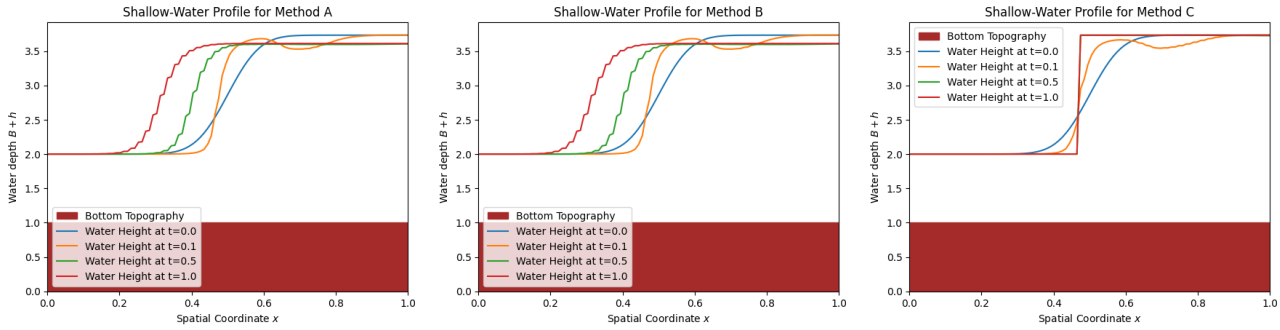

Figure 5: Evolution towards a "moving water" equilibrium with flat bottom topography.

We now take as initial profile the final profile of method C from the previous experiment, which we judge to be a near-steady-state. Figure 6 represents the evolution of such a standing wave for each method. We again observe that the waves in methods A and B shift to the left, and flatten out a little, whilst method C maintains the equilibrium quite precisely.

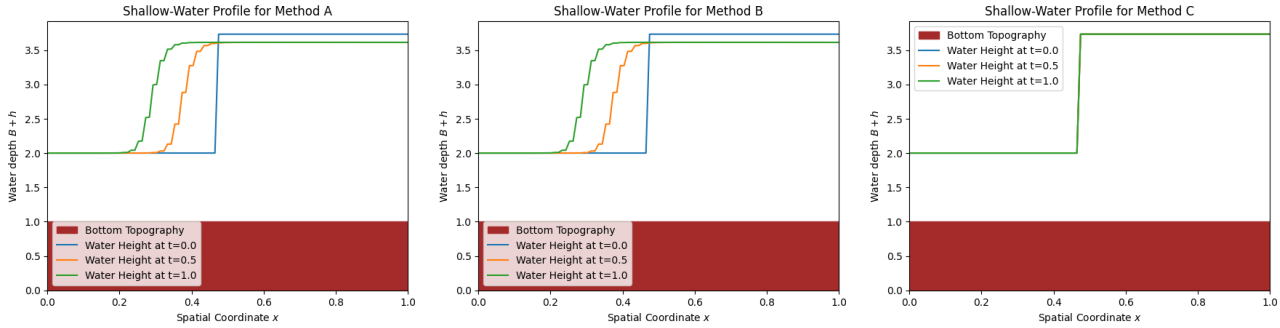

Figure 6: "Moving water" equilibrium with flat bottom topography and non-constant initial profile.

5.3 Non-Equilibria

We now present only method C, applied to various bottom topographies with different initial height and velocity profiles. The results at several timesteps are shown in Figure 7. The first experiment runs on a flat bottom topography, which an initial flat water surface. The velocity is given by

$$u(x,0) = \begin{cases} 1 & \text{if } 0 \le x \le 0.3 \\ 0 & \text{if } 0.3 < x < 0.7 \\ -1 & \text{if } 0.7 \le x \le 1 \end{cases}.$$

As such, the water will collide and form a spike around the center of the domain. In the second example, the initial water height is represented by a Gaussian, and the initial water velocity is constantly zero. We observe that the water flattens out with time. In the final example, the total water height is initially constant, the velocity is constantly 1, and the bottom topography is given by a Gaussian. We observe a wave creating.

