# A very easy high-order well-balanced reconstruction for hyperbolic systems with source terms

Christophe Berthon, Solène Bulteau, Françoise Foucher, Meissa M'Baye, Victor Michel-Dansac

# A very easy high-order well-balanced reconstruction for hyperbolic systems with source terms

Christophe Berthon* Solène Bulteau† Françoise Foucher‡* Meissa M'Baye§*
Victor Michel-Dansac¶‖

**Abstract**

When adopting high-order finite volume schemes based on MUSCL reconstruction techniques to approximate the weak solutions of hyperbolic systems with source terms, the preservation of the steady states turns out to be very challenging. Indeed, the designed reconstruction must preserve the steady states under consideration in order to get the required well-balancedness property. A priori, to capture such a steady state, one needs to solve some strongly nonlinear equations. Here, we design a very easy correction to high-order finite volume methods. This correction can be applied to any scheme of order greater than or equal to 2, such as a MUSCL-type scheme, and ensures that this scheme exactly preserves the steady solutions. The main discrepancy with usual techniques lies in avoiding the inversion of the nonlinear function that governs the steady solutions. Moreover, for under-determined steady solutions, several nonlinear functions must be considered simultaneously. Since the derived correction only considers the evaluation of the governing nonlinear functions, we are able to deal with under-determined stationary systems. Several numerical experiments illustrate the relevance of the proposed well-balanced correction, as well as its main limitation, namely the fact that it may fail at being both well-balanced and more than second-order accurate for a specific class of initial conditions.

## 1 Introduction

### 1.1 General framework

The present work is devoted to the numerical approximation of the weak solutions of an evolution law of the form

$$\partial_t w + \partial_x f(w) = S(w,x), \quad x \in \mathbb{R}, \quad t > 0, \tag{1.1}$$

where $w : \mathbb{R} \times \mathbb{R}^+ \to \Omega \subset \mathbb{R}^N$ denotes the unknown vector. The set $\Omega$ stands for the set of the admissible states and it is assumed to be convex. The flux function $f : \Omega \to \mathbb{R}^n$ is assumed to be smooth enough, say $C^1$. The source term $S : \Omega \times \mathbb{R} \to \mathbb{R}^N$ is also supposed to be smooth enough for $w \in \operatorname{Int}\Omega$. We also emphasize the need for the function $x \in \mathbb{R} \mapsto S(w,x)$ to be a known smooth function for all $w \in \Omega$. For stability reasons, in the present work, the matrix $\nabla_w f(w)$ is assumed to be diagonalizable in $\mathbb{R}$ so that the homogeneous system

*Université de Nantes, CNRS UMR 6629, Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière, BP 92208, 44322 Nantes, France (christophe.berthon@univ-nantes.fr, https://www.math.sciences.univ-nantes.fr/~berthon/WEBenglish/berthon.html).

†Maison de la Simulation, USR 3441, FR-91191 Gif-sur-Yvette (solene.bulteau1@gmail.com, https://solenebulteau.wordpress.com).

‡École Centrale de Nantes, CNRS UMR 6629, Laboratoire de Mathématiques Jean Leray, 1 rue de La Noë, BP 92101, 44321 Nantes Cedex 3, France (francoise.foucher@ec-nantes.fr).

§Laboratoire de Mathématiques de la Décision et d'Analyse Numérique (LMDAN), FASEG, Université Cheikh Anta Diop, BP 16889, Dakar, Sénégal (meissaths@gmail.com).

¶Université de Strasbourg, CNRS, Inria, IRMA, F-67000 Strasbourg, France (victor.michel-dansac@inria.fr, http://irma.math.unistra.fr/~micheldansac/index_en.html)

‖corresponding author

extracted from (1.1) is hyperbolic. The PDE system (1.1) is endowed with initial data $w(x, t = 0) = w^0(x)$, where $w^0(x) \in \Omega$ for all $x \in \mathbb{R}$ is a given function.

Because of the source term $S(w, x)$, there exists non-trivial steady solutions governed by

$$\begin{cases} \partial_x f(w) = S(w, x), \\ w(x_0) = w_0, \end{cases} \tag{1.2}$$

where $w_0$ is a given state in $\Omega$ and $x_0$ a given point in $\mathbb{R}$.

Now, if the above system can be integrated, there exists $G : \Omega \times \mathbb{R} \to \mathbb{R}^N$ such that the stationary solutions are governed by

$$\begin{cases} \partial_x G(w, x) = 0, \\ w(x_0) = w_0. \end{cases}$$

In fact, it is not always possible to integrate (1.2). Even then, $G$ is not necessarily unique and, generally, it must be restricted according to some invariant domains. Usually, the steady solutions are restricted to some particular definition of $G$. Then, we have to deal with a sequence $(G_\ell)_{1 \leq \ell \leq L}$ with $L < +\infty$ given.

Equipped with these comments, in this work, we only consider steady solutions defined by

$$\begin{cases} \partial_x G_\ell(w, x) = 0, \quad 1 \leq \ell \leq L, \\ w(x_0) = \Pi_\ell^{eq}(w_0), \end{cases} \tag{1.3}$$

where we have denoted by $\Pi^{eq}(w)$ the projection of $w$ over the invariant domain under consideration. More broadly, each steady state is characterized by two sets of equations. The first one, involving $G$, controls the space variations of the steady-state solution. The second one, which involves $\Pi$, controls the solution in a pointwise fashion.

## 1.2 Illustrating models

In order to illustrate the relevance of such a definition of the steady states, let us present some examples of particular interest. First, let us adopt the well-known shallow water model given by

$$w = \begin{pmatrix} h \\ q \end{pmatrix}, \quad f(w) = \begin{pmatrix} q \\ \dfrac{q^2}{h} + g\dfrac{h^2}{2} \end{pmatrix}, \quad S(w, x) = \begin{pmatrix} 0 \\ -gh\partial_x z \end{pmatrix}, \tag{1.4}$$

where $g > 0$ is the gravity constant, $z(x)$ the given smooth topography function, $h$ is the water height and $q$ is the water discharge. The smooth steady solutions (see [4, 11, 27]) are given by

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left( \dfrac{q^2}{2h^2} + g(h + z) \right) = 0, \end{cases} \tag{1.5}$$

with $w(x_0) = w_0$ for a given $w_0$ in $\Omega$, where the set $\Omega$ of admissible states is defined as follows:

$$\Omega = \left\{ {}^t(h, q) \in \mathbb{R}^2;\ h \geq 0,\ q \in \mathbb{R} \right\}. \tag{1.6}$$

As a consequence, we immediately obtain

$$G(w, x) = \begin{pmatrix} q \\ \dfrac{q^2}{2h^2} + g(h + z) \end{pmatrix} \quad \text{and} \quad \Pi^{eq}(w) = w. \tag{1.7}$$

Now, if we restrict the definition of the steady states by only considering the usual lake at rest given by

$$\begin{cases} q = 0, \\ \partial_x(h + z) = 0, \end{cases}$$

then we easily get

$$G(w,x) = \begin{pmatrix} 0 \\ h+z \end{pmatrix} \quad \text{and} \quad \Pi^{eq}(w) = \begin{pmatrix} h \\ 0 \end{pmatrix}. \tag{1.8}$$

The second example we present is given by the shallow water equations with a Manning friction source term (see [38, 36, 20]) and a flat bottom. This model reads

$$w = \begin{pmatrix} h \\ q \end{pmatrix}, \quad f(w) = \begin{pmatrix} q \\ \frac{q^2}{h} + g\frac{h^2}{2} \end{pmatrix}, \quad S(w,x) = \begin{pmatrix} 0 \\ -\kappa \frac{q|q|}{h^\eta} \end{pmatrix}, \tag{1.9}$$

where $\kappa > 0$ is the friction coefficient and $\eta \neq 1$ is the Manning exponent. The set of the admissible states is given by (1.6). After [38], the stationary solutions are given as follows:

$$\begin{cases} \partial_x q = 0, \\ \partial_x \left( -q^2 \dfrac{h^{\eta-1}}{\eta-1} + g\dfrac{h^{\eta+2}}{\eta+2} + \kappa x q|q| \right) = 0, \end{cases}$$

with $w(x_0) = w_0$ for a given $w_0 \in \Omega$, so that we immediately obtain

$$G(w,x) = \begin{pmatrix} q \\ -q^2 \dfrac{h^{\eta-1}}{\eta-1} + g\dfrac{h^{\eta+2}}{\eta+2} + \kappa x q|q| \end{pmatrix} \quad \text{and} \quad \Pi^{eq}(w) = w. \tag{1.10}$$

Next, we present a system involving a non-unique definition of $G$, the Euler model with gravity (see [21, 49]). This model reads as follows:

$$w = \begin{pmatrix} \rho \\ q \\ E \end{pmatrix}, \quad f(w) = \begin{pmatrix} q \\ \frac{q^2}{\rho} + p \\ (E+p)\frac{q}{\rho} \end{pmatrix}, \quad S(w,x) = \begin{pmatrix} 0 \\ -\rho\partial_x\varphi \\ -q\partial_x\varphi \end{pmatrix}, \tag{1.11}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ stands for a given smooth gravitational potential and $p := p(\rho, E - \frac{1}{2}\frac{q^2}{\rho})$ denotes the pressure law, with $E$ the total energy. The set of admissible states is defined here by

$$\Omega = \left\{ {}^t(\rho, q, E) \in \mathbb{R}^3;\ \rho > 0,\ q \in \mathbb{R},\ E - \frac{1}{2}\frac{q^2}{\rho} > 0 \right\}.$$

Concerning the steady solutions, we are concerned by steady solutions at rest governed by (see [21, 49])

$$\begin{cases} q = 0, \\ \partial_x p + \rho\partial_x\varphi = 0, \end{cases} \quad \text{with} \quad w(x_0) = \begin{pmatrix} \rho_0 \\ 0 \\ E_0 \end{pmatrix}.$$

Once again, the system to govern the steady-state solutions turns out to be under-determined and we have to focus on particular families of steady solutions. According to [21, 49], three families of steady states are of prime importance. The first family is given by

$$\begin{cases} q = 0, \\ \partial_x \rho = 0, \\ \partial_x(p + \rho\varphi) = 0, \end{cases} \quad \text{with} \quad w(x_0) = \begin{pmatrix} \rho_0 \\ 0 \\ E_0 \end{pmatrix},$$

for all $\rho_0 > 0$ and $E_0 > 0$.

In order to define both second and third families of steady states, we have to impose that the function $E \mapsto p(\rho, E)$ is invertible and we denote by $p_E^{-1}(\rho, .)$ this inverse function so that $p_E^{-1}(\rho, p(\rho, E)) = E$. Now, the second steady state family reads

$$\begin{cases} q = 0, \\ \partial_x(p - \kappa\rho) = 0, \\ \partial_x(\varphi + \kappa \ln\rho) = 0, \end{cases} \quad \text{with} \quad w(x_0) = \begin{pmatrix} \rho_0 \\ 0 \\ p_E^{-1}(\rho_0, \kappa\rho_0) \end{pmatrix},$$

where $\kappa > 0$ is a given parameter. The last steady state family is defined by

$$\begin{cases} q = 0, \\ \partial_x(p - \kappa\rho^\gamma) = 0, \\ \partial_x\left(\dfrac{\kappa\gamma}{\gamma - 1}\rho^{\gamma-1} + \varphi\right) = 0, \end{cases} \quad \text{with} \quad w(x_0) = \begin{pmatrix} \rho_0 \\ 0 \\ p_E^{-1}(\rho_0, \kappa\rho_0^\gamma) \end{pmatrix}, \tag{1.12}$$

where $\gamma > 1$ is a given parameter.

As a consequence, we get

$$G_1(w, x) = \begin{pmatrix} q \\ \rho \\ p + \rho\varphi \end{pmatrix}, \quad G_2(w, x) = \begin{pmatrix} q \\ p - \kappa\rho \\ \varphi + \kappa \ln\rho \end{pmatrix}, \quad G_3(w, x) = \begin{pmatrix} q \\ p - \kappa\rho^\gamma \\ \dfrac{\kappa\gamma}{\gamma - 1}\rho^{\gamma-1} + \varphi \end{pmatrix}, \tag{1.13}$$

with

$$\Pi_1^{eq}(w) = \begin{pmatrix} \rho \\ 0 \\ E \end{pmatrix}, \quad \Pi_2^{eq}(w) = \begin{pmatrix} \rho \\ 0 \\ p_E^{-1}(\rho, \kappa\rho) \end{pmatrix} \quad \text{and} \quad \Pi_3^{eq}(w) = \begin{pmatrix} \rho \\ 0 \\ p_E^{-1}(\rho, \kappa\rho^\gamma) \end{pmatrix}. \tag{1.14}$$

## 1.3 Main motivation

Now, considering the derivation of numerical schemes approximating the solutions of (1.1) and able to accurately, or even exactly, capture the steady solutions defined by (1.3) has been an important challenge during the two last decades. Numerous techniques have been designed for the shallow water model with topography (1.4) supplemented by the lake at rest (1.8). For a non-exhaustive bibliography, the reader is referred to [4, 2, 11, 41, 24, 9, 34, 13, 8, 18, 19]. More recently, in [27, 40, 48, 46, 5, 6, 37, 38], extensions are given in order to deal with moving steady states given by (1.7). In [12, 38, 30], the Manning-type friction source term is adopted and suitable discretizations are introduced to capture steady states given by (1.10). Regarding the discretization of the Euler model with gravity (1.11), the reader is referred to [14, 17, 49, 35, 47, 31, 21, 3, 33, 43] where numerical strategies are developed to capture steady states according to the pairs $(G_\ell, \Pi_\ell^{eq})$ defined by (1.13) – (1.14).

In the present work, we are not interested in the derivation of well-balanced schemes, namely schemes able to capture steady solutions given by (1.3). Here, our purpose concerns the high-order extensions obtained by involving a polynomial reconstruction procedure. Indeed, as soon as the well-balancedness property must be preserved, the reconstruction may involve strong difficulties. In particular, to be well-balanced, the usual reconstruction approaches need to invert $G_\ell(w, x)$ with respect to $w$ for one given $\ell$ as long as the function $w \mapsto G_\ell(w, x)$ is invertible. We immediately understand that dealing simultaneously with distinct functions $G_\ell(w, x)$ does not seem reachable. Moreover, imposing that the application $w \mapsto G_\ell(w, x)$ is invertible is a strong assumption, not satisfied in general by physical models.

In this work, we present a very easy strategy to force any reconstruction procedure to preserve the steady solutions given by (1.3) just evaluating the applications $G_\ell(w, x)$ according to the projection $\Pi_\ell^{eq}(w)$. Let us emphasize that this technique still requires the prior knowledge of a well-balanced first-order scheme, which may be a challenging endeavor. To address such an issue, the present work is organized as follows. In

order to set the framework and the main notations, in Section 2 we introduce the numerical schemes and the usual MUSCL second-order strategy [45, 32, 44]. In addition, we present the main issues when enforcing the polynomial reconstruction to be well-balanced. Section 3 is then devoted to the strategy designed here, which ensures that the expected well-balancedness property is satisfied by any reconstruction. The proposed improvement comes from a suitable evaluation of the pairs $(G_\ell, \Pi_\ell^{eq})_{1\leq \ell \leq L}$; $G_\ell$ is never inverted. In Section 4, we present a high-order well-balanced extension. Finally, Section 5 is devoted to several numerical experiments to illustrate the relevance of the designed high-order reconstruction improvement. We also illustrate the current main limitation of the scheme. Indeed, starting from a perturbed steady state solution, we have so far not been able to correctly discretize the initial condition to achieve both the well-balance property and an order of accuracy greater than two. Section 6 concludes this discussion and suggests some perspectives.

