# Well-balanced mesh-based and meshless schemes for the shallow-water equations

Alexander Bihlo Scott MacLachlan

November 28, 2017

**Abstract**

We formulate a general criterion for the exact preservation of the "lake at rest" solution in general mesh-based and meshless numerical schemes for the strong form of the shallow-water equations with bottom topography. The main idea is a careful mimetic design for the spatial derivative operators in the momentum flux equation that is paired with a compatible averaging rule for the water column height arising in the bottom topography source term. We prove consistency of the mimetic difference operators analytically and demonstrate the well-balanced property numerically using finite difference and RBF-FD schemes in the one- and two-dimensional cases.

## 1 Introduction

The shallow-water equations are a central model in geophysical fluid dynamics that is extensively used in the numerical simulation of propagating long waves such as tsunamis. A peculiarity of the shallow-water equations as used in ocean modeling is the presence of a source term arising due to a non-flat ocean bottom topography. An important exact steady-state solution of the shallow-water equations in this context is the *lake at rest solution*, i.e. that the total water height over arbitrary bottom topography remains constant and flat in the absence of a horizontal velocity field. From the numerical point of view, exactly preserving the lake at rest solution is usually the first benchmark for the quality of a discretization of the shallow-water equations. Since essentially all small amplitude wave solutions of the shallow-water equations can be regarded as perturbations of the lake at rest solution, the importance of preserving the lake at rest solution exactly cannot be overestimated, as it avoids the occurrence of spurious numerical waves that can render wholly inaccurate computed solutions. Numerical schemes that can preserve the lake at rest solution are called *well-balanced*. Designing such well-balanced numerical schemes for the shallow-water equations has attracted extensive interest in the literature, in particular in the finite-volume and discontinuous Galerkin methods communities. Examples of well-balanced numerical schemes are reported in, e.g., [1, 13, 17, 18, 26].

Besides finite-volume methods, several nodal-based discretization methodologies approximate derivatives of the field functions at a given point as linear combinations of certain weights with the field functions evaluated at nearby points. Examples for such methods include classical finite differences, meshless finite differences (where the weights are found using polynomial interpolation) and radial basis function based finite differences (RBF-FD; where the weights are found using radial basis function interpolation). It is this framework of nodal-based discretization for the shallow-water equations that we are interested in here.

Several numerical schemes for the shallow-water equations have been constructed over the past 20 years within the framework of the RBF methodology, see e.g. [15, 28, 30], but to the best of our knowledge none of these papers has explicitly studied the well-balanced properties of the constructed schemes. It was pointed out in [29] within the framework of a smoothed-particle hydrodynamics scheme that preserving the lake at rest solution in meshless approximations to the

shallow-water equations is a nontrivial endeavor. We show in the present paper that constructing well-balanced schemes for the shallow-water equations requires a careful design for the spatial derivatives in the momentum flux equations that is paired with a compatible averaging rule for the water column height arising in the momentum flux source terms. While mostly focusing on finite difference and RBF-FD derivative approximations, the derived conditions are applicable to any nodal-based derivative approximation for the shallow-water equations in the strong form.

Numerical conservation of physical properties, such as mass, momentum, and energy, has received significant attention in recent years, including many contributions to the literature of so-called mimetic discretizations; see, for example [3, 5, 16] and the references therein. For grid-based schemes, generalizations of finite-volume (or staggered finite-difference) approaches have been applied to the shallow-water equations in [24, 25], yielding schemes that conserve mass, momentum, and energy. Similar techniques have also been applied to discretizations in Lagrangian formulations [6, 7, 12], with conservation of potential vorticity for the smoothed-particle hydrodynamics discretization of the shallow-water equations shown in [12]. Very little work has been done, however, in terms of combining the mimetic methodology with general meshless methods, such as RBF-FD discretizations. The work presented here can be seen as a necessary first step in such development, although much more research is needed in this direction to see if a similarly broad class of mimetic meshless schemes can be realized.

The further organization of this paper is as follows. In Section 2, we present the rigorous theoretical analysis underlying well-balanced nodal derivative approximations for the shallow-water equations along with several examples of such well-balanced schemes, both for mesh-based finite differences and meshless RBF-FD approximations. Section 3 contains numerical simulations for well-balanced mesh-based and meshless schemes for the one-dimensional shallow-water equations as well as meshless schemes for the two-dimensional shallow-water equations. Section 4 is devoted to the conclusions of this work.

## 2 Well-balanced shallow-water equation discretizations

In this section, we review the shallow-water equations together with the lake at rest solution. We then proceed to introduce the general formalism for finding derivative approximations using weighted nodal-based approximations. Within this framework, we then derive general criteria for obtaining well-balanced discretization schemes for the shallow-water equations.

### 2.1 The shallow-water equations

The shallow-water equations with variable bottom topography are given by the transport equations for mass and momentum in the following form [21],

$$\boldsymbol{\rho}_t + \mathbf{F}_x + \mathbf{G}_y = \mathbf{S}, \tag{1}$$

where $\rho = (h, hu, hv)^{\mathrm{T}}$ is the vector of mass and momentum, $\mathbf{F} = (hu, hu^2 + h^2/2, huv)^{\mathrm{T}}$ and $\mathbf{G} = (hv, huv, hv^2 + h^2/2)^{\mathrm{T}}$ are the flux vectors, and $\mathbf{S} = (0, -hb_x, -hb_y)^{\mathrm{T}}$ is the source term. Here $h = h(t, x, y)$ denotes the depth of a water column of constant density, $(u, v)^{\mathrm{T}} = (u(t, x, y), v(t, x, y))^{\mathrm{T}}$ is the (horizontal) vector of vertically averaged fluid velocity and $b = b(x, y)$ is the prescribed bottom topography. Here and in the following, partial derivatives with respect to the independent variables $t$, $x$ and $y$ are denoted by subscripts. Note that for notational simplicity we apply the scaling $g = 1$, i.e. the gravitational constant is set to one.

The lake at rest solution is the steady state solution of (1) given by

$$u = v = 0, \quad h + b = \text{const}\,.$$

It states that in the absence of horizontal motion, the total height of the water column and bottom topography over every point in the spatial domain is constant and independent of time.

While it is usually straightforward to numerically preserve this steady state solution in the case of flat topography, $b = 0$, arbitrary sea bottom elevations are notoriously challenging to handle for typical shallow-water discretization schemes.