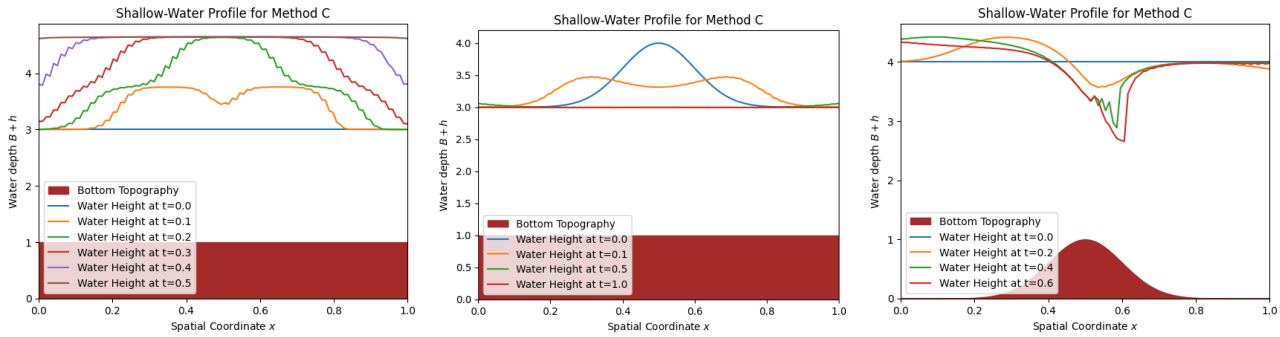

Figure 7: Non-equilibrium states.

6 Conclusion

This report develops a well-balanced scheme for the one-dimensional hyperbolic shallow water equations, given by the Saint-Venant equations. We develop a finite volume scheme for the spatial discretization, which yields a semi-discrete system. The source term is then discretized carefully in order to first achieve a "still-water" equilibrium preserving scheme, and next to achieve a "moving-water" equilibium preserving scheme. The notion of *strong stability preserving schemes* are discussed, and used to solve the semi-discrete system. Finally the methods are tested and compared to each other.

6.1 Further Work

Although this report covered a large portion of the basics within well-balanced methods for shallow water equations, it also allows for further improvements on multiple axis.

The methods developped are not *postitivity preserving*, meaning that the water height remains constant at all times. Indeed, due to the linear reconstruction of the bottom topography, it might happen that $w < B$ at certain points, in which case the water height becomes negative. After this the equations that follow become ill-posed, and the entire system suffers an instabilility. In the same fashion, our method does not allow for dry or near-dry areas in the domain.

On another hand, we only developped a one-dimensional scheme. Although the ideas remain comparable when extending to two spatial dimensions, this becomes unarguable more complex and requires more subtlety and computations.

6.2 Reproducability

The code created and used for the numerical experiments has been uploaded as a public Python package. Download instructions and documentation are available via https://pypi.org/project/SWE-Solver/.

References

[1] M. Chalfen and A. Niemiec. Analytical and numerical solution of Saint-Venant equations. *Journal of Hydrology*, 86(1-2):1–13, 1986.

[2] Y. Cheng and A. Kurganov. Moving-water equilibria preserving central-upwind schemes for the shallow water equations. *Communications in Mathematical Sciences*, 14(6):1643–1663, 2016.

[3] A. J.-C. de Saint-Venant et al. Théorie du mouvement non-permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. *Comptes Rendus de l'Académie des Sciences Paris*, 73(147-154):237–240, 1871.

[4] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.

[5] A. Kurganov and D. Levy. Central-upwind schemes for the saint-venant system. *ESAIM: Mathematical Modelling and Numerical Analysis*, 36(3):397–425, 2002.

[6] A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the saint-venant system. *Communications in Mathematical Sciences*, 5(1):133–160, 2007.

[7] R. J. LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge University Press, 2002.