## 2 Issues of the well-balanced second-order MUSCL schemes

To approximate the solutions of (1.1), the space is discretized by introducing a sequence of cells $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, for all $i \in \mathbb{Z}$, with a constant size $\Delta x$. We denote by $x_i = (x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})/2$ the center of each cell. We set $t^{n+1} = t^n + \Delta t$ to discretize the time domain with a time step $\Delta t$. In general, $\Delta t$ is restricted according to a CFL condition.

At time $t^n$, we denote by $w_i^n$ a constant approximation of the solution of (1.1) over the cell $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$. To evolve this approximation in time, we adopt a finite volume scheme of the form

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n\right) + \frac{\Delta t}{2}\left(S_{i-\frac{1}{2}}^n + S_{i+\frac{1}{2}}^n\right), \tag{2.1}$$

where we have set

$$f_{i+\frac{1}{2}}^n = f_\Delta(w_i^n, w_{i+1}^n) \quad \text{and} \quad S_{i+\frac{1}{2}}^n = S_\Delta(w_i^n, w_{i+1}^n, x_i, x_{i+1}, \Delta x).$$

In order to get a consistent scheme, the numerical flux function $f_\Delta$ and the discrete source term $S_\Delta$ are assumed to be Lipschitz-continuous and to verify

$$f_\Delta(w, w) = f(w) \quad \text{and} \quad S_\Delta(w, w, x, x, 0) = S(w, x). \tag{2.2}$$

At this level, the reader is referred to the large literature devoted to the derivation of a well-balanced scheme according to the system of interest. Here, we have imposed the well-balancedness property according to the definition (1.3) of the steady states. As a consequence, we get $w_i^{n+1} = w_i^n$ as long as, for all $i$ in $\mathbb{Z}$, we have

$$G_\ell(w_i^n, x_i) = G_\ell(w_{i+1}^n, x_{i+1}) \quad \text{and} \quad w_i^n = \Pi_\ell^{eq}(w_i^n) \quad \text{with} \quad 1 \leq \ell \leq L, \tag{2.3}$$

to enforce the invariant domain according to (1.3).

Now, we focus on a second-order extension, see for instance [44, 45, 7]. To address such an issue, we have to introduce suitable reconstructed states, denoted by $w_{i+\frac{1}{2}}^{\pm}$, on each side of the interface $x_{i+\frac{1}{2}}$. This reconstruction is said to be second-order accurate in space if, for all $i \in \mathbb{Z}$, we have

$$w_{i+\frac{1}{2}}^{-} = w(x_{i+\frac{1}{2}}, t^n) + \mathcal{O}(\Delta x^2) \quad \text{and} \quad w_{i+\frac{1}{2}}^{+} = w(x_{i+\frac{1}{2}}, t^n) + \mathcal{O}(\Delta x^2), \tag{2.4}$$

for some smooth function $x \mapsto w(x, t^n)$ such that

$$w_i^n = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x, t^n) dx.$$

Equipped with this second-order reconstruction, from the first-order scheme (2.1), we define a second-order scheme as follows:

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}}^{\pm} - f_{i-\frac{1}{2}}^{\pm}\right) + \Delta t S_i^{\pm}, \tag{2.5}$$

where we have set $f^{\pm}_{i+\frac{1}{2}} = f_\Delta(w^-_{i+\frac{1}{2}}, w^+_{i+\frac{1}{2}})$, and where $S^{\pm}_i$ is a second-order approximation of the source term average, i.e.

$$S^{\pm}_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(w(x,t^n),x)\,dx + \mathcal{O}(\Delta x^2). \tag{2.6}$$

A classical choice for such second-order accurate schemes is to use the second-order midpoint approximation:

$$S^{\pm}_i = S\left(\frac{1}{2}\left(w^+_{i-\frac{1}{2}} + w^-_{i+\frac{1}{2}}\right), x_i\right). \tag{2.7}$$

It is clear that second-order accuracy is achieved as soon as the reconstructed states are defined. At the interface $x_{i+\frac{1}{2}}$, the reconstructed states read (for instance, see [7, 44, 45])

$$\begin{aligned} w^-_{i+\frac{1}{2}} &= w^n_i + \frac{1}{2}\mathcal{L}(w^n_i - w^n_{i-1}, w^n_{i+1} - w^n_i),\\ w^+_{i+\frac{1}{2}} &= w^n_{i+1} - \frac{1}{2}\mathcal{L}(w^n_{i+1} - w^n_i, w^n_{i+2} - w^n_{i+1}), \end{aligned} \tag{2.8}$$

where $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ are Lipschitz-continuous functions, which satisfy

$$\begin{aligned} &\mathcal{L}(w,w) = w \quad \text{for all } w \in \mathbb{R}^N,\\ &\exists M > 0 \text{ such that } \|\mathcal{L}(w_L, w_R)\| \le M\max(\|w_L\|, \|w_R\|), \quad \forall\, w_L, w_R \in \mathbb{R}^N. \end{aligned}$$

A large body of literature is devoted to introduce suitable definition of $\mathcal{L}$ (for instance, see [32] and references therein).

Now, by adopting (2.8), the steady states are, in general, not preserved by such a reconstruction. Indeed, in order to recover the expected well-balancedness property, we require

$$w^-_{i+\frac{1}{2}} = w^n_i \quad \text{and} \quad w^+_{i+\frac{1}{2}} = w^n_{i+1} \quad \text{for all } i \in \mathbb{Z}, \tag{2.9}$$

as soon as $(w^n_i)_{i\in\mathbb{Z}}$ defines a steady state according to (2.3). Except for linear steady states, the steady condition (2.9) is lost. As a consequence, a particular attention must be paid on the definition of $\mathcal{L}$ to preserves the steady states.

Currently, the reconstruction on $G_\ell$ instead of $w$ is preferred (see [2, 40]). For a fixed $\ell$, denoted by $\ell^\star$, we perform the reconstruction as follows:

$$\begin{aligned} G^-_{i+\frac{1}{2}} &= \quad G_{\ell^\star}(w^n_i, x_i) \quad + \frac{1}{2}\mathcal{L}\Big(G_{\ell^\star}(w^n_i, x_i) - G_{\ell^\star}(w^n_{i-1}, x_{i-1}),\, G_{\ell^\star}(w^n_{i+1}, x_{i+1}) - G_{\ell^\star}(w^n_i, x_i)\Big),\\ G^+_{i+\frac{1}{2}} &= G_{\ell^\star}(w^n_{i+1}, x_{i+1}) - \frac{1}{2}\mathcal{L}\Big(G_{\ell^\star}(w^n_{i+1}, x_{i+1}) - G_{\ell^\star}(w^n_i, x_i),\, G_{\ell^\star}(w^n_{i+2}, x_{i+2}) - G_{\ell^\star}(w^n_{i+1}, x_{i+1})\Big), \end{aligned}$$

The reconstructed states $w^{\pm}_{i+\frac{1}{2}}$ at the interface $x_{i+\frac{1}{2}}$ are then defined by

$$\begin{cases} G_{\ell^\star}(w^-_{i+\frac{1}{2}}, x_{i+\frac{1}{2}}) = G^-_{i+\frac{1}{2}},\\ G_{\ell^\star}(w^+_{i+\frac{1}{2}}, x_{i+\frac{1}{2}}) = G^+_{i+\frac{1}{2}}. \end{cases} \tag{2.10}$$

We immediately remark that (2.9) holds as soon as $(w^n_i)_{i\in\mathbb{Z}}$ defines a steady state for $G_{\ell^\star}$ according to (2.3). However, the function $w \mapsto G_{\ell^\star}(w,x)$ must be inverted. Such a procedure may be very costly, or even impossible to carry out if $G_{\ell^\star}(.,x)$ is not invertible.

In fact, in the simpler situation of the lake at rest for the shallow water equation, where $G$ is given by (1.8), we have to solve a linear $2\times 2$ system. But for a moving steady state, i.e. with $G$ defined by (1.7), the uniqueness of the reconstructed states $w^{\pm}_{i+\frac{1}{2}}$ is no longer ensured. Next, considering (1.10), neither the existence nor the uniqueness of $w^{\pm}_{i+\frac{1}{2}}$ is ensured.

Moreover, adopting such a procedure needs to fix $\ell$. As a consequence, it is not possible to deal with steady states governed by several families $G_\ell(w,x)$ with $1 \leq \ell \leq L$ for $L \geq 2$. Such a restriction arises for instance for the Euler equations with gravity, where we consider three steady state families.

To summarize the failure of the usual well-balanced reconstruction technique, the reconstructed states, solution of (2.10), may not exist or not be unique. Moreover, since we have to solve a nonlinear system, the evaluation of the reconstructed states turns out to be computationally expensive. In addition, such a reconstruction technique preserves only one steady state family while some systems involve several steady state families.

# 3 A very easy well-balanced MUSCL reconstruction

The objective is now to derive a reconstruction technique able to preserve the steady states but never involving an inversion of $G_\ell$. To address such an issue, we suggest to improve the usual reconstruction (2.8) as follows:

$$\begin{aligned} \tilde{w}^-_{i+\frac{1}{2}} &= w^n_i + \frac{1}{2}\,\theta^n_{i+\frac{1}{2}}\,\mathcal{L}(w^n_i - w^n_{i-1}, w^n_{i+1} - w^n_i), \\ \tilde{w}^+_{i+\frac{1}{2}} &= w^n_{i+1} - \frac{1}{2}\,\theta^n_{i+\frac{1}{2}}\,\mathcal{L}(w^n_{i+1} - w^n_i, w^n_{i+2} - w^n_{i+1}), \end{aligned} \tag{3.1}$$

where the correction $\theta^n_{i+\frac{1}{2}}$ must be an approximation of 1, with at least second-order accuracy, which vanishes for pairs $(w^n_i, w^n_{i+1})$ satisfying (2.3). We propose the following formulation of $\theta^n_{i+\frac{1}{2}}$:

$$\theta^n_{i+\frac{1}{2}} = \frac{\varepsilon^n_{i+\frac{1}{2}}}{\varepsilon^n_{i+\frac{1}{2}} + \left(\dfrac{\Delta x}{C^n_{i+\frac{1}{2}}}\right)^k}, \text{ with} \tag{3.2}$$

$$\varepsilon^n_{i+\frac{1}{2}} = \prod_{\ell=1}^{L} \Big( \|G_\ell(w^n_{i+1}, x_{i+1}) - G_\ell(w^n_i, x_i)\| + \|w^n_{i+1} - \Pi^{eq}_\ell(w^n_{i+1})\| + \|w^n_i - \Pi^{eq}_\ell(w^n_i)\| \Big), \tag{3.3}$$

where $k \geq 2$ must be selected and where $C^n_{i+\frac{1}{2}} \neq 0$ is any expression independent from $\Delta x$. We shall suggest an expression of $C^n_{i+\frac{1}{2}}$ in the numerical experiments. From now on, it is worth noting that $\varepsilon^n_{i+\frac{1}{2}} = 0$ if and only if the pair $(w^n_i, w^n_{i+1})$ defines a local steady state, according to (2.3), at the interface $x_{i+\frac{1}{2}}$.

Concerning the source term discretization, we adopt the following definition:

$$\tilde{S}^\pm_i = \frac{1}{2}\left(\left(1 - \theta^n_{i-\frac{1}{2}}\right) S^n_{i-\frac{1}{2}} + \left(1 - \theta^n_{i+\frac{1}{2}}\right) S^n_{i+\frac{1}{2}}\right) + \frac{1}{2}\left(\theta^n_{i-\frac{1}{2}} + \theta^n_{i+\frac{1}{2}}\right) S^\pm_i, \tag{3.4}$$

where $S^\pm_i$ is given by (2.7) and where $S^n_{i\pm\frac{1}{2}}$ comes from the first-order discretization (2.1). As a consequence, the second-order MUSCL scheme now reads

$$w^{n+1}_i = w^n_i - \frac{\Delta t}{\Delta x}\left(f^\pm_{i+\frac{1}{2}} - f^\pm_{i-\frac{1}{2}}\right) + \Delta t \tilde{S}^\pm_i, \tag{3.5}$$

where we have set

$$f^\pm_{i+\frac{1}{2}} = f_\Delta(\tilde{w}^-_{i+\frac{1}{2}}, \tilde{w}^+_{i+\frac{1}{2}}). \tag{3.6}$$

Before we establish the main properties satisfied by the second-order MUSCL scheme (3.5) – (3.6) with the reconstructed states (3.1) and the source term discretization (3.4), let us recall the definition of the order of accuracy that is adopted here (for instance, see [11]).

**Definition 1.** *For some smooth solution $w(x,t)$ of (1.1), let us consider*

$$w^n_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x,t^n)\,dx. \tag{3.7}$$

*Define $w_i^{n+1}$ by (2.5). The scheme (2.5) is said to be of order $\tau$ in time and $\delta$ in space if, for all $i$ in $\mathbb{Z}$, we have*

$$w_i^{n+1} = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x, t^n + \Delta t)\, dx - \frac{\Delta t}{\Delta x}\left(\mathcal{F}_{i+\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}\right) + \Delta t \mathcal{S}_i, \tag{3.8}$$

*where $\mathcal{F}_{i+\frac{1}{2}} = \mathcal{O}(\Delta t^\tau) + \mathcal{O}(\Delta x^\delta)$ and $\mathcal{S}_i = \mathcal{O}(\Delta t^\tau) + \mathcal{O}(\Delta x^\delta)$.*

Now, arguing the above definition of the order of accuracy, the improved reconstruction technique based on $\theta_{i+\frac{1}{2}}$ is established to yield a second-order accurate and well-balanced scheme.

**Theorem 2.** *Given a well-balanced first-order scheme, the scheme (3.5) – (3.6), with reconstructed states given by (3.1) and a source term discretization given by (3.4), satisfies the following properties:*

*(i) it is second-order accurate in space for unsteady solutions;*

*(ii) it is well-balanced, i.e. it exactly preserves steady solutions: if $(w_i^n)_{i\in\mathbb{Z}}$ defines a steady state according to (2.3), then $w_i^{n+1} = w_i^n$ for all $i$ in $\mathbb{Z}$;*

*(iii) it is robust, i.e. if the original reconstruction (2.8) preserves $\Omega$, then $\Omega$ remains invariant by the improved reconstruction (3.1).*

*Proof.* We establish properties (i), (ii) and (iii) in order.

(i) To establish the order of accuracy, let us consider $w(x,t)$ a smooth unsteady solution of (1.1). By integration of (1.1) over $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times (t^n, t^n + \Delta t)$, we get

$$\begin{aligned}
&\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x, t^n + \Delta t)\, dx - \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x, t^n)\, dx \\
&\qquad + \frac{\Delta t}{\Delta x}\left(\frac{1}{\Delta t}\int_{t^n}^{t^n+\Delta t} f\left(w(x_{i+\frac{1}{2}}, t)\right)\, dt - \frac{1}{\Delta t}\int_{t^n}^{t^n+\Delta t} f\left(w(x_{i-\frac{1}{2}}, t)\right)\, dt\right) \\
&\qquad = \Delta t \frac{1}{\Delta t \Delta x}\int_{t^n}^{t^n+\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(w(x,t), x)\, dx\, dt.
\end{aligned} \tag{3.9}$$

With $(w_i^n)_{i\in\mathbb{Z}}$ given by (3.7) and $(w_i^{n+1})_{i\in\mathbb{Z}}$ given by (3.5) – (3.6), a straightforward computation gives the expected relation (3.8), where we have set

$$\begin{aligned}
\mathcal{F}_{i+\frac{1}{2}} &= f^{\pm}_{i+\frac{1}{2}} - \frac{1}{\Delta t}\int_0^{\Delta t} f\left(w(x_{i+\frac{1}{2}}, t^n + t)\right)\, dt, \\
\mathcal{S}_i &= \tilde{S}_i^{\pm} - \frac{1}{\Delta t \Delta x}\int_0^{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(w(x, t^n + t), x)\, dx\, dt,
\end{aligned}$$

where $f^{\pm}_{i+\frac{1}{2}}$ and $\tilde{S}_i^{\pm}$ are respectively given by (3.6) and (3.4).