### 2.2 Computation of weights in nodal-based derivative approximations

The discretization framework we are interested in here is a slight generalization of the framework usually used for conventional finite difference approximations, see e.g. [8–10] for further details and a more in-depth discussion. Suppose we are given $n$ points $x_1 < x_2 < \cdots < x_n$ covering the (one-dimensional) spatial domain $\Omega = [x_1, x_n]$ as well as the values of a field function $f(x)$ at these points, $f_j = f(x_j)$, we want to find the weights $w_{ij}^{\mathcal{L}}$, $i, j = 1, \ldots, n$, such that for a given linear differential operator $\mathcal{L}$ we have

$$\mathcal{L}f|_{x=x_i} \approx \sum_{j=1}^{n} w_{ij}^{\mathcal{L}} f_j \tag{2}$$

To obtain the weights $w_{ij}^{\mathcal{L}}$ for the stencil of the point $x_i$, one assumes that the approximation (2) is exact for a given set of basis functions $\{\psi_k(x)\}$ over the entire stencil of $x_i$, i.e.

$$\mathcal{L}\psi_k(x_i) = \sum_{j=1}^{n} w_{ij}^{\mathcal{L}} \psi_k(x_j), \quad k = 1, \ldots, n. \tag{3}$$

This defines a linear system for $\left\{w_{ij}^{\mathcal{L}}\right\}_{j=1}^{n}$ for each node $i$. In general, care must be taken in the choice of nodes and basis functions so that this system has a solution that also yields a stable discretization (see, for example, [22]). Unique solvability is guaranteed when the coefficient matrix $(\psi_k(x_j))$ (restricted to points where $w_{ij}^{\mathcal{L}}$ is allowed to be nonzero) is square and non-singular, although more general situations are possible.

In the following we restrict ourselves to polynomial basis functions, i.e. $\psi_k(x) = x^k$ (when the nodes are on an interval of the real line), as well as to radial basis functions (RBFs), $\psi_k(x) = \phi(||x - x_k||)$, although the conditions on well-balanced shallow-water discretizations derived in Section 2.3 do not depend on the type of basis functions involved. For RBFs, it is clearly not essential that $x_i, x_k \in \mathbb{R}$, i.e. $x_j$ and $x_k$ could be vectors $\mathbf{x}_j$ and $\mathbf{x}_k$ in $\mathbb{R}^d$ as well, $d \geqslant 1$. For polynomial basis functions, standard finite-difference discretizations are obtained with appropriately chosen monomial basis functions on tensor-product meshes, and similar ideas can be extended to meshless finite-difference schemes assuming the points are suitably distributed through the domain (see, for example, [22]).

For polynomial basis functions in the one-dimensional case, the matrix $(\psi_k(x_j))$ in Eq. (3) is the Vandermonde matrix and hence non-singular if the points are distinct. The non-singularity of this matrix is also guaranteed for RBFs, again provided that no two points, $x_i$ and $x_j$, coincide.

It is also possible to consider a family of basis functions that includes both RBFs and polynomials. Such a combination is relevant in meshless RBF schemes as derivative approximations derived solely based on RBFs typically cannot reproduce trivial derivatives such as $\mathcal{L}c = 0$ exactly, for $\mathcal{L} \in \{\partial_x, \partial_{xx}, \ldots\}$ and $c = \text{const}$.

Note that once the system (3) is solved at all points $x_i$, $i = 1, \ldots, n$, we can assemble the weights $w_{ij}^{\mathcal{L}}$ in a differentiation matrix $W^{\mathcal{L}} = (w_{ij}^{\mathcal{L}})$. The derivatives of a field function $f$ at the nodal points $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ are thus approximated as $\mathcal{L}\mathbf{f} \approx W^{\mathcal{L}}\mathbf{f}$, where $\mathbf{f} = (f(x_1), \ldots, f(x_n))^{\mathrm{T}}$.

### 2.3 Well-balanced discretizations for the shallow-water equations

For the sake of simplicity of the following exposition, we consider the one-dimensional form of the shallow-water equation (1), i.e.

$$h_t + (hu)_x = 0, \qquad (hu)_t + \left(hu^2 + \frac{1}{2}h^2\right)_x = -hb_x.$$

Extension to the two-dimensional case is straightforward by enforcing that Eqs. (4) and (5) below have to hold for the $y$-derivative approximation as well.

Since $u = 0$ in the lake at rest solution, we need to preserve the property

$$\frac{1}{2}\partial_x h^2 = -h\partial_x b,$$

numerically for the case when $h + b = c$, where $c = \text{const}$. At the discrete level, this translates to the requirement that, at all nodal points,

$$\frac{1}{2}\left(\mathrm{D}_x^{\mathrm{f}} h^2\right)_i = -\overline{h}_i \left(\mathrm{D}_x^{\mathrm{s}}(c-h)\right)_i,$$

is satisfied, where $\mathrm{D}_x^{\mathrm{f}}$ and $\mathrm{D}_x^{\mathrm{s}}$ are the discrete first derivative operators for the partial derivatives with respect to $x$ arising in the flux and source terms of the shallow-water equations (not necessarily the same), and $\overline{h}_i = \sum_{j=1}^n m_{ij} h_j$ denotes an appropriate average over the field function $h$ in the stencil of $x_i$. Note that consistency of the average requires that $\sum_{j=1}^n m_{ij} = 1$.

The above equality is naturally satisfied if the following two conditions hold for all $i$

$$(\mathrm{D}_x^{\mathrm{s}} c)_i = 0, \qquad \frac{1}{2}\left(\mathrm{D}_x^{\mathrm{f}} h^2\right)_i = \overline{h}_i \left(\mathrm{D}_x^{\mathrm{s}} h\right)_i. \tag{4}$$

The first condition arises naturally as a consistency condition on $\mathrm{D}_x^{\mathrm{s}}$, and it is the second condition that requires more effort to achieve. A key step to this is to generalize $\mathrm{D}_x^{\mathrm{f}} h^2$ so that, rather than discretizing this derivative to act on a vector of values of $h^2$, it acts as a bilinear form,

$$\left(\frac{1}{2}\mathrm{D}_x^{\mathrm{f}} h^2\right)_i = \frac{1}{2}\mathbf{h}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{h}.$$

In this way, we define a *differentiation tensor* of order 3, $\mathbf{W}^{\mathrm{f}}$, whose $i^{\text{th}}$ slice is the matrix $W_i^{\mathrm{f}}$ used above. There are two important properties of $W_i^{\mathrm{f}}$ to note. First, the classical case, where the derivative operator acts on the vector of values of $h^2$, is still allowed, simply by choosing $W_i^{\mathrm{f}}$ to be a diagonal matrix. Secondly, since we only consider values of $\mathbf{h}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{h}$, only the symmetric part of $W_i^{\mathrm{f}}$ matters. In what follows, we assume $W_i^{\mathrm{f}}$ to be symmetric, except where noted.