We first treat the approximation of the flux function. By definition of $\theta^n_{i+\frac{1}{2}}$, given by (3.2), as long as $\varepsilon^n_{i+\frac{1}{2}}$ does not vanish, we have $\theta^n_{i+\frac{1}{2}} = 1 + \mathcal{O}(\Delta x^k)$. As a consequence, in the current unsteady context, we get

$$\tilde{w}^-_{i+\frac{1}{2}} = w^-_{i+\frac{1}{2}} + \mathcal{O}(\Delta x^k) \quad \text{and} \quad \tilde{w}^+_{i+\frac{1}{2}} = w^+_{i+\frac{1}{2}} + \mathcal{O}(\Delta x^k),$$

where $w^{\pm}_{i+\frac{1}{2}}$ are given by (2.8), and with $k \geq 2$. Since (2.4) holds for the second-order polynomial reconstruction, we immediately obtain

$$\tilde{w}^-_{i+\frac{1}{2}} = w(x_{i+\frac{1}{2}}, t^n) + \mathcal{O}(\Delta x^2) \quad \text{and} \quad \tilde{w}^+_{i+\frac{1}{2}} = w(x_{i+\frac{1}{2}}, t^n) + \mathcal{O}(\Delta x^2).$$

Assuming a Lipschitz-continuous numerical flux function such that the consistency condition (2.2) holds, we have

$$f^{\pm}_{i+\frac{1}{2}} = f(w(x_{i+\frac{1}{2}}, t^n)) + \mathcal{O}(\Delta x^2),$$

and we get $\mathcal{F}_{i+\frac{1}{2}} = \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2)$.

Next, we study the accuracy of the source term discretization. By definition of the source term reconstruction (3.4), we have

$$\tilde{S}^{\pm}_i = S^{\pm}_i + \mathcal{O}(\Delta x^2),$$

and arguing (2.6) immediately yields $\mathcal{S}_i = \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta t)$.

Arguing Theorem 1, the second-order space accuracy is thus established.

(ii) Concerning the preservation of the steady states, as soon as $(w^n_i)_{i\in\mathbb{Z}}$ satisfy (2.3), we easily get $\varepsilon^n_{i+\frac{1}{2}} = 0$ for all $i$ in $\mathbb{Z}$. As a consequence, we have $\theta^n_{i+\frac{1}{2}} = 0$, which leads to

$$\tilde{w}^-_{i+\frac{1}{2}} = w^n_i \quad \text{and} \quad \tilde{w}^+_{i+\frac{1}{2}} = w^n_{i+1},$$

while $\tilde{S}^{\pm}_i = \frac{1}{2}(S^n_{i-\frac{1}{2}} + S^n_{i+\frac{1}{2}})$. Put in other words, the reconstruction vanishes for steady states. Then, the original well-balanced first-order scheme (2.1) is recovered and the steady states are preserved.

(iii) We finally turn to the robustness of the improved reconstructed states $\tilde{w}^{\pm}_{i+\frac{1}{2}}$. We remark that

$$\tilde{w}^-_{i+\frac{1}{2}} = (1-\theta^n_{i+\frac{1}{2}})w^n_i + \theta^n_{i+\frac{1}{2}} w^-_{i+\frac{1}{2}} \quad \text{and} \quad \tilde{w}^+_{i+\frac{1}{2}} = (1-\theta^n_{i+\frac{1}{2}})w^n_{i+1} + \theta^n_{i+\frac{1}{2}} w^+_{i+\frac{1}{2}},$$

where $\theta^n_{i+\frac{1}{2}}$, defined by (3.2), belongs to $[0,1]$, and where $w^{\pm}_{i+\frac{1}{2}}$ are given by the initial reconstruction (2.8). Since the states $w^n_i$, $w^n_{i+1}$ and $w^{\pm}_{i+\frac{1}{2}}$ belong to $\Omega$, the states $\tilde{w}^{\pm}_{i+\frac{1}{2}}$ turn out to be convex combinations of states in $\Omega$. With $\Omega$ a convex set, we immediately deduce that $\tilde{w}^{\pm}_{i+\frac{1}{2}}$ are in $\Omega$.

The proof is thus completed. □

To conclude this section, we emphasize that we have designed a well-balanced reconstruction procedure by only evaluating $(G_\ell(w^n_i, x_i))_{1\le\ell\le L}$ and never solving some nonlinear equations. Moreover, the introduced procedure simultaneously deals with all the involved families of steady states and it is not necessary to give more importance to one than to another.

# 4 Well-balanced high-order finite volume extension

The above well-balanced improvement for the second-order MUSCL scheme is easily extended to yield a well-balanced and high-order accurate scheme. Note that in the high-order case, particular attention should be paid to the initialization. Indeed, on the one hand, using usual high-order initialization with cell averages will render the scheme unable to exactly preserve steady states given pointwise by (1.3). On the other hand, using pointwise initialization will cap the scheme to second-order accuracy. Instead, we propose the following way to compute the discrete initial condition in cell $i$:

$$w^0_i = (1-\theta_i)w^0(x_i) + \theta_i \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} w^0(x)\,dx, \tag{4.1}$$

where $\theta_i = \max(\theta_{i-1/2}, \theta_{i+1/2})$. This initial condition is given by a cell average far from a steady state and by pointwise values close to a steady state, which lifts both the issues mentioned above.

**Remark 3.** *We remark that, in some situations, even this initialization procedure may fail to be high-order accurate and accurate. For instance, consider an initial condition made of a perturbation of a steady solution. In some cell $i$ where the perturbation starts, one interface (for instance $x_{i-\frac{1}{2}}$) corresponds to a steady solution, while the other one (for instance $x_{i+\frac{1}{2}}$) corresponds to the unsteady perturbation. If $\theta_i = 0$, then the high-order approximation of the perturbation at interface $x_{i+\frac{1}{2}}$ is lost; otherwise, the steady solution at interface $x_{i-\frac{1}{2}}$ is lost. Also, if $\theta_i = 0$, the second-order accurate zone will extend by one cell every time iteration. This means that, due to this failure in initialization that we have, so far, been unable to correct, the well-balanced scheme may be limited to second-order accuracy in experiments consisting in perturbed steady solutions. This situation is reflected in the upcoming Theorem 4 and in the numerical experiments from Section 5.2.4.*

Now, to build the well-balanced and high-order accurate scheme, let us first introduce a high-order reconstruction according to existing work, see for instance [22, 23]. With $w(x,t)$ a given smooth function, we define $w_i^n$ by adopting (3.7). Now, we consider the following polynomial reconstruction of degree $d$ in space:

$$p_w^n(x;i) = w_i^n + \pi_i^w(x - x_i), \tag{4.2}$$

where $\pi_i^w$ is a polynomial function of degree $d$ such that, for all $x \in (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, we have

$$p_w^n(x;i) = w(x,t^n) + \mathcal{O}(\Delta x^{d+1}) \quad \text{and} \quad \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} p_w^n(x;i)\,dx = w_i^n. \tag{4.3}$$

Equipped with this reconstruction of degree $d$, a scheme of space order $\delta = d+1$ is derived. The reader is referred to [22, 23] where such reconstruction techniques are derived.

From this high-order reconstruction, we now give the associated high-order well-balanced scheme to approximate the weak solutions of (1.1) as follows:

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}\left(f^{\pm}_{i+\frac{1}{2}} - f^{\pm}_{i-\frac{1}{2}}\right) + \Delta t \bar{S}_i^{\pm}, \tag{4.4}$$

with the numerical flux function given by (3.6), with

$$w^-_{i+\frac{1}{2}} = (\tilde{p}_w^n)_i^+ \quad \text{and} \quad w^+_{i+\frac{1}{2}} = (\tilde{p}_w^n)_{i+1}^-, \tag{4.5}$$

where $(\tilde{p}_w^n)_i^{\pm}$ is the following well-balanced modification of the high-order polynomial reconstruction (4.2) evaluated at the interface point $x_{i\pm\frac{1}{2}}$:

$$(\tilde{p}_w^n)_i^{\pm} = w_i^n + \theta^n_{i\pm\frac{1}{2}}\pi_i^w\left(\pm\frac{\Delta x}{2}\right), \tag{4.6}$$

with $\theta^n_{i+\frac{1}{2}}$ defined by (3.2) – (3.3). In (3.2), the parameter $k$ must be fixed larger than $\delta = d+1$ in order to preserve the order $\delta$ of the polynomial reconstruction. Concerning the source term approximation, we start with an approximation of order $\delta$ of the source term average, as follows:

$$S_i^{\pm} = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(p_w^n(x;i),x)\,dx + \mathcal{O}(\Delta x^{\delta}). \tag{4.7}$$

In practice, this approximation is nothing but a quadrature formula of order $\delta$, see for instance [1]. We then adopt the following well-balanced modification of this approximation:

$$\bar{S}_i^{\pm} = \frac{1}{2}\left(\left(1-\theta^n_{i-\frac{1}{2}}\right)S^n_{i-\frac{1}{2}} + \left(1-\theta^n_{i+\frac{1}{2}}\right)S^n_{i+\frac{1}{2}}\right) + \frac{1}{2}\left(\theta^n_{i-\frac{1}{2}} + \theta^n_{i+\frac{1}{2}}\right)S_i^{\pm}. \tag{4.8}$$

At this level, it is worth noting that the $\delta$-order numerical scheme designed here is obtained arguing a very easy modification (4.6) and (4.8) of any polynomial reconstruction (4.2) of degree $d$ and any source term integration (4.7) of order $\delta$. However, this minor correction ensures that the scheme is well-balanced, is of order $\delta$ in space, and preserves the set of admissible states as soon as the original high-order scheme does.

**Theorem 4.** *Given a well-balanced first-order scheme, the scheme (4.4), with the reconstructed states given by (4.5) and the source term approximation (4.8), satisfies the following properties:*

*(i) if $w_i^n$ is an approximation of $w(x_i, t^n)$ up to $\mathcal{O}(\Delta x^\delta)$ for all $i$, then the scheme (4.4) is of order $\delta = d+1$ in space;*

*(ii) if $(w_i^n)_{i\in\mathbb{Z}}$ defines a steady state according to (2.3), then the scheme (4.4) is well-balanced, i.e. $w_i^{n+1} = w_i^n$ for all $i \in \mathbb{Z}$;*

*(iii) the scheme (4.4) is robust, i.e. if the original reconstruction (4.2) of degree $d$ preserves $\Omega$, then $\Omega$ remains invariant by the well-balanced improvement of the reconstruction (4.6).*

We stress that, according to Theorem 3, properties (i) and (ii) cannot be satisfied for perturbed steady states unless the initial condition is computed using (4.1) with, respectively, $\theta_i = 1$ for all $i$ and $\theta_i = 0$ for all $i$.

*Proof of Theorem 4.* We establish properties (i), (ii) and (iii) in order.

(i) We first establish the order of accuracy, as defined in Theorem 1. To address such an issue, we consider $w(x,t)$ a smooth unsteady solution of (1.1) so that the relation (3.9) holds. Next, with $(w_i^n)_{i\in\mathbb{Z}}$ given by (3.7) and $(w_i^{n+1})_{i\in\mathbb{Z}}$ given by the high-order scheme (4.4), the relation (3.8) holds for

$$\mathcal{F}_{i+\frac12} = f^\pm_{i+\frac12} - \frac{1}{\Delta t}\int_0^{\Delta t} f(w(x_{i+\frac12}, t^n + t))\,dt,$$
$$\mathcal{S}_i = \bar{S}_i^\pm - \frac{1}{\Delta t \Delta x}\int_0^{\Delta t}\int_{x_{i-\frac12}}^{x_{i+\frac12}} S(w(x, t^n + t), x)\,dx\,dt,$$

where $\bar{S}_i^\pm$ is defined by (4.8) and $f^\pm_{i+\frac12}$ by (3.6), with the high-order reconstructed states $w^\pm_{i+\frac12}$ given by (4.5).

Next, we establish the order of accuracy associated with the numerical flux function. First, because of the definition (3.2) of $\theta^n_{i+\frac12}$, as long as $\varepsilon^n_{i+\frac12}$ does not vanish, a Taylor expansion yields

$$\theta^n_{i+\frac12} = 1 + \mathcal{O}(\Delta x^k), \qquad \text{with } k \geq \delta = d+1.$$

As a consequence, in the current unsteady context, by definition of the polynomial reconstruction according to (4.3), we obtain

$$w^\pm_{i+\frac12} = w(x_{i+\frac12}, t^n) + \mathcal{O}(\Delta x^\delta).$$

Next, from (2.2), we know that the numerical flux function is Lipschitz-continuous and consistent. Therefore,

$$f_\Delta(w^-_{i+\frac12}, w^+_{i+\frac12}) = f(w(x_{i+\frac12}, t^n)) + \mathcal{O}(\Delta x^\delta),$$

and we get $\mathcal{F}_{i+\frac12} = \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^\delta)$.

Concerning the order of accuracy of the source term, since (4.8) reduces to $\bar{S}_i^\pm = S_i^\pm + \mathcal{O}(\Delta x^\delta)$, arguing (4.7) yields

$$\bar{S}_i^\pm = \frac{1}{\Delta x}\int_{x_{i-\frac12}}^{x_{i+\frac12}} S(w(x, t^n), x)\,dx + \mathcal{O}(\Delta x^\delta). \tag{4.9}$$

Plugging (4.9) into the definition of $\mathcal{S}_i$, we get $\mathcal{S}_i = \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^\delta)$.

The establishment of the order of accuracy is thus completed.

(ii) For the proof of the well-balancedness property, let us consider $(w_i^n)_{i\in\mathbb{Z}}$ to define a steady state according to (2.3). By definition of the correction, given by (3.2) – (3.3), we easily obtain $\theta^n_{i+\frac{1}{2}} = 0$ for all $i$ in $\mathbb{Z}$ so that the reconstructed states now read

$$w^-_{i+\frac{1}{2}} = w^n_i \quad \text{and} \quad w^+_{i+\frac{1}{2}} = w^n_{i+1}.$$

Similarly, regarding the source term reconstruction given by (4.8), we now have $\bar{S}^{\pm}_i = \frac{1}{2}(S^n_{i-\frac{1}{2}} + S^n_{i+\frac{1}{2}})$. As a consequence, the high-order scheme (4.4) coincides with the first-order well-balanced scheme (2.1), and the preservation of the steady states immediately follows.

(iii) To conclude the proof, we now establish that the improvement (4.6) preserves the convex set $\Omega$ as long as the original polynomial reconstruction (4.2) preserves $\Omega$. Indeed, we have

$$(\tilde{p}^n_w)^{\pm}_i = \left(1 - \theta^n_{i\pm\frac{1}{2}}\right) w^n_i + \theta^n_{i\pm\frac{1}{2}} p^n_w\left(x \pm \frac{\Delta x}{2}; i\right).$$

Since $p^n_w(x;i) \in \Omega$ for all $x \in (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, $w^n_i \in \Omega$ and $\theta^n_i \in [0,1]$, we immediately get $(\tilde{p}^n_w)^{\pm}_i \in \Omega$.

The proof is thus completed. □

# 5 Numerical experiments

To assess the performance of the scheme developed above, we now perform several numerical experiments. First, we describe in Section 5.1 the setup used to assess both the order of accuracy and the well-balancedness property of the schemes under consideration. Then, we apply the high-order well-balanced strategy to several systems, namely the shallow water equations with topography (1.4) in Section 5.2, the shallow water equations with friction (1.9) in Section 5.3, and the Euler equations with gravity (1.11) in Section 5.4. We also include, in Section 5.2.4, a discussion around the issues encountered with initialization.