For the right-hand (source) derivative, $\mathrm{D}_x^{\mathrm{s}}$, we use a standard discretization as a matrix, writing

$$(\mathrm{D}_x^{\mathrm{s}} h)_i = (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{h},$$

where we write $\mathbf{w}_i^{\mathrm{s}} = (W_{ij}^{\mathrm{s}})_{1 \leqslant j \leqslant n}$ for the $i$th row of the matrix $W^{\mathrm{s}}$. Similarly writing $\mathbf{m}_i = (m_{ij})_{1 \leqslant j \leqslant n}$ for the averaging stencil, the second condition in (4) can be represented as

$$\frac{1}{2}\mathbf{h}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{h} = (\mathbf{m}_i^{\mathrm{T}} \mathbf{h})((\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{h}) \text{ for all } i.$$

From this it follows that

$$\frac{1}{2}\mathbf{h}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{h} = \mathbf{h}^{\mathrm{T}} (\mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}}) \mathbf{h},$$

and thus the following relation among the weights in the derivative approximations and the averaging relation has to hold for all $i$:

$$W_i^{\mathrm{f}} = \mathbf{m}_i \left(\mathbf{w}_i^{\mathrm{s}}\right)^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}}. \tag{5}$$

In other words, specifying an averaging matrix $M$ and the weights for the derivative matrix, $\mathrm{W}_x^{\mathrm{s}}$, Eq. (5) prescribes weights in the flux derivative $\mathrm{D}_x^{\mathrm{f}} h^2$, represented by the tensor $\mathbf{W}^{\mathrm{f}}$, such that the resulting numerical scheme for the shallow-water equations will be well-balanced. Alternately, given $\mathbf{W}^{\mathrm{f}}$ and one of the matrices $M$ and $W^{\mathrm{s}}$, it can be used to check if the other matrix can be defined in such a way as to yield a well-balanced scheme. We further motivate this approach by stating the following theorem.

**Theorem 2.1.** *Let the derivative tensor, $\mathbf{W}^{\mathrm{f}}$, derivative matrix, $W^{\mathrm{s}}$, and averaging matrix, $M$, be given. The resulting discretization is well-balanced, satisfying (4) at every nodal point, if and only if $W^{\mathrm{s}}\mathbf{1} = \mathbf{0}$ (where $\mathbf{1}$ and $\mathbf{0}$ represent the vectors with all entries equal to 1 and 0, respectively) and, for every $i$,*

$$(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{m}_i = \frac{(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}}}{2(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}}}, \tag{6a}$$

$$\mathbf{v}^{\mathrm{T}} \mathbf{m}_i = \frac{\mathbf{v}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}}}{(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}}} \text{ for any } \mathbf{v} \perp \mathbf{w}_i^{\mathrm{s}}, \tag{6b}$$

*and*

$$\mathbf{v}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{u} = 0 \text{ for any } \mathbf{u}, \mathbf{v} \perp \mathbf{w}_i^{\mathrm{s}}. \tag{7}$$

*Proof.* First note that $W^s\mathbf{1} = \mathbf{0}$ naturally implies the first condition in (4). Next, from (5), recalling that $W_i^{\mathrm{f}}$ is symmetric, we have that

$$(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}} = 2((\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{m}_i) \left((\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}}\right).$$

Similarly, we also have that for any $\mathbf{v} \perp \mathbf{w}_i^{\mathrm{s}}$,

$$\mathbf{v}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}} = \mathbf{v}^{\mathrm{T}} \mathbf{m}_i \, (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \, \mathbf{w}_i^{\mathrm{s}}.$$

Finally, we derive (7) by noting that (5) implies that

$$\mathbf{v}^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{u} = 0, \tag{8}$$

whenever $\mathbf{u}, \mathbf{v} \perp \mathbf{w}_i^{\mathrm{s}}$. □

When $W^{\mathrm{s}}$ and $M$ are specified, Eq. (5) directly prescribes the flux differentiation tensor, $\mathbf{W}^{\mathrm{f}}$, so that its slices are given by symmetric outer products of the rows of $W^{\mathrm{s}}$ and $M$. When both differentiation rules, $\mathbf{W}^{\mathrm{f}}$ and $W^{\mathrm{s}}$, are specified, then an algorithmic form of Theorem 2.1 can be expressed by introducing a basis $\langle \mathbf{w}_i^{\mathrm{s}}, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1} \rangle$ where the $n-1$ vectors $\mathbf{v}_j$ are pairwise orthogonal as well as orthogonal to $\mathbf{w}_i^{\mathrm{s}}$, i.e.

$$\mathbf{v}_j^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} = 0, \quad \text{for} \quad 1 \leqslant j \leqslant n-1, \qquad \mathbf{v}_j^{\mathrm{T}} \mathbf{v}_k = 0, \quad \text{for} \quad j \neq k. \tag{9}$$

Then, the existence of a compatible averaging rule is guaranteed by Eq. (7), which can be expressed as

$$\mathbf{v}_j^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{v}_k = 0, \tag{10}$$

for $1 \leqslant j, k \leqslant n-1$. If these conditions are satisfied, then the averaging rule itself is specified for point $i$ by specializing (6) to the basis, giving

$$(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{m}_i = \frac{(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}}}{2(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}}} \text{ and } \mathbf{v}_j^{\mathrm{T}} \mathbf{m}_i = \frac{\mathbf{v}_j^{\mathrm{T}} W_i^{\mathrm{f}} (\mathbf{w}_i^{\mathrm{s}})}{(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}}} \text{ for } 1 \leqslant j \leqslant n-1. \tag{11}$$

The derivation of Eqs. (6) and (7), as well as of the basis discussed above, can be restated in the obvious way to show both existence and the definition of a compatible source differentiation matrix, $W^{\mathrm{s}}$, given $\mathbf{W}^{\mathrm{f}}$ and $M$.

Additionally, Eq. (5) can be used to understand the consistency and accuracy of the rules derived as described above in relation to their continuum counterparts. To do so, we define vectors $\mathbf{x}^{(p)}$ for $p \geqslant 0$ such that $\left(\mathbf{x}^{(p)}\right)_i = (x_i)^p$ for all $i$, noting that the case of $p = 0$ corresponds to the vector, $\mathbf{1}$, of all ones. We define the following three consistency/accuracy conditions:

$$\begin{aligned}
A_P: &\quad \mathbf{m}_i^{\mathrm{T}}\mathbf{x}^{(p)} = (x_i)^p && 0 \leqslant p \leqslant P \\
B_P: &\quad (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}}\mathbf{x}^{(p)} = p(x_i)^{p-1} && 0 \leqslant p \leqslant P \\
C_{P,Q}: &\quad (\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)} = (p+q)(x_i)^{p+q-1} && 0 \leqslant p \leqslant P,\ 0 \leqslant q \leqslant Q
\end{aligned}$$

For basic consistency, we would require that $A_0$ holds for all $i$ (meaning that $M$ defines a true averaging (row stochastic) matrix), $B_1$ hold for all $i$ (so that the discrete derivative of a constant is zero and of a linear function is its slope). Consistency is somewhat less natural for $\mathbf{W}^{\mathrm{f}}$, and could be expressed either as $C_{1,0}$ (which is equivalent to $C_{0,1}$ since $W_i^{\mathrm{f}}$ is symmetric) or $C_{1,1}$ holding for all $i$. In the former case, this requires that the discrete flux derivative reproduce the true derivative on constants (the case when $p = q = 0$) and that the scheme produced by "flattening" the flux derivative matrix at point $i$ into a row vector as $(\mathbf{x}^{(0)})^{\mathrm{T}} W_i^{\mathrm{f}}$ exactly reproduces the derivative of a linear function. Requiring the stronger condition $C_{1,1}$ requires the flux derivative also to be faithful to the true derivative of a quadratic function, which is counter to our usual expectation of consistency of the first derivative, but more natural when recalling this is an approximation to the derivative of $h^2$ and not $h$ itself. When these conditions hold for all $i$ and larger values of $P$ and $Q$, they express accuracy conditions that are natural in the usual sense for meshless finite differences, requiring that they be accurate pointwise for monomials up to a given order.

The natural question to be answered in the context of Theorem 2.1 is whether or not the conditions given there, together with consistency and accuracy in the sense of conditions $A_P$, $B_P$, and/or $C_{P,Q}$ yield consistency and accuracy for the third term in the well-balanced scheme. The following results present each possible implication.

**Theorem 2.2.** *Let $M$ and $W^{\mathrm{s}}$ be given, and assume conditions $A_P$ and $B_Q$ are satisfied for each nodal point $i$ with $P, Q \geqslant 0$. Let $\mathbf{W}^{\mathrm{f}}$ be determined by Eq. (5). Then condition $C_{R,R}$ holds for all $i$ with $R = \min(P, Q)$.*

*Proof.* For any nodal point $i$, consider $(\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)}$ for $0 \leqslant p, q \leqslant \min(P, Q)$:

$$\begin{aligned}
(\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)} &= (\mathbf{x}^{(q)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(p)} + (\mathbf{x}^{(q)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(p)} \\
&= (x_i^q)(p x_i^{p-1}) + (q x_i^{q-1})(x_i^p) = (p+q)(x_i)^{p+q-1}.
\end{aligned}$$

□

**Theorem 2.3.** *Let $M$ and $\mathbf{W}^{\mathrm{f}}$ be given, and assume conditions $A_R$ and $C_{P,Q}$ are satisfied for each nodal point $i$ with $P, Q, R \geqslant 0$. Assume there exists a matrix $W^{\mathrm{s}}$ such that Eq. (5) holds and the conditions of Theorem 2.1 are satisfied. Then condition $B_S$ holds for all $i$ with $S = \max(P, Q)$.*

*Proof.* Without loss of generality, we consider the case where $P \geqslant Q$. First consider $(\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)}$ for $p = q = 0$, which gives

$$0 = (\mathbf{x}^{(0)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(0)} = (\mathbf{x}^{(0)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(0)} + (\mathbf{x}^{(0)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(0)} = 2(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(0)}.$$

Thus, $(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(0)} = 0$. With this, we consider $(\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)}$ for $q = 0$, $1 \leqslant p \leqslant P$, giving

$$p x_i^{p-1} = (\mathbf{x}^{(0)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)} = (\mathbf{x}^{(0)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(p)} + (\mathbf{x}^{(0)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(p)} = 1(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(p)} + 0.$$

Thus, $(\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(p)} = p(x_i)^{p-1}$ for $1 \leqslant p \leqslant P$.

□

**Theorem 2.4.** *Let $W^{\mathrm{s}}$ and $\mathbf{W}^{\mathrm{f}}$ be given, and assume conditions $B_R$ and $C_{P,Q}$ are satisfied for each nodal point $i$ with $P, Q, R \geqslant 1$. Assume there exists a matrix $M$ such that Eq. (5) holds and the conditions of Theorem 2.1 are satisfied. Then condition $A_S$ holds for all $i$ with $S = \min(\max(P, Q), R)$.*

*Proof.* Without loss of generality, we consider the case where $P \geqslant Q$. First consider $(\mathbf{x}^{(q)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)}$ for $p = 0$, $q = 1$, which gives

$$1 = (\mathbf{x}^{(1)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(0)} = (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(0)} + (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(0)} = 0 + 1 \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(0)}.$$

Thus, $\mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(0)} = 1$. Similarly, taking $p = q = 1$ gives

$$2x_i = (\mathbf{x}^{(1)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(1)} = (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(1)} + (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(1)} = 2 \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(1)}.$$

Thus, $\mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(1)} = x_i$. Finally, considering the general case with $q = 1$ and $2 \leqslant p \leqslant \min(P, R)$, we have

$$(p+1) x_i^p = (\mathbf{x}^{(1)})^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{x}^{(p)} = (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{m}_i (\mathbf{w}_i^{\mathrm{s}})^{\mathrm{T}} \mathbf{x}^{(p)} + (\mathbf{x}^{(1)})^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(p)} = p x_i^p + \mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(p)}.$$

This implies that $\mathbf{m}_i^{\mathrm{T}} \mathbf{x}^{(p)} = x_i^p$ for all $i$. □

These results illustrate a natural asymmetry between the consistency/accuracy of the various terms defined via Eq. (5) and Theorem 2.1. This is most noticeable in Theorem 2.3, where only basic consistency of $M$ is needed for $W^{\mathrm{s}}$ to inherit the full accuracy of $\mathbf{W}^{\mathrm{f}}$. In contrast, if both $W^{\mathrm{s}}$ and $\mathbf{W}^{\mathrm{f}}$ are consistent, Theorem 2.4 shows that $M$ inherits only the lower level of accuracy from them.

We now provide several examples from classical finite differences on a uniform mesh with spacing $\Delta x$ to demonstrate the consequences of Eq. (5) for prescribing $\mathbf{W}^{\mathrm{f}}$ when $W^{\mathrm{s}}$ and $M$ are given. In what follows, $\mathbf{e}_i$ denotes the canonical $i$th unit vector.