## 5.1 Setup

To justify the relevance of the procedure outlined in Sections 3 and 4, we wish to compare the results of a given first-order well-balanced scheme to the ones produced by its second-order and high-order extensions, with and without the well-balancedness correction. For the sake of simplicity, we introduce the following notations:

- the $\mathbb{P}_d$ scheme is the scheme of order $d+1$ without the well-balancedness correction,
- the $\mathbb{P}^{\mathrm{WB}}_d$ scheme is the scheme of order $d+1$ with the well-balancedness correction.

Note that the $\mathbb{P}_0$ and $\mathbb{P}^{\mathrm{WB}}_0$ schemes are identical. Furthermore, note that forcing $\theta = 1$ on the whole space-time domain in the $\mathbb{P}^{\mathrm{WB}}_d$ scheme is enough to yield the $\mathbb{P}_d$ scheme. In this paper, we consider high-order schemes up to a third-order accurate $\mathbb{P}_2$ scheme. This is enough to justify both the high-order accuracy and the steady state preservation.

To use the $\mathbb{P}_d$ scheme, we need to define three elements: the polynomial reconstruction from (4.2), the approximation of the source term average from (4.7), and the time integration. These elements are summarized in Table 5.1.

Moreover, recall that the $\mathbb{P}^{\mathrm{WB}}_d$ scheme is defined up to the choice of $C^n_{i+\frac{1}{2}}$ in the definition (3.2) of $\theta^n_{i+\frac{1}{2}}$. From this definition, one notes that larger values of $C^n_{i+\frac{1}{2}}$ increase $\theta^n_{i+\frac{1}{2}}$, while smaller values of $C^n_{i+\frac{1}{2}}$ decrease $\theta^n_{i+\frac{1}{2}}$. Although Theorem 4 holds for any expression of $C^n_{i+\frac{1}{2}}$, this choice will impact numerical experiments. Indeed, if $C^n_{i+\frac{1}{2}}$ is too large, then $\theta^n_{i+\frac{1}{2}}$ will also be too large, and steady states that should be captured will not be captured. Similarly, if $C^n_{i+\frac{1}{2}}$ is too small, $\theta^n_{i+\frac{1}{2}}$ will also be too small, and steady states

| | polynomial reconstruction | source term average | time integration |
|---|---|---|---|
| $\mathbb{P}_1$ scheme | MUSCL [45] | midpoint method | SSPRK2 [28, 29] |
| $\mathbb{P}_2$ scheme | third-order [42] | Simpson's method | SSPRK3 [28, 29] |

Table 5.1: Polynomial reconstruction from (4.2), source term average from (4.7), and time integrator for the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes.

that should not be captured will be captured. Basically, taking a very large $C^n_{i+\frac{1}{2}}$ means that the scheme is never well-balanced, while taking a very small $C^n_{i+\frac{1}{2}}$ means that the scheme is never high-order accurate.

Investigating several expressions of $C^n_{i+\frac{1}{2}}$ led us to choose the numerical time derivative of the solution, since it is large when the solution is unsteady (and we want a large $\theta^n_{i+\frac{1}{2}}$), and small when the solution is steady (and we want a small $\theta^n_{i+\frac{1}{2}}$). As a consequence, it is the most generic choice we found that provided good numerical approximations for each system under consideration.

To that end, we define, for $n \geq 1$,

$$C^n_{i+\frac{1}{2}} = C_\theta \, \frac{1}{2} \left( \frac{\left\| w^n_{i+1} - w^{n-1}_{i+1} \right\|}{\Delta t} + \frac{\left\| w^n_i - w^{n-1}_i \right\|}{\Delta t} \right),$$

with $C_\theta$ a constant parameter, which can be interpreted as a normalization of the time derivative. The choice of $C_\theta$ depends on the numerical experiment under consideration (unless otherwise mentioned, we take $C_\theta = 1$). We also take $C^0_{i+\frac{1}{2}} = 1$.

Note that, equipped with this definition of $C^n_{i+\frac{1}{2}}$, the expression (3.2) of $\theta^n_{i+\frac{1}{2}}$ reads:

$$\theta^n_{i+\frac{1}{2}} = \frac{\varepsilon^n_{i+\frac{1}{2}} (C^n_{i+\frac{1}{2}})^k}{\varepsilon^n_{i+\frac{1}{2}} (C^n_{i+\frac{1}{2}})^k + \Delta x^k}.$$

Therefore, we get $\theta^n_{i+\frac{1}{2}} = 0$ if $\varepsilon^n_{i+\frac{1}{2}} = 0$ or if $C^n_{i+\frac{1}{2}} = 0$. This is justified in each case, as follows.

- If $\varepsilon^n_{i+\frac{1}{2}} = 0$, then a steady solution of the equations has been reached, and the first-order well-balanced scheme should be used to ensure the preservation of this solution. Taking $\theta^n_{i+\frac{1}{2}} = 0$ enables this behavior.

- If $C^n_{i+\frac{1}{2}} = 0$, then a local steady solution of the $\mathbb{P}_d$ scheme has been reached. Regardless of whether $\varepsilon^n_{i+\frac{1}{2}} = 0$, we should get $\theta^n_{i+\frac{1}{2}} = 0$ in this case, since a steady solution for the $\mathbb{P}_d$ scheme will not, in general, be a steady solution for the equations. Setting $\theta^n_{i+\frac{1}{2}} = 0$ for such cases perturbs the steady numerical solution and allows it to converge towards the true steady solution.

In practice, to avoid round-off errors, we take $\theta^n_{i+\frac{1}{2}} = 0$ as soon as the expression above yields a value smaller than $10^{-12}$.

This new expression of $\theta^n_{i+\frac{1}{2}}$ gives a partial answer to the additional question of the interpretation of intermediate values of $\theta^n_{i+\frac{1}{2}}$. Take $\theta^n_{i+\frac{1}{2}} = \nu$, with $0 < \nu < 1$ some constant independent from $\Delta x$. In this case, we get $\varepsilon^n_{i+\frac{1}{2}} (C^n_{i+\frac{1}{2}})^k = \mathcal{O}(\Delta x^k)$. Since both quantities $\varepsilon^n_{i+\frac{1}{2}}$ and $C^n_{i+\frac{1}{2}}$ measure an error to a steady state, then at least one of these quantities is small, and the numerical solution is $\Delta x^k$-close to a steady state. Thus, heuristically, the first-order scheme is at least as accurate as the high-order scheme, since it is exact on steady states and the numerical solution is $\Delta x^k$-close to a steady state.

We shall perform three experiments for each system, in order to validate both the high-order accuracy and the well-balancedness property. These experiments are detailed below; system-specific parameters (such as the final physical time, for instance) will be given in the relevant sections.

In each experiment, the space domain is $(0,1)$ and the simulation is run until some final time $t_{\text{end}}$. Each experiment relies on the following compactly supported $\mathcal{C}^\infty$ bump function:

$$\omega(x) = \begin{cases} \exp\left(1 - \dfrac{1}{1 - \left(4\left(x - \dfrac{1}{2}\right)\right)^2}\right) & \text{if } \left|x - \dfrac{1}{2}\right| < \dfrac{1}{4}, \\ 0 & \text{otherwise.} \end{cases}$$

In addition, unless otherwise mentioned, all errors computed in the remainder of the text are $L^2$ errors between the approximate solution and the exact or reference solution.

The first experiment we perform yields a measure of the order of accuracy of the schemes, and it is designed to show that the well-balancedness correction does not reduce the accuracy for unsteady solutions. To correctly measure the order of accuracy, no slope limitation is added to the $\mathbb{P}_d$ and $\mathbb{P}_d^{\text{WB}}$ schemes. Since we do not necessarily know an exact solution of the system under consideration, we compute a reference solution using a very fine grid made of $20 \times 2^{12}$ cells. Then, after computing the approximate solution on a coarser dyadic grid made of $N = 20 \times 2^k$ cells, $0 \leq k < 12$, the fine solution is projected onto the coarser grid to measure the error between the reference solution and its approximation. To ensure that no shock waves form, the initial condition is smooth; its expression is given for each system. The initial condition is then evolved until the final time $t_{\text{end}} = 5 \cdot 10^{-3}$. Periodic boundary conditions are prescribed for this experiment.

The second and third experiments assess the well-balancedness property. To that end, we study the dissipation of a perturbation applied to an initially steady solution. Here, we add a slope limitation to the $\mathbb{P}_d$ and $\mathbb{P}_d^{\text{WB}}$ schemes, namely the MC limiter from [32] for $d = 1$ and the limiter from [42] for $d = 2$. The initial condition is a steady solution $w$, computed by solving the nonlinear system (1.3). This steady solution is then perturbed using the bump function $\omega$. Namely, each variable in $w$ is multiplied by $(1 + \alpha\omega(x))$. We distinguish two cases: a small and a large perturbation. For the small perturbation, we take $\alpha = \beta\Delta x^3$, with $\beta$ a constant given for each system. For the large perturbation, we take $\alpha = 0.25$. The steady solution is imposed on the boundaries, and the experiment is run until the numerical solution becomes steady; this final time is given for each system. We take 50 discretization cells for this experiment.

## 5.2 Application: the shallow water equations with topography

The first application concerns the shallow water equations with topography (1.4). The first-order well-balanced scheme comes from [37, 10]. It contains a parameter $C$, set here to $+\infty$, or rather to the upper bound of the double-precision floating-point numbers in practice. In addition, the topography function is set to $z(x) = \omega(x)$ and we take $g = 9.81$.

### 5.2.1 Order of accuracy assessment

For this experiment, the initial condition is given by

$$h_0(x) = 2 - z(x) + \cos^2(2\pi x) \quad \text{and} \quad q_0(x) = \sin(2\pi x).$$

In Figure 5.1, we display the reference solution and the approximations given by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\text{WB}}$ schemes with 40 discretization cells. We observe that the third-order schemes are very close to the reference solution, even with such a few cells.

This observation is confirmed in Table 5.2 and Figure 5.2, where we report the errors on $h$ and $q$, as well as the orders of accuracy. As expected, the well-balancedness procedure does not alter the order of accuracy of the scheme, since the solution produced by the $\mathbb{P}_2^{\text{WB}}$ scheme is almost the same as the one produced by the $\mathbb{P}_2$ scheme in this unsteady context. We even observe a slight over-convergence, possibly explained by the use of the fourth-order accurate Simpson's method in the source term approximation.

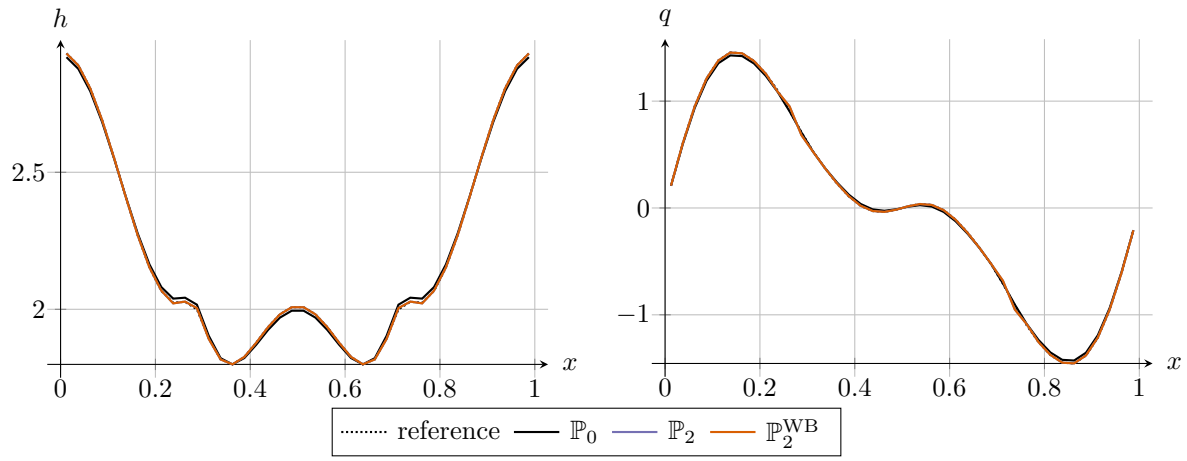


Figure 5.1: Shallow water equations with topography: comparison between the reference and approximate solutions for the dyadic experiment with 40 discretization cells, at time $t_{\mathrm{end}} = 5 \cdot 10^{-3}$. Left panel: water height $h$; right panel: discharge $q$.

| $N$ | error, $\mathbb{P}_0$ | order, $\mathbb{P}_0$ | error, $\mathbb{P}_2$ | order, $\mathbb{P}_2$ | error, $\mathbb{P}_2^{\mathrm{WB}}$ | order, $\mathbb{P}_2^{\mathrm{WB}}$ |
|---|---|---|---|---|---|---|
| 40 | $1.04 \cdot 10^{-2}$ | — | $1.12 \cdot 10^{-3}$ | — | $1.12 \cdot 10^{-3}$ | — |
| 80 | $5.24 \cdot 10^{-3}$ | 0.99 | $3.25 \cdot 10^{-4}$ | 1.78 | $3.26 \cdot 10^{-4}$ | 1.78 |
| 160 | $2.58 \cdot 10^{-3}$ | 1.02 | $4.06 \cdot 10^{-5}$ | 3.00 | $4.08 \cdot 10^{-5}$ | 3.00 |
| 320 | $1.29 \cdot 10^{-3}$ | 1.00 | $2.74 \cdot 10^{-6}$ | 3.89 | $2.73 \cdot 10^{-6}$ | 3.90 |
| 640 | $6.42 \cdot 10^{-4}$ | 1.00 | $1.76 \cdot 10^{-7}$ | 3.96 | $1.76 \cdot 10^{-7}$ | 3.96 |
| 1280 | $3.21 \cdot 10^{-4}$ | 1.00 | $1.34 \cdot 10^{-8}$ | 3.72 | $1.36 \cdot 10^{-8}$ | 3.69 |
| 2560 | $1.60 \cdot 10^{-4}$ | 1.00 | $1.50 \cdot 10^{-9}$ | 3.16 | $1.54 \cdot 10^{-9}$ | 3.14 |

Table 5.2: Shallow water equations with topography: errors and orders of accuracy for the dyadic experiment. For the sake of conciseness, only the errors on $h$ are reported in this table; the reader is referred to Figure 5.2 for a visualization of the errors on $q$.

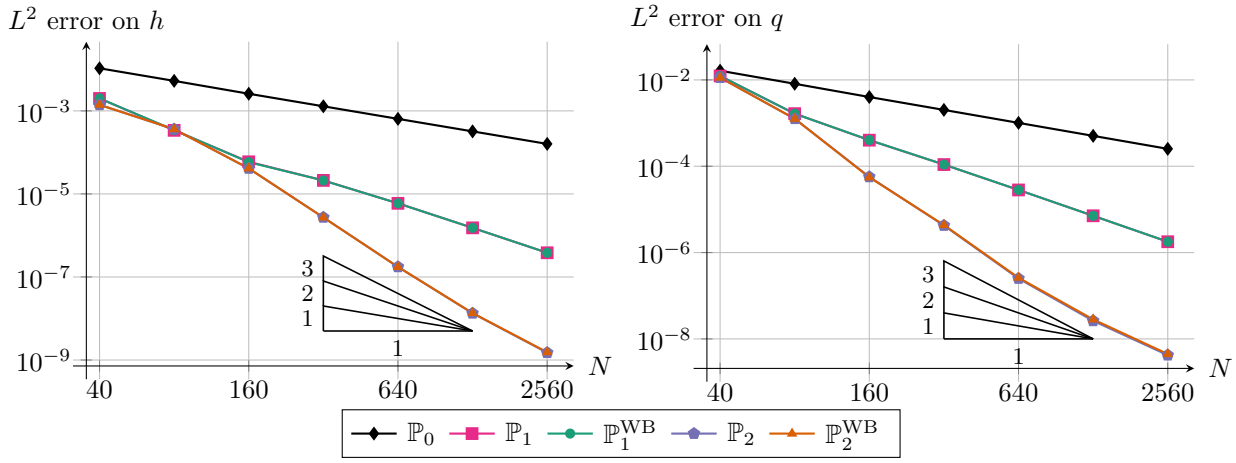


Figure 5.2: Shallow water equations with topography: error lines for the dyadic experiment. Left panel: error on $h$; right panel: error on $q$.