**Example 2.1.** Suppose we wish to take both derivatives to be given by first-order upwind discretizations, with

$$\mathbf{w}_i^{\mathrm{s}} = \frac{1}{\Delta x} (\mathbf{e}_i - \mathbf{e}_{i-1}) \tag{12a}$$

and

$$W_i^{\mathrm{f}} = \frac{1}{\Delta x} (\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} - \mathbf{e}_{i-1} \mathbf{e}_{i-1}^{\mathrm{T}}). \tag{12b}$$

To satisfy the orthogonality condition in (9), we take $\mathbf{v}_1 = \mathbf{e}_{i-1} + \mathbf{e}_i$, and complete $\mathbf{v}_2$ through $\mathbf{v}_{n-1}$ with the unit vectors $\mathbf{e}_j$ for $j \neq i$ and $j \neq i-1$. It is straightforward to see that $W_i^{\mathrm{f}} \mathbf{v}_j = \mathbf{0}$ for $2 \leqslant j \leqslant n-1$, meaning that we only need to verify (10) for $\mathbf{v}_1$ and then use (11) to define $\mathbf{m}_i = m_{i-1} \mathbf{e}_{i-1} + m_i \mathbf{e}_i$.

To verify (10), we see that $W_i^{\mathrm{f}} \mathbf{v}_1 = (\mathbf{e}_i - \mathbf{e}_{i-1})/\Delta x$ and, so $\mathbf{v}_1^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{v}_1 = 0$. Since $W_i^{\mathrm{f}} \mathbf{w}_i^{\mathrm{s}} = (\mathbf{e}_i + \mathbf{e}_{i-1})/\Delta x^2$, the first equation in (11) forces $m_{i-1} = m_i$. Computing from the second, we find that these both take value $1/2$, giving $\mathbf{m}_i = (\mathbf{e}_i + \mathbf{e}_{i-1})/2$. Direct calculation shows that Eq. (5) is satisfied for this $\mathbf{m}_i$.

Written in component form, the associated upwind scheme defined through (12) reads

$$\frac{1}{2} \left( \frac{h_i^2 - h_{i-1}^2}{\Delta x} \right) = \bar{h}_i \frac{b_i - b_{i-1}}{\Delta x}, \quad \bar{h}_i = \frac{1}{2} (h_i + h_{i-1}).$$

which is directly seen to be well-balanced. From (12), we can verify that conditions $B_1$ and $C_{1,0}$ are satisfied for all $i$ by the upwind discretizations. In this case (since $C_{1,1}$ does not hold for all $i$), Theorem 2.4 does not apply, and it can easily be seen that $M$ satisfies $A_0$ for all $i$, but not $A_1$. Considering the alternate implications, if we were to specify $M$ and $W^{\mathrm{s}}$, Theorem 2.2 would confirm that $A_0$ and $B_1$ for all $i$ implies $C_{0,0}$ for all $i$, but not $C_{1,1}$ for all $i$. Similarly, since $A_0$ and $C_{1,0}$ hold for all $i$, Theorem 2.3 implies that $B_1$ holds for all $i$.

**Example 2.2.** A similar calculation verifies that taking a centered averaging for $\mathbf{m}_i$ yields a well-balanced scheme when the two derivatives are approximated by centered finite differences. Setting

$$\mathbf{m}_i = \frac{1}{2}(\mathbf{e}_{i+1} + \mathbf{e}_{i-1}), \quad \mathbf{w}_i^{\mathrm{s}} = \frac{1}{2\Delta x}(\mathbf{e}_{i+1} - \mathbf{e}_{i-1}), \tag{13a}$$

then Eq. (5), gives

$$W_i^{\mathrm{f}} = \frac{1}{2\Delta x}(\mathbf{e}_{i+1}\mathbf{e}_{i+1}^{\mathrm{T}} - \mathbf{e}_{i-1}\mathbf{e}_{i-1}^{\mathrm{T}}). \tag{13b}$$

Component-wise the differentiation and averaging rule (13) imply that, at the node $x_i$, our well-balanced scheme is given by

$$\frac{1}{2}\left(\frac{h_{i+1}^2 - h_{i-1}^2}{2\Delta x}\right) = \bar{h}_i \frac{b_{i+1} - b_{i-1}}{2\Delta x}, \quad \bar{h}_i = \frac{1}{2}(h_{i+1} + h_{i-1}).$$

which obviously satisfies the conditions (4). Considering the consistency/accuracy conditions for these rules, we can directly verify that $A_1$, $B_2$, $C_{1,1}$, and $C_{2,0}$ hold for all $i$. (Note that neither $C_{1,1}$ nor $C_{2,0}$ implies the other, and that $C_{2,1}$ does not hold for all $i$ for this choice of $\mathbf{W}^{\mathrm{f}}$.) Theorem 2.2 states that $A_1$ and $B_2$ together imply $C_{1,1}$, Theorem 2.3 states that $A_1$ and $C_{2,0}$ together imply $B_2$ , and Theorem 2.4 states that $B_2$ and $C_{1,1}$ imply $A_1$. We note that the conclusions of these theorems naturally depend differently on $P$ and $Q$ in $C_{P,Q}$, with $\max(P, Q)$ appearing in Theorem 2.3, but $\min(P, Q)$ in Theorem 2.4.

**Example 2.3.** When we choose centered differencing for the source derivative,

$$\mathbf{w}_i^{\mathrm{s}} = \frac{1}{2\Delta x}(\mathbf{e}_{i+1} - \mathbf{e}_{i-1}), \tag{14}$$

we can consider which values for $W_i^{\mathrm{f}}$ are possible to achieve a well-balanced scheme. If we restrict $W_i^{\mathrm{f}}$ to have a nonzero pattern over only the three points $i-1$, $i$, and $i+1$, we can write

$$\begin{aligned} W_i^{\mathrm{f}} &= w_{i-1,i-1}\mathbf{e}_{i-1}\mathbf{e}_{i-1}^{\mathrm{T}} + w_{i,i}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} + w_{i+1,i+1}\mathbf{e}_{i+1}\mathbf{e}_{i+1}^{\mathrm{T}} \\ &\quad + w_{i-1,i}(\mathbf{e}_{i-1}\mathbf{e}_i^{\mathrm{T}} + \mathbf{e}_i\mathbf{e}_{i-1}^{\mathrm{T}}) + w_{i-1,i+1}(\mathbf{e}_{i-1}\mathbf{e}_{i+1}^{\mathrm{T}} + \mathbf{e}_{i+1}\mathbf{e}_{i-1}^{\mathrm{T}}) + w_{i,i+1}(\mathbf{e}_i\mathbf{e}_{i+1}^{\mathrm{T}} + \mathbf{e}_{i+1}\mathbf{e}_i^{\mathrm{T}}). \end{aligned} \tag{15}$$

Take $\mathbf{v}_1 = \mathbf{e}_i$, $\mathbf{v}_2 = \mathbf{e}_{i-1} + \mathbf{e}_{i+1}$, and complete $\mathbf{v}_3$ through $\mathbf{v}_{n-1}$ with $\mathbf{e}_j$ for $j \neq i$ and $j \neq i \pm 1$. From (10), we have $\mathbf{v}_1^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{v}_1 = w_{i,i} = 0$, $\mathbf{v}_2^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{v}_1 = w_{i-1,i} + w_{i,i+1} = 0$, and $\mathbf{v}_2^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{v}_2 = w_{i-1,i-1} + 2w_{i-1,i+1} + w_{i+1,i+1} = 0$. Simplifying (15), we then see a restricted form of

$$\begin{aligned} W_i^{\mathrm{f}} = w_{i-1,i-1}\mathbf{e}_{i-1}\mathbf{e}_{i-1}^{\mathrm{T}} + w_{i+1,i+1}\mathbf{e}_{i+1}\mathbf{e}_{i+1}^{\mathrm{T}} + w_{i-1,i}(\mathbf{e}_{i-1}\mathbf{e}_i^{\mathrm{T}} + \mathbf{e}_i\mathbf{e}_{i-1}^{\mathrm{T}} - \mathbf{e}_i\mathbf{e}_{i+1}^{\mathrm{T}} - \mathbf{e}_{i+1}\mathbf{e}_i^{\mathrm{T}}) \\ - \frac{1}{2}(w_{i-1,i-1} + w_{i+1,i+1})(\mathbf{e}_{i-1}\mathbf{e}_{i+1}^{\mathrm{T}} + \mathbf{e}_{i+1}\mathbf{e}_{i-1}^{\mathrm{T}}). \end{aligned}$$

In order to not break the flux form of the shallow-water equations, we need the off-diagonal terms in $W_i^{\mathrm{f}}$ to vanish, forcing both $w_{i-1,i+1} = 0$ and $w_{i-1,i-1} = -w_{i+1,i+1}$. In other words, the only consistent well-balanced discretization in flux form that uses centered differences for the source derivatives occurs when also using centered differences for the flux derivative, as in Example 2.2. Even if we were to allow breaking of the flux form, straightforward calculation shows that we cannot enforce consistency condition $C_{1,1}$ without also requiring that $w_{i-1,i+1} = 0$ and $w_{i-1,i-1} = -w_{i+1,i+1}$.

**Example 2.4.** To extend the above example, we consider the case where the right-hand (source) derivative is given by centered differencing, but the left-hand (flux) derivative is given by second-order upwinding, with

$$W_i^{\mathrm{f}} = \frac{1}{2\Delta x}(3\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} - 4\mathbf{e}_{i-1}\mathbf{e}_{i-1}^{\mathrm{T}} + \mathbf{e}_{i-2}\mathbf{e}_{i-2}^{\mathrm{T}}). \tag{16}$$

Note that $\mathbf{e}_i^{\mathrm{T}} \mathbf{w}_i^{\mathrm{s}} = 0$, but $\mathbf{e}_i^{\mathrm{T}} W_i^{\mathrm{f}} \mathbf{e}_i = 3/(2\Delta x)$. Thus, Theorem 2.1 states that no possible choice of $\mathbf{m}_i$ exists that yields a well-balanced scheme.

**Example 2.5.** We now consider the case where both derivatives are given by second-order upwinding, with

$$W_i^{\mathrm{f}} = \frac{1}{2\Delta x}(3\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} - 4\mathbf{e}_{i-1}\mathbf{e}_{i-1}^{\mathrm{T}} + \mathbf{e}_{i-2}\mathbf{e}_{i-2}^{\mathrm{T}}), \tag{17}$$

and

$$\mathbf{w}_i^{\mathrm{s}} = \frac{1}{2\Delta x}(3\mathbf{e}_i - 4\mathbf{e}_{i-1} + \mathbf{e}_{i-2}). \tag{18}$$

Note that we can naturally take $\mathbf{v}_1 = \mathbf{e}_i + \mathbf{e}_{i-1} + \mathbf{e}_{i+1}$, and can find $\mathbf{v}_2$ orthogonal to both $\mathbf{w}_i^{\mathrm{s}}$ and $\mathbf{v}_1$ by taking the cross-product of the two three-dimensional restrictions of these vectors, giving $\mathbf{v}_2 = 5\mathbf{e}_i + 2\mathbf{e}_{i-1} - 7\mathbf{e}_{i-2}$. By construction, $\mathbf{v}_2^{\mathrm{T}}\mathbf{w}_i^{\mathrm{s}} = 0$, but $\mathbf{v}_2^{\mathrm{T}}W_i^{\mathrm{f}}\mathbf{v}_2 = 108/(2\Delta x) \neq 0$. Thus, Theorem 2.1 again states that no possible choice of $\mathbf{m}_i$ exists that yields a well-balanced scheme for these choices.

The last two examples raise the question of whether higher-order well-balanced schemes are possible. This is easily addressed as a consequence of Equation (5).

**Theorem 2.5.** *Let the differentiation tensor, $\mathbf{W}^{\mathrm{f}}$, be given. If there exists an $i$ such that $W_i^{\mathrm{f}}$ is a diagonal matrix with more than 2 nonzero entries, then no well-balanced scheme exists.*

*Proof.* Equation (5) states that a scheme is well-balanced if and only if the symmetric part of $W_i^{\mathrm{f}}$ is a rank-two matrix for all $i$. When $W_i^{\mathrm{f}}$ is diagonal, then it is its own symmetric part. The rank of a diagonal matrix equals the number of nonzero entries in the matrix. Thus, if more than two nonzero entries appear in such a $W_i^{\mathrm{f}}$, the scheme cannot be well-balanced. □

This result highlights another asymmetry in the construction of well-balanced schemes, that the flux derivative cannot be freely prescribed. In particular, for flux form discretizations, where $W_i^{\mathrm{f}}$ is constrained to be diagonal, no higher-order finite-difference stencil can be accommodated under the restriction of only two nonzero weights. In constrast, Equation (5) and Theorem 2.2 state that for *any* choice of source derivative, $W^{\mathrm{s}}$, and averaging matrix, $M$, a well-balanced scheme can be defined, inheriting the lower of the consistency orders of $W^{\mathrm{s}}$ and $M$, albeit with no expectation that $W_i^{\mathrm{f}}$ be in flux form. Thus, of the possible ways to complete a well-balanced scheme, we exclusively adopt this latter one, which allows us to make free choices of $W^{\mathrm{s}}$ and $M$, prescribing a well-balanced scheme via Equation (5).