### 5.2.2 Well-balancedness property – large perturbation

The initial condition is a large perturbation of the steady solution implicitly given by (1.5), with

$$\begin{cases} q = 1, \\ \dfrac{q^2}{2h^2} + g(h + Z) = 2. \end{cases}$$

We take $C_\theta = 9 \cdot 10^{-3}$, and the final physical time is $t_{\text{end}} = 20$.

In Figure 5.3, we display the initial condition, as well as the approximations given by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\text{WB}}$ schemes at time $t = 2.5 \cdot 10^{-2}$. We observe that the solutions of the $\mathbb{P}_2$ and $\mathbb{P}_2^{\text{WB}}$ are quite close, even in this case of a perturbed steady solution, and that they are less diffusive than the solution given by the $\mathbb{P}_0$ scheme. To get a more precise error quantification, the errors between the approximate solutions and a reference solution computed with a fine mesh are presented in Table 5.3. We confirm the observations from Figure 5.3, since the well-balanced correction does not degrade the error produced by the corresponding high-order scheme. Note that small oscillations appear around the extrema with the third-order methods, both with and without the well-balanced correction. Such small oscillations are expected in this case, since the limiter from [42] is less viscous than the MC limiter used in the second-order methods.

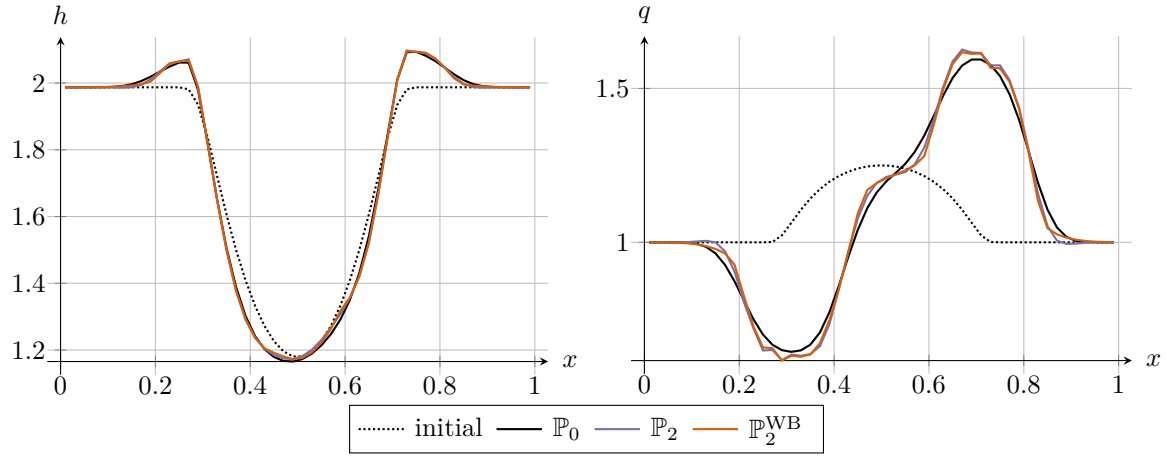


Figure 5.3: Shallow water equations with topography: comparison between the initial condition and the approximate solutions at time $t = 2.5 \cdot 10^{-2}$, for the perturbed steady state experiment with 50 cells. Left panel: water height $h$; right panel: discharge $q$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $2.44 \cdot 10^{-2}$ | $1.53 \cdot 10^{-2}$ | $1.60 \cdot 10^{-2}$ | $8.63 \cdot 10^{-3}$ | $9.46 \cdot 10^{-3}$ |
| error on $q$ | $8.84 \cdot 10^{-2}$ | $5.39 \cdot 10^{-2}$ | $5.44 \cdot 10^{-2}$ | $4.02 \cdot 10^{-2}$ | $4.44 \cdot 10^{-2}$ |

Table 5.3: Shallow water equations with topography: $L^\infty$ errors between the steady solution and the approximate solutions at time $t = 2.5 \cdot 10^{-2}$, for the steady state with a large perturbation, using 50 cells.

Then, in Figure 5.4 and Table 5.4, we report the errors on $h$ and $q$ at the final time $t_{\text{end}}$, as well as the CPU time taken by each method. We observe that the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_2^{\text{WB}}$ schemes have all converged towards the exact steady solution up to machine precision, while a non-zero error remains for the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes. These observations validate the proposed well-balancedness correction. We also note that, as expected, the well-balanced correction of the high-order methods is not computationally costly. Indeed, it corresponds to a 10% increase for the $\mathbb{P}_1^{\text{WB}}$ scheme and a 1% increase for the $\mathbb{P}_2^{\text{WB}}$ scheme. This can be contrasted to usual high-order well-balanced methods, where a costly nonlinear inversion is needed for the reconstruction.

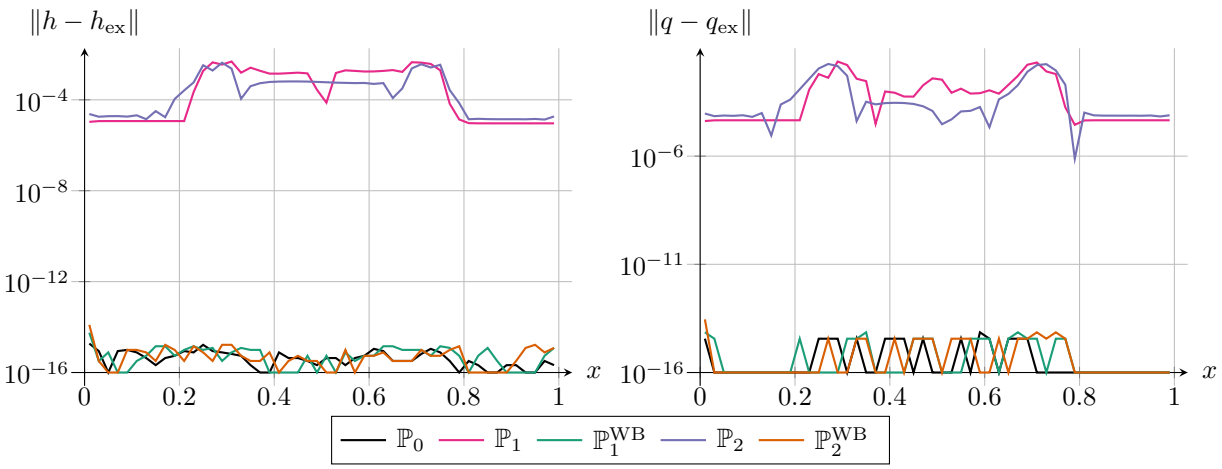


Figure 5.4: Shallow water equations with topography: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a large perturbation, using 50 cells. Left panel: error on $h$; right panel: error on $q$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $5.72 \cdot 10^{-16}$ | $1.85 \cdot 10^{-3}$ | $1.08 \cdot 10^{-15}$ | $1.30 \cdot 10^{-3}$ | $1.91 \cdot 10^{-15}$ |
| error on $q$ | $2.15 \cdot 10^{-15}$ | $6.07 \cdot 10^{-3}$ | $2.64 \cdot 10^{-15}$ | $5.42 \cdot 10^{-3}$ | $4.73 \cdot 10^{-15}$ |
| CPU time (s) | 2.91 | 8.59 | 9.5 | 23.78 | 24 |

Table 5.4: Shallow water equations with topography: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a large perturbation, using 50 cells.

### 5.2.3 Well-balancedness property – small perturbation

The initial condition is computed the same way as in the previous test case, but using a small perturbation of amplitude $\beta\Delta x^3$, with $\beta = 10^{-2}$, instead of a large perturbation. We take $C_\theta = 3.5 \cdot 10^{-2}$, and the final physical time is $t_{\text{end}} = 5 \cdot 10^{-3}$.

In Figure 5.5, we display the errors between the numerical solution and the underlying steady solution at the final time $t_{\text{end}}$. The $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes produce errors that destroy both the underlying steady state and the small perturbation. Indeed, these errors are actually larger than the perturbation, and the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes behave as though there were no perturbation. Conversely, the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_2^{\text{WB}}$ schemes yield comparable results, and the small perturbation is clearly visible against the underlying, exactly preserved steady state. Here, it was expected for the results of these three schemes to be close, since the convex combination procedure correctly switches from the locally less accurate high-order scheme to the locally more accurate well-balanced scheme. These observations are confirmed in Table 5.5, where the errors between the numerical solutions and the underlying steady states are reported. As expected, the well-balanced schemes produce an error of the order of magnitude of the perturbation amplitude (around $10^{-7}$), while the errors made by using the non-well-balanced schemes are several orders of magnitude larger.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $4.28 \cdot 10^{-8}$ | $1.03 \cdot 10^{-3}$ | $4.27 \cdot 10^{-8}$ | $7.70 \cdot 10^{-4}$ | $4.54 \cdot 10^{-8}$ |
| error on $q$ | $4.66 \cdot 10^{-8}$ | $7.94 \cdot 10^{-3}$ | $4.48 \cdot 10^{-8}$ | $6.63 \cdot 10^{-3}$ | $9.46 \cdot 10^{-8}$ |

Table 5.5: Shallow water equations with topography: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells.

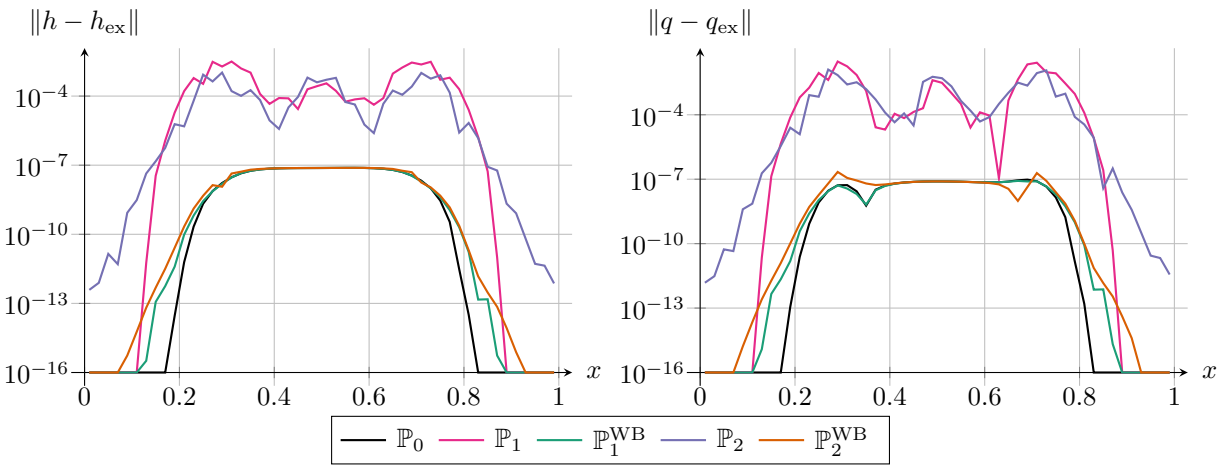


Figure 5.5: Shallow water equations with topography: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells. Left panel: error on $h$; right panel: error on $q$.

### 5.2.4 Issues with initialization

For this last experiment on the shallow water system with topography (1.4), we wish to highlight the specific issue of initialization. discussed in Theorem 3. Indeed, let us underline that the scheme endowed with the correction is formally high-order accurate and exactly well-balanced. However, issues with the initialization procedure (4.1) mean that, on specific cases such as a perturbed steady state, the resulting discretization of the initial condition will either be exactly well-balanced but second-order accurate, or high-order accurate but non-well-balanced. We once again stress that the issues discussed in this paragraph stem from a failure of the initialization procedure, rather than a failure of the well-balanced correction itself.

For this experiment, we consider a large perturbation of a steady state. The initial condition is a perturbed steady solution, but this time the support of the perturbation is $(3/8, 5/8)$ rather than $(1/4, 3/4)$, which means that the steady solution is multiplied by $(1 + \alpha\omega(2x - 1/2))$. We compare the initialization given by (4.1) (called well-balanced initialization) to the high-order initialization, given by integrating the fine reference solution over each cell. To ensure that the solution remains smooth enough, we take $t_{\text{end}} = 10^{-3}$. In addition, we take $C_\theta = 0.5$.

The error lines are displayed in Figure 5.6. In the top left panel, the error on $h$ clearly shows that the approximate solution obtained by the $\mathbb{P}_2$ and $\mathbb{P}_2^{\text{WB}}$ schemes is capped to second order accuracy with the well-balanced initialization (4.1). The bottom left panel shows that using cell averages increases the order of accuracy to 3, as expected. Note that this behavior occurs whether or not the well-balanced correction is implemented: this confirms the fact that this is a failure of the initialization, rather than of the well-balanced correction. In the right panels, the same conclusions are drawn for $q$, although the effect is less visible. We also emphasize that, because the high-order scheme we consider in the experiments is third-order accurate, we only lose one order of accuracy by switching to the well-balanced initialization. For high-order schemes, more orders of accuracy would be lost on such test cases, and the discrepancy shown in Figure 5.6 would increase.

## 5.3 Application: the shallow water equations with Manning friction

The next application concerns the shallow water equations with Manning friction, governed by (1.9). The first-order well-balanced scheme comes from [38]. It also contains a parameter $C$, also set here to $+\infty$. We take the friction exponent $\eta = 7/3$ according to Manning's model [36, 20], and we set $\kappa = 1$ as well as $g = 9.81$.

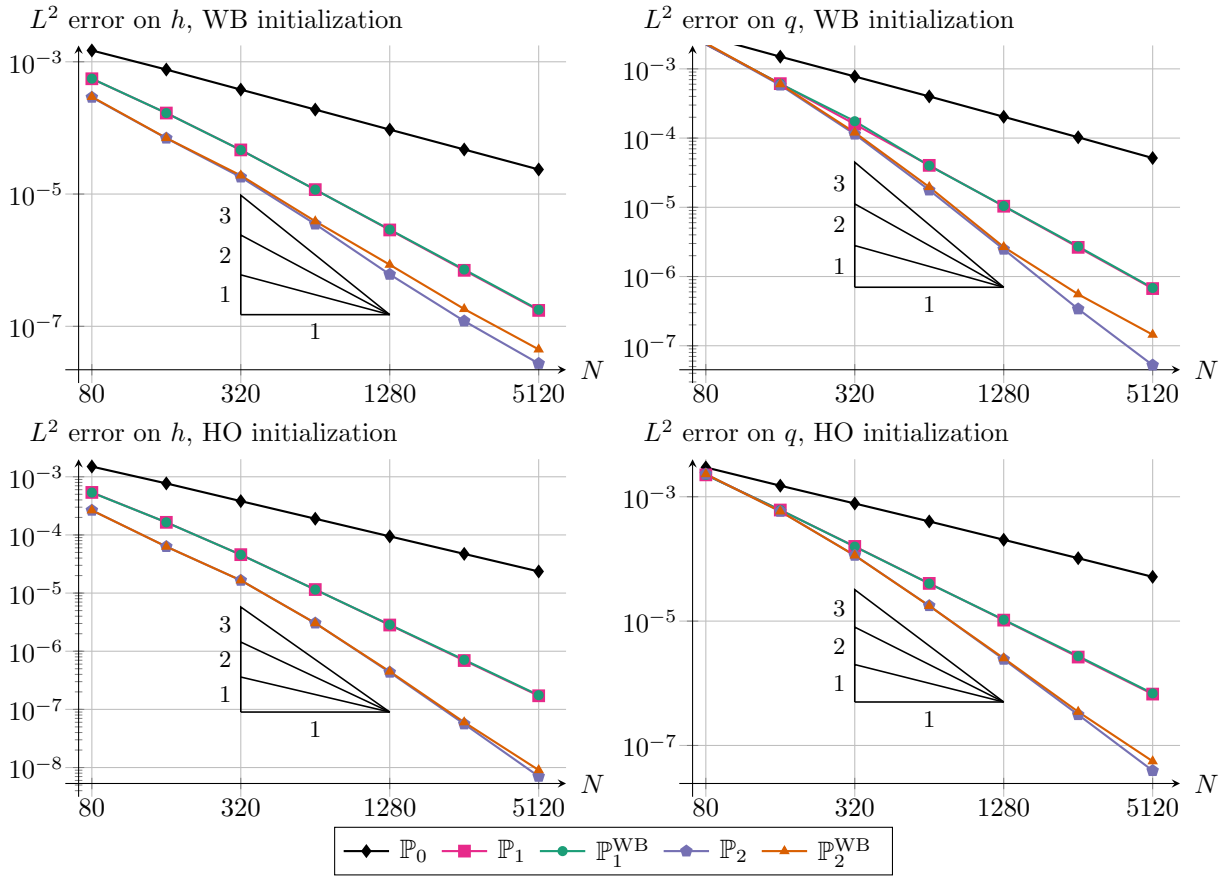


Figure 5.6: Shallow water equations with topography: error lines for the perturbed steady solution. Left panels: error on $h$; right panels: error on $q$. Top panels: well-balanced initialization with pointwise values; bottom panels: high-order accurate initialization with cell averages.

### 5.3.1 Order of accuracy assessment

The initial condition for this experiment is given by

$$h_0(x) = 2 + \cos^2(2\pi x) \quad \text{and} \quad q_0(x) = \sin(2\pi x).$$

In Figure 5.7, we display the reference solution and the approximations given by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ schemes with 40 discretization cells. We once again observe that the third-order schemes are very close to the reference solution.

We report the errors on $h$ and $q$, as well as the orders of accuracy, in Table 5.6 and Figure 5.8. The previous observation is once again confirmed since the $\mathbb{P}_d$ and $\mathbb{P}_d^{\mathrm{WB}}$ schemes produce almost exactly the same solution for this experiment.

### 5.3.2 Well-balancedness property – large perturbation

The initial condition is a perturbation of the steady solution implicitly given by (1.2), with

$$\begin{cases} q = 1, \\ -q^2 \dfrac{h^{\eta-1}}{\eta-1} + g\dfrac{h^{\eta+2}}{\eta+2} + kq|q|x = 3. \end{cases}$$

The final physical time is $t_{\mathrm{end}} = 20$.

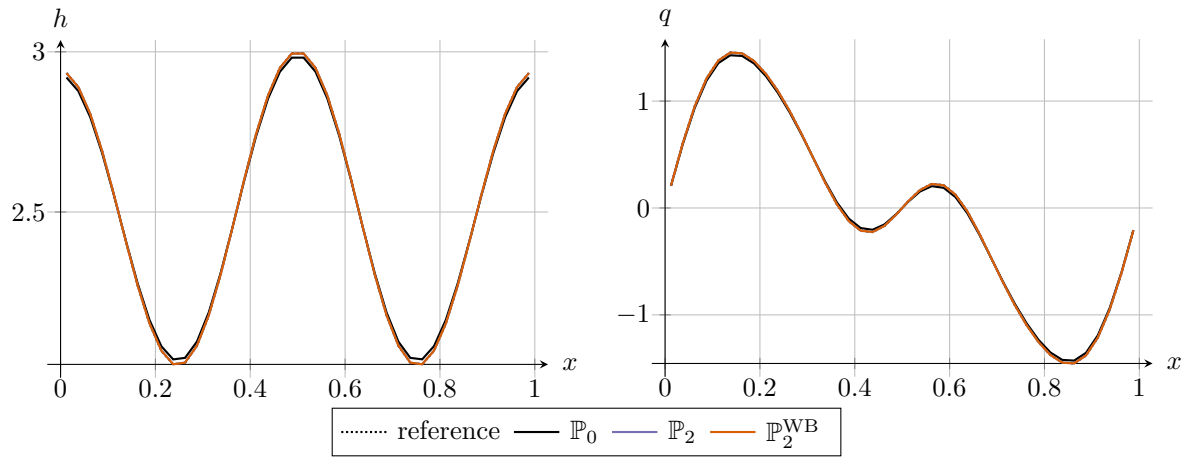


Figure 5.7: Shallow water equations with Manning friction: comparison between the reference and approximate solutions for the dyadic experiment with 40 discretization cells, at time $t_{\mathrm{end}} = 5 \cdot 10^{-3}$. Left panel: water height $h$; right panel: discharge $q$.

| $N$ | error, $\mathbb{P}_0$ | order, $\mathbb{P}_0$ | error, $\mathbb{P}_2$ | order, $\mathbb{P}_2$ | error, $\mathbb{P}_2^{\mathrm{WB}}$ | order, $\mathbb{P}_2^{\mathrm{WB}}$ |
|---|---|---|---|---|---|---|
| 40 | $1.04 \cdot 10^{-2}$ | — | $2.82 \cdot 10^{-4}$ | — | $2.82 \cdot 10^{-4}$ | — |
| 80 | $5.23 \cdot 10^{-3}$ | 1.00 | $3.64 \cdot 10^{-5}$ | 2.95 | $3.64 \cdot 10^{-5}$ | 2.95 |
| 160 | $2.57 \cdot 10^{-3}$ | 1.03 | $4.62 \cdot 10^{-6}$ | 2.98 | $4.62 \cdot 10^{-6}$ | 2.98 |
| 320 | $1.28 \cdot 10^{-3}$ | 1.00 | $5.80 \cdot 10^{-7}$ | 2.99 | $5.80 \cdot 10^{-7}$ | 2.99 |
| 640 | $6.40 \cdot 10^{-4}$ | 1.00 | $7.28 \cdot 10^{-8}$ | 3.00 | $7.28 \cdot 10^{-8}$ | 3.00 |
| 1280 | $3.20 \cdot 10^{-4}$ | 1.00 | $9.11 \cdot 10^{-9}$ | 3.00 | $9.11 \cdot 10^{-9}$ | 3.00 |
| 2560 | $1.60 \cdot 10^{-4}$ | 1.00 | $1.14 \cdot 10^{-9}$ | 3.00 | $1.14 \cdot 10^{-9}$ | 3.00 |

Table 5.6: Shallow water equations with Manning friction: errors and orders of accuracy for the dyadic experiment. For the sake of conciseness, only the errors on $h$ are reported in this table; the reader is referred to Figure 5.8 for a visualization of the errors on $q$.

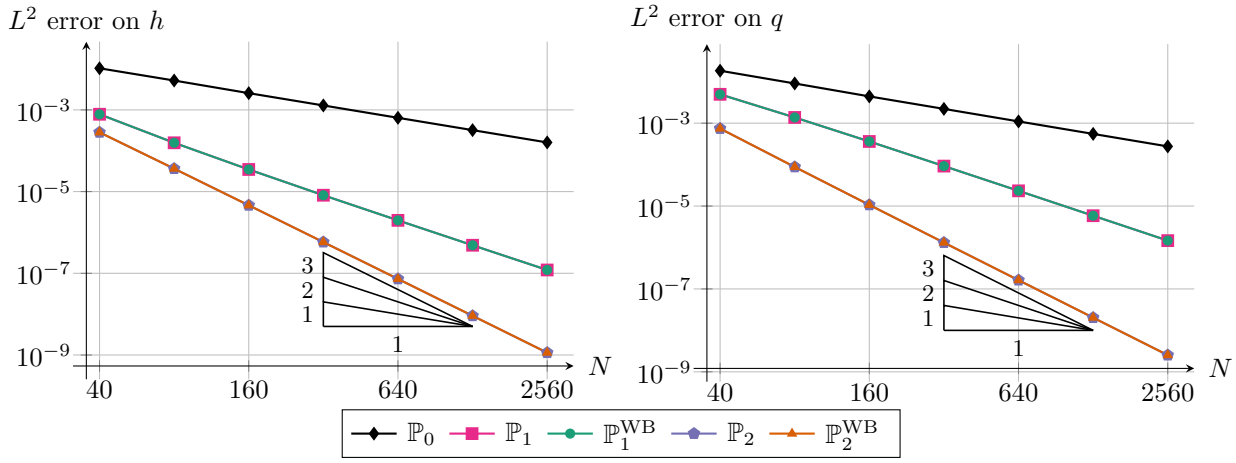


Figure 5.8: Shallow water equations with Manning friction: error lines for the dyadic experiment. Left panel: error on $h$; right panel: error on $q$.

We display the initial condition, as well as the approximations produced by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ schemes at time $t = 2.5 \cdot 10^{-2}$, in Figure 5.9. We observe that the solutions of the $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ are indistinguishable,

and that they are less diffusive than the solution given by the $\mathbb{P}_0$ scheme. These observations are confirmed by the errors reported in Table 5.7.

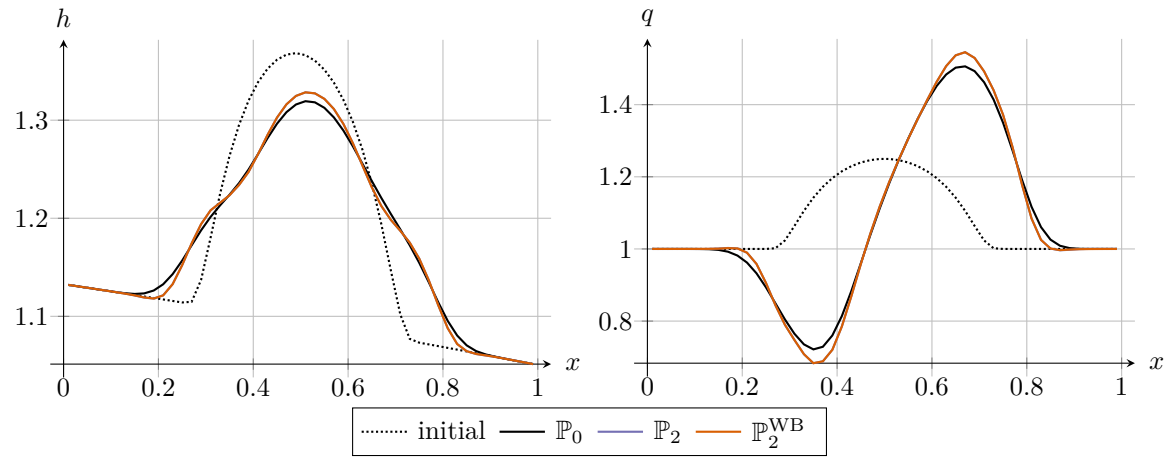


Figure 5.9: Shallow water equations with Manning friction: comparison between the initial condition and the approximate solutions at time $t = 2.5 \cdot 10^{-2}$, for the perturbed steady state experiment with 50 cells. Left panel: water height $h$; right panel: discharge $q$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $1.42 \cdot 10^{-2}$ | $6.21 \cdot 10^{-3}$ | $6.21 \cdot 10^{-3}$ | $3.40 \cdot 10^{-3}$ | $3.40 \cdot 10^{-3}$ |
| error on $q$ | $5.32 \cdot 10^{-2}$ | $2.25 \cdot 10^{-2}$ | $2.25 \cdot 10^{-2}$ | $1.47 \cdot 10^{-2}$ | $1.45 \cdot 10^{-2}$ |

Table 5.7: Shallow water equations with Manning friction: $L^\infty$ errors between the steady solution and the approximate solutions at time $t = 2.5 \cdot 10^{-2}$, for the perturbed steady state experiment with 50 cells.

In Figure 5.10 and Table 5.8, the errors on $h$ and $q$ at the final time $t_{\text{end}}$ are reported alongside the CPU time. Once again, the $\mathbb{P}_0$, $\mathbb{P}_1^{\text{WB}}$ and $\mathbb{P}_2^{\text{WB}}$ schemes have all converged towards the exact steady solution, and the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes produce a non-zero error. Like the previous test case, we note that the well-balanced correction is not costly.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $3.20 \cdot 10^{-15}$ | $4.08 \cdot 10^{-7}$ | $4.44 \cdot 10^{-15}$ | $2.57 \cdot 10^{-9}$ | $3.22 \cdot 10^{-15}$ |
| error on $q$ | $2.35 \cdot 10^{-14}$ | $5.55 \cdot 10^{-6}$ | $3.32 \cdot 10^{-14}$ | $2.46 \cdot 10^{-8}$ | $2.36 \cdot 10^{-14}$ |
| CPU time (s) | 3.54 | 9.77 | 10.96 | 23.84 | 24.87 |

Table 5.8: Shallow water equations with Manning friction: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the perturbed steady state experiment with 50 cells.

### 5.3.3 Well-balancedness property – small perturbation

We consider a small perturbation of a steady solution, with $\beta = 10^{-5}$; the initial condition is computed in a similar fashion as in the large perturbation case. The final physical time is $t_{\text{end}} = 0.005$, and we take $C_\theta = 0.25$.

The errors between the underlying steady solution and the numerical approximation are reported in Figure 5.11. As in the shallow water equations with topography, the large errors made by the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes destroy both the perturbation and the underlying steady state, as confirmed in Table 5.5.

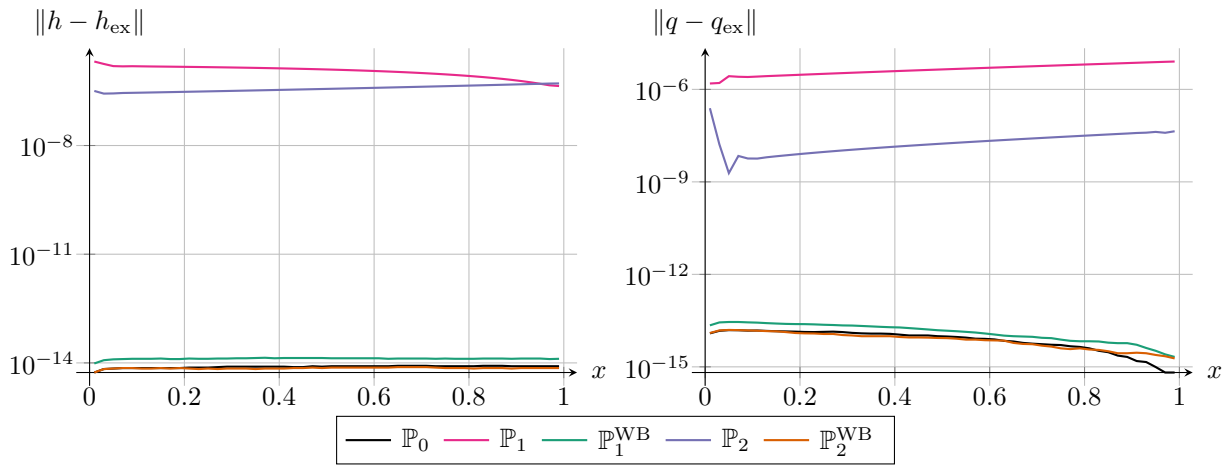


Figure 5.10: Shallow water equations with Manning friction: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the perturbed steady state experiment with 50 cells. Left panel: error on $h$; right panel: error on $q$.