**Remark 2.1.** It follows from the first condition in (4) that well-balanced shallow-water equation discretizations need to employ derivative approximations that are exact for constants. Thus, if the RBF-FD or global RBF collocation methodology is invoked, the underlying RBF interpolant for the field function $f$ should be of the form $f(x) = \sum_{i=1}^{n} \alpha_i \phi(||x - x_i||) + \alpha_{n+1}$ with the constraint that $\sum_{i=1}^{n} \alpha_i = 0$. In other words, the RBF basis should be supplemented with the monomial $\{1\}$. Higher-order polynomials can be included in the basis as well for accuracy considerations, see, e.g., [2]. Such higher-order polynomials may play an important role for the accurate representation of more complicated, non-stationary solutions of the shallow-water equations.

**Remark 2.2.** It is well-known that the application of RBF-FD methods to purely convective PDEs is prone to numerical instabilities since the eigenvalues of the differentiation matrices tend to scatter to the right half of the complex plane. As a remedy, the inclusion of hyperviscosity was proposed [11], which allows shifting the eigenspectrum of the convective operators back into the left-half plane, thus allowing for the use of explicit time-stepping methods. We note here that adding hyperviscosity in the momentum equations (specifically, terms like $\Delta^k(uh)$, where $\Delta^k$ is the $k$-th power of the Laplacian operator) to the above well-balanced schemes is perfectly possible without tampering with the well-balanced property for the lake at rest solution (where $u = 0$).

**Remark 2.3.** As mentioned above, we note that the extension of these results to the two-dimensional case follows simply by applying the one-dimensional results twice, independently in each coordinate direction. This follows from substituting the lake-at-rest solution into Equation (1), yielding two independent conditions, that $\frac{1}{2}\frac{\partial h^2}{\partial x} = h\frac{\partial h}{\partial x}$ and $\frac{1}{2}\frac{\partial h^2}{\partial y} = h\frac{\partial h}{\partial y}$. Thus, no cross-derivative terms or other coupling arises in the development of well-balanced schemes in two dimensions.

# 3 Numerical simulations

In this section, we present some numerical verification for the above theoretical construction of well-balanced schemes for the shallow-water equations in the one- and two-dimensional case.

## 3.1 One-dimensional lake at rest solution

We solve the shallow-water equations using either centered finite differences with the averaging rule as defined in Example 2.2, or with the RBF-FD method. In the latter case, we exclusively use the multiquadric RBF, i.e. $\phi(r) = \sqrt{1+(\epsilon r)^2}$, augmented with the monomial $\{1\}$ (see Remark 2.1), where we set the shape parameter $\epsilon = 0.1$ in all experiments. The stencil size of the RBF-FD method is three (center point and the immediate neighbors to the left and to the right) and the averaging rule is a normalized Gaussian filter, with weights given by $C_i e^{-|x_j - x_i|}$ assigned at all points, $j$, that appear in the stencil for point $i$, and constant $C_i$ chosen so these weights sum to 1. The discretization for the flux derivative $\frac{1}{2}\mathrm{D}_x h^2$ is then computed using the condition (5). We note the flexibility in the framework defined above allows us to independently choose the source derivative and averaging rule, and this approach will always yield a well-balanced scheme. As time stepping, we use the Heun scheme. A total of $n = 100$ equally spaced points are used on the domain $\Omega = [-3, 3]$ where reflective boundary conditions were employed. The bottom topography is a cosine bump of amplitude $A = 7$ extending over the interval $[-1, 1]$, which is superimposed with white noise generated independently at each node using normally distributed random numbers with zero mean and unit variance. The initial total water height is $h_0 + b = 10$.

In Figure 1 we show the results of the numerical computations at $t = 10$ using the RBF-FD method. The results of the classical centered finite difference scheme presented in Example 2.2 are essentially the same, with slightly smaller errors overall, and are, hence, not displayed here.

Note that for irregular nodal layouts, the eigenvalues of the derivative matrices for the RBF-FD method scatter into the right half of the complex plane. This is particularly prominent in the multidimensional case and in the case that several neighboring nodal points are used for the computation of the derivative matrices. To improve the stability of the numerical schemes in these cases, hyperviscosity or other stabilization should be used.

Figure 1 shows that the RBF-FD scheme is indeed well-balanced, being able to maintain the constant water height even in the presence of quite rough bottom topography. We also monitored the conservation of total mass $\mathcal{M} = \sum_i h_i \Delta x_i$ and found conservation with relative errors of the magnitude $10^{-16}$ and hence machine precision (not shown here). The conservation of mass for the lake at rest solution is a particularly nice feature of the present well-balanced schemes as it is straightforward to check that mass is in general not conserved in numerical schemes for the shallow-water equations using the RBF-FD methodology.

In contrast, Figure 2 depicts the results obtained when the standard RBF-FD approximation is used for both the source and flux derivatives (i.e., not employing the well-balanced condition derived above). As expected, the violation of balance leads to the emergence of spurious waves that travel through the entire computational domain. The emergence of these waves is typically not tolerable in numerical schemes for the shallow-water equations, as they can lead to wrong

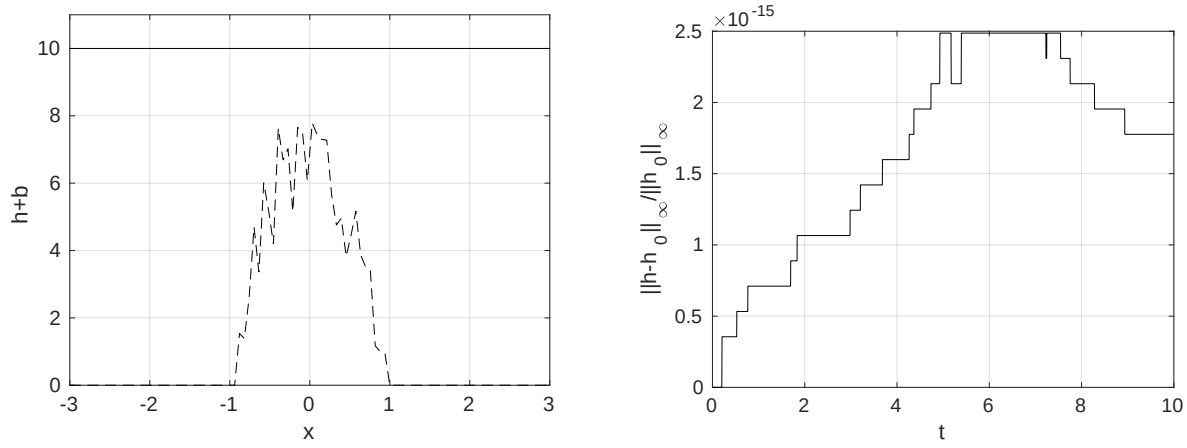


Figure 1. Numerical integration of the shallow-water equations using the *balanced* RBF-FD method on $n = 100$ regularly spaced nodes, integrated up to $t = 10$ with the Heun scheme. **Left:** Total water height at $t = 10$ (solid line) and bottom topography (dashed line). **Right:** Relative $l_\infty$-error in the water height.

run-up heights and, thus, to unphysical results estimating factors such as tsunami inundation or the stress on coastal structures.