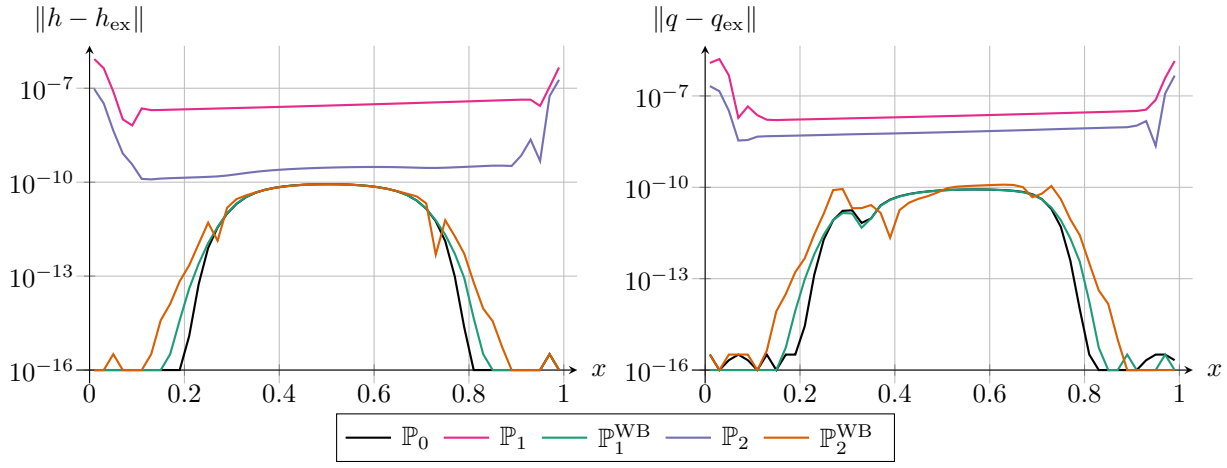


Figure 5.11: Shallow water equations with Manning friction: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells. Left panel: error on $h$; right panel: error on $q$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $h$ | $4.28 \cdot 10^{-11}$ | $1.54 \cdot 10^{-7}$ | $4.27 \cdot 10^{-11}$ | $3.09 \cdot 10^{-8}$ | $4.28 \cdot 10^{-11}$ |
| error on $q$ | $4.25 \cdot 10^{-11}$ | $3.60 \cdot 10^{-7}$ | $4.20 \cdot 10^{-11}$ | $7.70 \cdot 10^{-8}$ | $5.63 \cdot 10^{-11}$ |

Table 5.9: Shallow water equations with Manning friction: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells.

## 5.4 Application: the Euler equations with gravity

For the last application, we turn to another system to highlight the genericity of our method. We choose the Euler equations with gravity (1.11). The first-order well-balanced scheme is based on the strategy from [37, 38, 25]. It is designed to exactly preserve and capture the family of steady states given by (1.12).

We consider an ideal gas pressure law, where the pressure $p$ is given by

$$p = (\gamma - 1)(E - \frac{1}{2}\frac{q^2}{\rho}).$$

Furthermore, we take the parameters $\gamma = 1.4$ and $\kappa = 1$. The gravity potential is given by $\varphi(x) = \omega(x)$.

### 5.4.1 Order of accuracy assessment

For this experiment, we take the following initial condition:

$$\rho_0(x) = 2 + \cos^2(2\pi x), \quad q_0(x) = \sin(2\pi x), \quad E_0(x) = 5 + \cos^2(2\pi x).$$

In Figure 5.12, we display the reference solution and the approximations given by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ schemes with 40 discretization cells. As usual, the third-order schemes are very close to the reference solution.

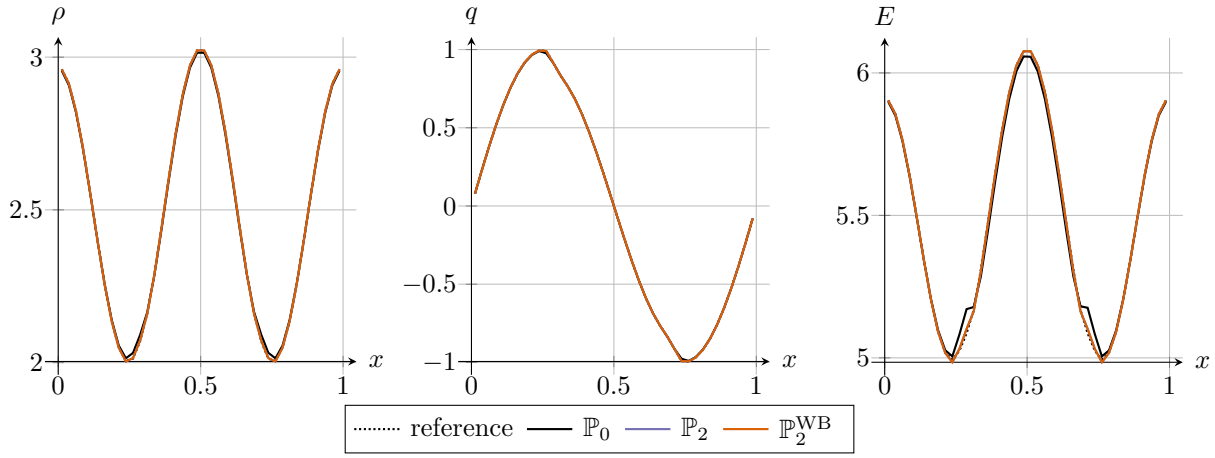


Figure 5.12: Euler equations with gravity: comparison between the reference and approximate solutions for the dyadic experiment with 40 discretization cells, at time $t_{\mathrm{end}} = 5 \cdot 10^{-3}$. Left panel: density $\rho$; middle panel: momentum $q$; right panel: energy $E$.

We also report the errors and the orders of accuracy for $\rho$, $q$ and $E$ in Table 5.10 and Figure 5.13. The same conclusion as for the shallow water system with topography or Manning friction is drawn.

| $N$ | error, $\mathbb{P}_0$ | order, $\mathbb{P}_0$ | error, $\mathbb{P}_2$ | order, $\mathbb{P}_2$ | error, $\mathbb{P}_2^{\mathrm{WB}}$ | order, $\mathbb{P}_2^{\mathrm{WB}}$ |
|---|---|---|---|---|---|---|
| 80 | $4.18 \cdot 10^{-3}$ | — | $2.61 \cdot 10^{-4}$ | — | $2.61 \cdot 10^{-4}$ | — |
| 160 | $2.07 \cdot 10^{-3}$ | 1.01 | $5.59 \cdot 10^{-5}$ | 2.22 | $5.59 \cdot 10^{-5}$ | 2.22 |
| 320 | $1.03 \cdot 10^{-3}$ | 1.01 | $1.09 \cdot 10^{-5}$ | 2.36 | $1.09 \cdot 10^{-5}$ | 2.36 |
| 640 | $5.14 \cdot 10^{-4}$ | 1.00 | $1.72 \cdot 10^{-6}$ | 2.66 | $1.72 \cdot 10^{-6}$ | 2.66 |
| 1280 | $2.56 \cdot 10^{-4}$ | 1.00 | $2.31 \cdot 10^{-7}$ | 2.89 | $2.31 \cdot 10^{-7}$ | 2.89 |
| 2560 | $1.28 \cdot 10^{-4}$ | 1.00 | $2.95 \cdot 10^{-8}$ | 2.97 | $2.95 \cdot 10^{-8}$ | 2.97 |
| 5120 | $6.41 \cdot 10^{-5}$ | 1.00 | $3.70 \cdot 10^{-9}$ | 2.99 | $3.70 \cdot 10^{-9}$ | 2.99 |

Table 5.10: Euler equations with gravity: errors and orders of accuracy for the dyadic experiment. For the sake of conciseness, only the errors on $\rho$ are reported in this table; the reader is referred to Figure 5.8 for a visualization of the errors on $q$ and $E$.

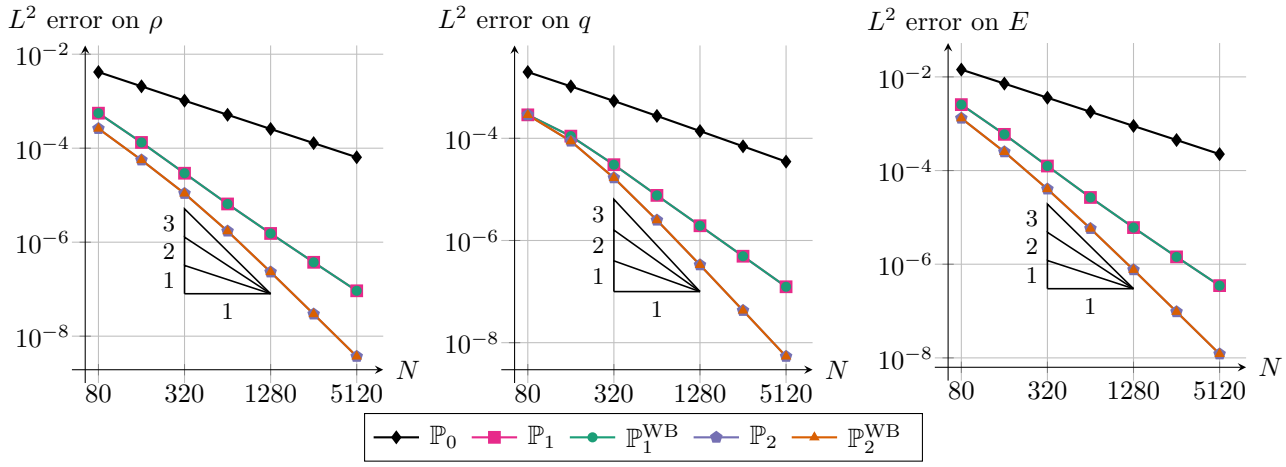


Figure 5.13: Euler equations with gravity: error lines for the dyadic experiment. Left panel: error on $\rho$; middle panel: error on $q$; right panel: error on $E$.

### 5.4.2 Well-balancedness property – large perturbation

The initial condition is a perturbation of the steady solution implicitly given by (1.12), with

$$\begin{cases} q = 0, \\ p - \kappa\rho = 1, \\ \kappa \dfrac{\rho^{\gamma-1}}{\gamma - 1} + \varphi = 5. \end{cases}$$

The final physical time is $t_{\mathrm{end}} = 300$.

We display the initial condition, as well as the approximations produced by the $\mathbb{P}_0$, $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ schemes at time $t = 1 \cdot 10^{-1}$, in Figure 5.14. The solutions of the $\mathbb{P}_2$ and $\mathbb{P}_2^{\mathrm{WB}}$ schemes are once again indistinguishable and less diffusive than the one given by the $\mathbb{P}_0$ scheme. As before, the errors reported in Table 5.11 confirm this observation.

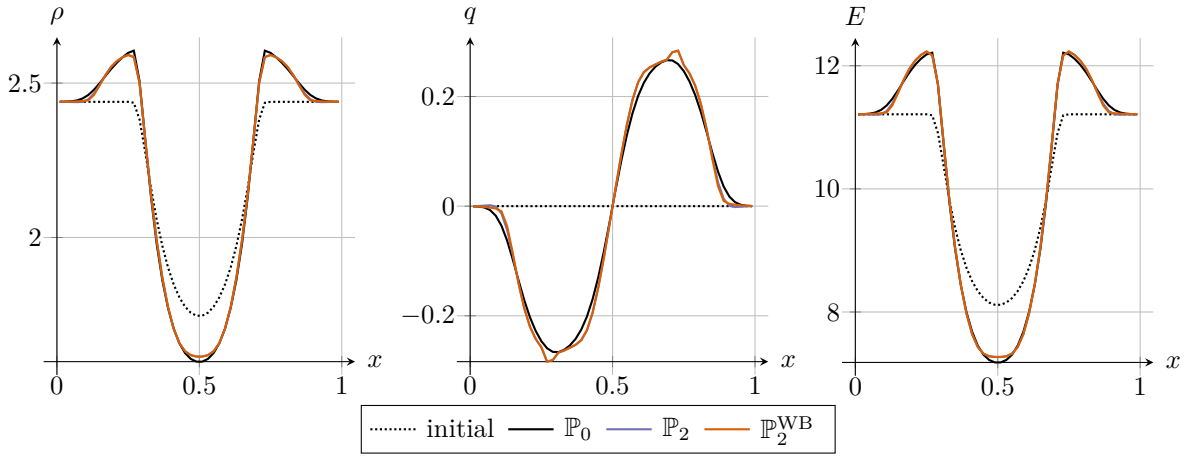


Figure 5.14: Euler equations with gravity: comparison between the initial condition and the approximate solutions at time $t = 1 \cdot 10^{-1}$, for the perturbed steady state experiment with 50 cells. Left panel: density $\rho$; middle panel: momentum $q$; right panel: energy $E$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $\rho$ | $2.02 \cdot 10^{-2}$ | $1.15 \cdot 10^{-2}$ | $1.17 \cdot 10^{-2}$ | $6.65 \cdot 10^{-3}$ | $6.82 \cdot 10^{-3}$ |
| error on $q$ | $3.34 \cdot 10^{-2}$ | $1.76 \cdot 10^{-2}$ | $1.60 \cdot 10^{-2}$ | $1.58 \cdot 10^{-2}$ | $1.56 \cdot 10^{-2}$ |
| error on $E$ | $1.31 \cdot 10^{-1}$ | $7.06 \cdot 10^{-2}$ | $6.43 \cdot 10^{-2}$ | $5.13 \cdot 10^{-2}$ | $5.18 \cdot 10^{-2}$ |

Table 5.11: Euler equations with gravity: $L^\infty$ errors between the steady solution and the approximate solutions at time $t = 1 \cdot 10^{-1}$, for the perturbed steady state experiment with 50 cells.

In Figure 5.15 and Table 5.12, the errors on $\rho$, $q$ and $E$ at the final time $t_{\text{end}}$ are reported, as well as the CPU time. The same conclusions as in the shallow water case are reached. We even observe a decrease in CPU time with the $\mathbb{P}_2^{\text{WB}}$ scheme compared to the $\mathbb{P}_2$ scheme. This is due to the following interesting optimization. When a steady state is reached at interface $x_{i+\frac{1}{2}}$, we get $\theta_{i+\frac{1}{2}} = 0$ and the high-order correction need not be computed, which saves some CPU time.

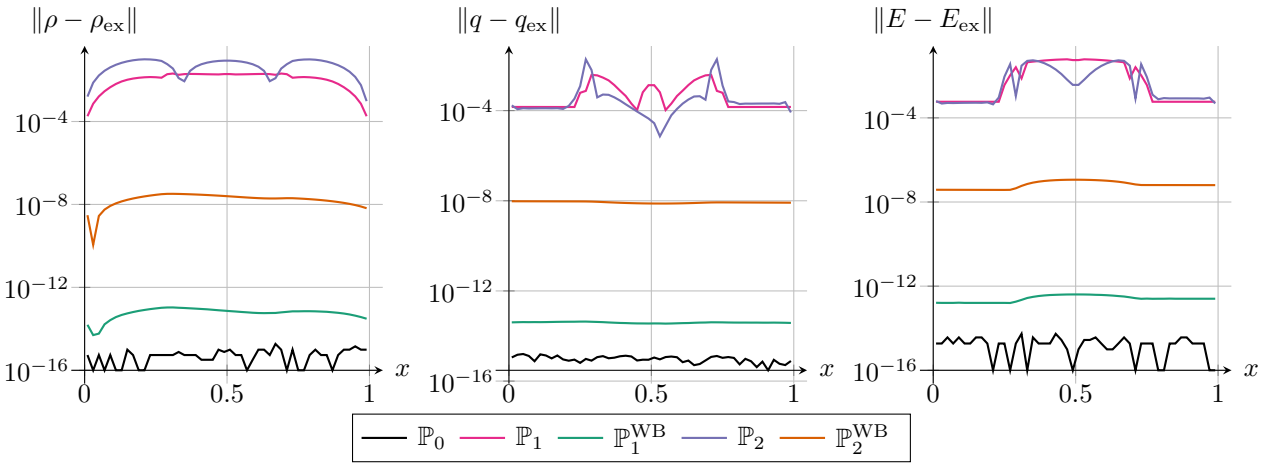


Figure 5.15: Euler equations with gravity: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the perturbed steady state experiment with 50 cells. Left panel: error on $\rho$; middle panel: error on $q$; right panel: error on $E$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $\rho$ | $4.03 \cdot 10^{-16}$ | $1.63 \cdot 10^{-2}$ | $4.96 \cdot 10^{-15}$ | $6.77 \cdot 10^{-2}$ | $3.19 \cdot 10^{-15}$ |
| error on $q$ | $5.85 \cdot 10^{-16}$ | $1.55 \cdot 10^{-3}$ | $1.12 \cdot 10^{-14}$ | $3.83 \cdot 10^{-3}$ | $6.80 \cdot 10^{-15}$ |
| error on $E$ | $8.42 \cdot 10^{-16}$ | $3.02 \cdot 10^{-2}$ | $4.76 \cdot 10^{-15}$ | $2.35 \cdot 10^{-2}$ | $2.89 \cdot 10^{-15}$ |
| CPU time (s) | 14.29 | 45.77 | 54.22 | 145.43 | 141.35 |

Table 5.12: Euler equations with gravity: errors between the steady solution and the approximate solutions at time $t_{\text{end}}$, for the perturbed steady state experiment with 50 cells.

### 5.4.3 Well-balancedness property – small perturbation

This last experiment is a small perturbation of a steady solution, with parameter $\beta = 10^{-2}$. We take $C_\theta = 7 \cdot 10^{-2}$, and the final physical time is $t_{\text{end}} = 5 \cdot 10^{-3}$.

The errors between the approximate solution and the underlying steady state are displayed in Figure 5.16. Once again, like for the previous systems, we note that the $\mathbb{P}_1$ and $\mathbb{P}_2$ schemes have produced errors several orders of magnitude larger than the small perturbation, and destroyed both the perturbation and the

underlying steady state. The numerical values of the errors reported in Table 5.5 once again validate this observation.

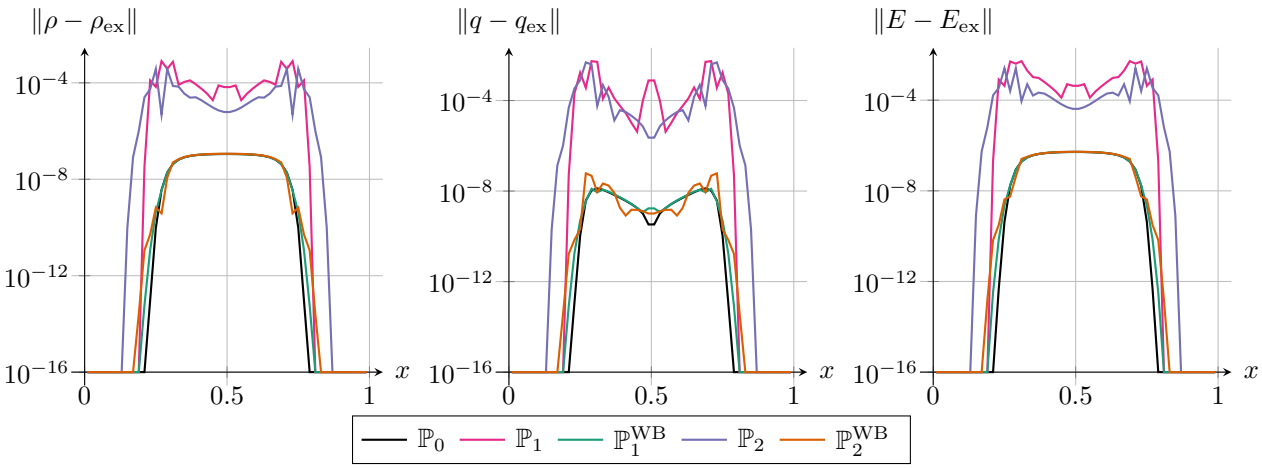


Figure 5.16: Euler equations with gravity: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells. Left panel: error on $\rho$; middle panel: error on $q$; right panel: error on $E$.

| | $\mathbb{P}_0$ scheme | $\mathbb{P}_1$ scheme | $\mathbb{P}_1^{\text{WB}}$ scheme | $\mathbb{P}_2$ scheme | $\mathbb{P}_2^{\text{WB}}$ scheme |
|---|---|---|---|---|---|
| error on $\rho$ | $6.04 \cdot 10^{-8}$ | $2.38 \cdot 10^{-4}$ | $6.07 \cdot 10^{-8}$ | $1.00 \cdot 10^{-4}$ | $6.20 \cdot 10^{-8}$ |
| error on $q$ | $4.90 \cdot 10^{-9}$ | $1.55 \cdot 10^{-3}$ | $4.69 \cdot 10^{-9}$ | $1.07 \cdot 10^{-3}$ | $2.79 \cdot 10^{-8}$ |
| error on $E$ | $2.78 \cdot 10^{-7}$ | $1.83 \cdot 10^{-3}$ | $2.79 \cdot 10^{-7}$ | $7.13 \cdot 10^{-4}$ | $2.85 \cdot 10^{-7}$ |

Table 5.13: Euler equations with gravity: errors between the underlying steady solution and the approximate solutions at time $t_{\text{end}}$, for the steady state with a small perturbation, using 50 cells.

# 6 Conclusion and perspectives

Usual high-order well-balanced methods involve costly inversions of nonlinear equations – or even systems of equations – to ensure both high-order accuracy and the preservation of complex steady solutions. The generic linear technique proposed in this manuscript avoids this extra cost while remaining able to exactly preserve steady states and to ensure high-order accuracy away from steady states.

Although this method has been presented in one space dimension, it can be naturally extended to more space dimensions. Indeed, consider the usual $D$-dimensional polynomial reconstruction of some variable $w$, which would replace the 1D polynomial reconstruction (4.2):

$$p_w^n(\mathbf{x}; i) = w_i^n + \pi_i^w(\mathbf{x} - \mathbf{x}_i),$$

where $\mathbf{x} \in \mathbb{R}^D$ and $\pi_i^w$ is a polynomial function of degree $d$, which satisfies some conditions detailed for instance in [22, 23, 39]. To compute the high-order $D$-dimensional numerical flux, the above reconstruction is evaluated at several Gauss points on the cell interfaces. Let $\boldsymbol{\zeta}_{i,j}$ be a Gauss point at the interface between cell $i$ and cell $j$. Then, in the spirit of (4.6), the evaluation of the polynomial at $\boldsymbol{\zeta}_{i,j}$ is modified as follows:

$$p_w^n(\boldsymbol{\zeta}_{i,j}; i) \coloneqq w_i^n + \theta_{i,j}\pi_i^w(\boldsymbol{\zeta}_{i,j} - \mathbf{x}_i),$$

where $\theta_{i,j}$ is nothing but the error to the steady state between cells $i$ and $j$. This $D$-dimensional extension is the object of future work.

Finally, numerical experiments show that the proposed technique performs as expected, i.e. we recover both high-order accuracy and steady state preservation at a negligible cost, for several hyperbolic systems.

The main advantages of this procedure thus are its genericity and its simplicity. Indeed, it is very easy to implement and, since it is linear, it does not involve much extra computational cost, if any. However, the technique requires prior knowledge of a first-order well-balanced scheme. Moreover, there remains one free parameter in the expression (3.2) of the steady state detector, but the value proposed in Section 5.1 seems relevant for each case treated so far.

An important difference between our method and existing literature, such as [15, 16, 26], also needs to be mentioned. In these papers, a local steady solution is obtained in each cell by solving an ODE similar to (1.2); then, the deviations with respect to this local steady solution are reconstructed; and the final reconstruction is given by adding the reconstructed deviations to the local steady solutions. This reconstruction strategy, although much more computationally expensive than the one introduced in the present paper, is able to preserve the distance to a perturbation of a global steady solution. Indeed, a reconstruction of a perturbation of amplitude $\varepsilon$ will remain $\varepsilon$-close to the underlying steady solution, which is not the case for our method. In our case, the perturbation is discretized with the high-order scheme, so with an error in $\mathcal{O}(\Delta x^\delta)$, which may be much larger than $\varepsilon$ for coarse meshes. Of course, this discussion becomes moot for fine meshes, since both strategies are consistent with a high order of accuracy.

## Acknowledgments

C. Berthon acknowledges the support of ANR MUFFIN ANR-19-CE46-0004.

## References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, 1992. Reprint of the 1972 edition.

[2] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[3] J. P. Berberich, P. Chandrashekar, C. Klingenberg, and F. K. Röpke. Second Order Finite Volume Scheme for Euler Equations with Gravity which is Well-Balanced for General Equations of State and Grid Systems. *Commun. Comput. Phys.*, 26(2):599–630, 2019.

[4] A. Bermudez and M. E. Vazquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. & Fluids*, 23(8):1049–1071, 1994.

[5] C. Berthon and C. Chalons. A fully well-balanced, positive and entropy-satisfying godunov-type method for the shallow-water equations. *Math. Comp.*, 85(299):1281–1307, 2016.

[6] C. Berthon, C. Chalons, S. Cornet, and G. Sperone. Fully well-balanced, positive and simple approximate Riemann solver for shallow water equations. *Bull. Braz. Math. Soc. (N.S.)*, 47(1):117–130, 2016.

[7] C. Berthon and V. Desveaux. An entropy preserving MOOD scheme for the Euler equations. *Int. J. Finite Vol.*, 11, 2014.

[8] C. Berthon and F. Foucher. Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *J. Comput. Phys.*, 231(15):4993–5015, 2012.

[9] C. Berthon and F. Marche. A positive preserving high order VFRoe scheme for shallow water equations: a class of relaxation schemes. *SIAM J. Sci. Comput.*, 30(5):2587–2612, 2008.

[10] C. Berthon and V. Michel-Dansac. A simple fully well-balanced and entropy preserving scheme for the shallow-water equations. *Appl. Math. Lett.*, 86:284–290, 2018.

[11] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2004.

[12] F. Bouchut, H. Ounaissa, and B. Perthame. Upwinding of the source term at interfaces for Euler equations with high friction. *Comput. Math. Appl.*, 53(3):361–375, 2007.

[13] S. Bryson, Y. Epshteyn, A. Kurganov, and G. Petrova. Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM Math. Model. Numer. Anal.*, 45(3):423–446, 2011.

[14] P. Cargo and A.-Y. Le Roux. Un schéma équilibre adapté au modèle d'atmosphère avec termes de gravité. *C. R. Acad. Sci., Paris, Sér. I*, 318(1):73–76, 1994.

[15] M. Castro, J. M. Gallardo, J. A. López-García, and C. Parés. Well-balanced high order extensions of Godunov's method for semilinear balance laws. *SIAM J. Numer. Anal.*, 46(2):1012–1039, 2008.

[16] M. J. Castro Díaz, J. A. López-García, and C. Parés. High order exactly well-balanced numerical methods for shallow water systems. *J. Comput. Phys.*, 246:242–264, 2013.

[17] C. Chalons, F. Coquel, E. Godlewski, P.-A. Raviart, and N. Seguin. Godunov-type schemes for hyperbolic systems with parameter-dependent source. The case of Euler system with friction. *Math. Models Methods Appl. Sci.*, 20(11):2109–2166, 2010.

[18] G. Chen and S. Noelle. A new hydrostatic reconstruction scheme based on subcell reconstructions. *SIAM J. Numer. Anal.*, 55(2):758–784, 2017.

[19] Y. Cheng, A. Chertock, M. Herty, A. Kurganov, and T. Wu. A new approach for designing moving-water equilibria preserving schemes for the shallow water equations. *J. Sci. Comput.*, 80(1):538–554, 2019.

[20] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N.-T. Vo, F. James, and S. Cordier. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *Internat. J. Numer. Methods Fluids*, 72(3):269–300, 2013.

[21] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *Internat. J. Numer. Methods Fluids*, 81(2):104–127, 2016.

[22] S. Diot, S. Clain, and R. Loubère. Improved detection criteria for the multi-dimensional optimal order detection (MOOD) on unstructured meshes with very high-order polynomials. *Comput. & Fluids*, 64:43–63, 2012.

[23] S. Diot, R. Loubère, and S. Clain. The multidimensional optimal order detection method in the three-dimensional case: very high-order finite volume method for hyperbolic systems. *Internat. J. Numer. Methods Fluids*, 73(4):362–392, 2013.

[24] J. M. Gallardo, C. Parés, and M. Castro. On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas. *J. Comput. Phys.*, 227(1):574–601, 2007.

[25] B. Ghitti, C. Berthon, M. H. Le, and E. F. Toro. A fully well-balanced scheme for the 1D blood flow equations with friction source term. *J. Comput. Phys.*, 421:109750, 2020.

[26] I. Gómez-Bueno, M. J. Castro, and C. Parés. High-order well-balanced methods for systems of balance laws: a control-based approach. *Appl. Math. Comput.*, 394:125820, apr 2021.

[27] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.*, 39(9-10):135–159, 2000.

[28] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.

[29] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[30] J.-L. Guermond, M. Quezada de Luna, B. Popov, C. E. Kees, and M. W. Farthing. Well-balanced second-order finite element approximation of the shallow water equations with friction. *SIAM J. Sci. Comput.*, 40(6):A3873–A3901, 2018.

[31] R. Käppeli and S. Mishra. Well-balanced schemes for the Euler equations with gravitation. *J. Comput. Phys.*, 259(0):199 – 219, 2014.

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

[33] G. Li and Y. Xing. Well-balanced discontinuous Galerkin methods with hydrostatic reconstruction for the Euler equations with gravitation. *J. Comput. Phys.*, 352:445–462, 2018.

[34] Q. Liang and F. Marche. Numerical resolution of well-balanced shallow water equations with complex source terms. *Adv. Water Resour.*, 32(6):873–884, 2009.

[35] J. Luo, K. Xu, and N. Liu. A well-balanced symplecticity-preserving gas-kinetic scheme for hydrodynamic equations under gravitational field. *SIAM J. Sci. Comput.*, 33(5):2356–2381, 2011.

[36] R. Manning. On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, 20:161–207, 1891.

[37] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Comput. Math. Appl.*, 72(3):568–593, 2016.

[38] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography or Manning friction. *J. Comput. Phys.*, 335:115–154, 2017.

[39] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A two-dimensional high-order well-balanced scheme for the shallow water equations with topography and Manning friction. *Comput. & Fluids*, 230:105152, 2021.

[40] S. Noelle, Y. Xing, and C.-W. Shu. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *J. Comput. Phys.*, 226(1):29–58, 2007.

[41] C. Parés and M. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *M2AN Math. Model. Numer. Anal.*, 38(5):821–852, 2004.

[42] B. Schmidtmann, B. Seibold, and M. Torrilhon. Relations Between WENO3 and Third-Order Limiting in Finite Volume Methods. *J. Sci. Comput.*, 68(2):624–652, 2015.

[43] A. Thomann, M. Zenk, and C. Klingenberg. A second-order positivity-preserving well-balanced finite volume scheme for Euler equations with gravity for arbitrary hydrostatic equilibria. *Internat. J. Numer. Methods Fluids*, 89(11):465–482, 2019.

[44] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics.* Springer-Verlag, Berlin, third edition, 2009. A practical introduction.

[45] B. van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *J. Comput. Phys.*, 32(1):101–136, 1979.

[46] Y. Xing. Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium. *J. Comput. Phys.*, 257(part A):536–553, 2014.

[47] Y. Xing and C.-W. Shu. High order well-balanced WENO scheme for the gas dynamics equations under gravitational fields. *J. Sci. Comput.*, 54(2-3):645–662, 2013.

[48] Y. Xing, C.-W. Shu, and S. Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *J. Sci. Comput.*, 48(1-3):339–349, 2011.

[49] K. Xu, J. Luo, and S. Chen. A well-balanced kinetic scheme for gas dynamic equations under gravitational field. *Adv. Appl. Math. Mech.*, 2:200–210, 2010.