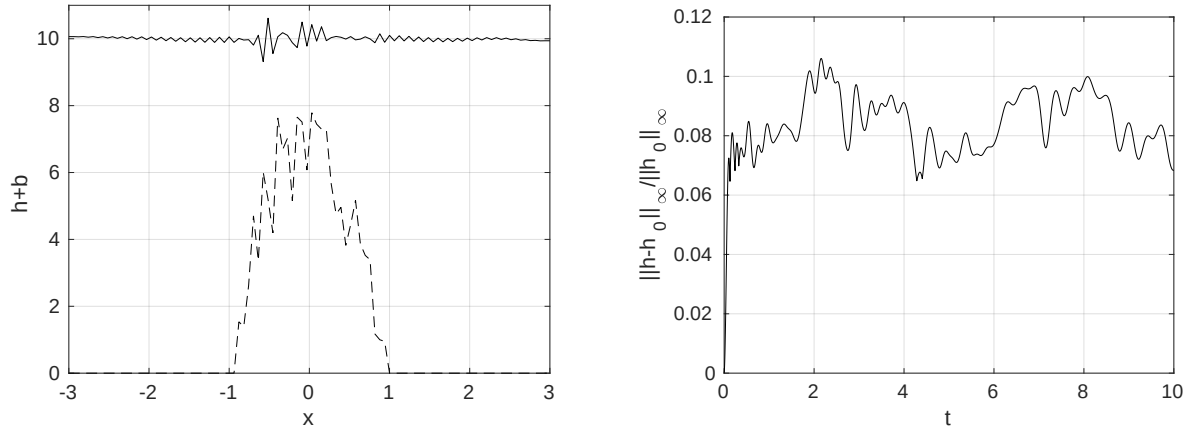


Figure 2. Numerical integration of the shallow-water equations using the *unbalanced* RBF-FD method on $n = 100$ regularly spaced nodes, integrated up to $t = 10$ with the Heun scheme. **Left:** Total water height at $t = 10$ (solid line) and bottom topography (dashed line). **Right:** Relative $l_\infty$-error in the water height.

### 3.2 Two-dimensional lake at rest solution

In the two-dimensional case, we constrain ourselves to the use of the RBF-FD method only. We consider the domain $\Omega = [-3, 3] \times [-3, 3]$ covered by $n = 1600$ nodes. To demonstrate the versatility and independence of the chosen nodal layout (mesh-based or meshfree) of the condition (5), we add $(0.1\Delta x\mathcal{N}(0, 1), 0.1\Delta y\mathcal{N}(0, 1))$ as disturbance to each nodal point originally lying on an orthogonal and equally spaced mesh, where $\mathcal{N}(0, 1)$ is a normally distributed random variable with zero mean and variance one. The resulting nodal layout is depicted in Figure 3.

The bottom topography used is a cosine bell on the area $[-1, 1] \times [-1, 1]$ with amplitude $A = 7$ and again superimposed with white noise obtained independently at each node from normally distributed random numbers with zero mean and unit variance.

We choose a stencil based on the 25 nearest neighbors of each point in the domain $\Omega$ for the computation of the RBF-FD differentiation matrices. Once again, the multiquadric RBF is used in all computations and while the current nodal layout might profit from a spatially variable shape parameter $\epsilon$ for improved accuracy, for the sake of simplicity we used $\varepsilon = 1$ in

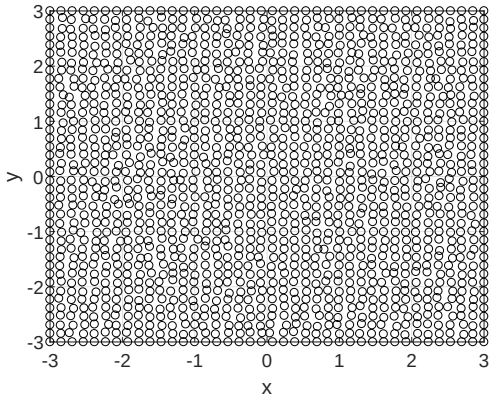


Figure 3. Nodal distribution for the balanced 2D shallow-water equations discretization.

all points. The averaging rule is a two-dimensional Gaussian filter over the 25 points in the stencil of each point. Once again, the flux derivative discretizations for $\frac{1}{2}\mathrm{D}_x h^2$ and $\frac{1}{2}\mathrm{D}_y h^2$ are obtained from the condition (5). We integrate the two-dimensional shallow-water equations with the Heun scheme up to $t = 10$. The results of this integration are displayed in Figure 4 and verify numerically that the scheme is indeed well-balanced.

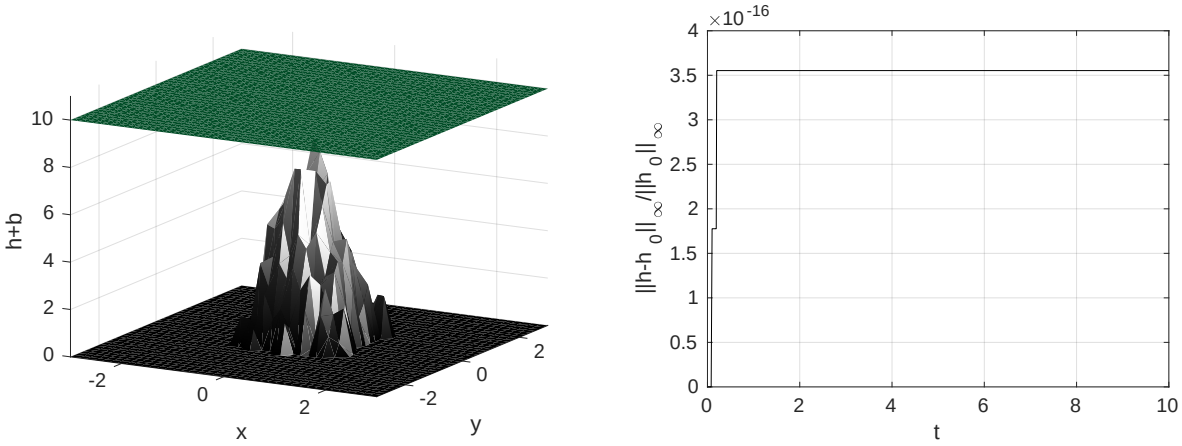


Figure 4. Numerical integration of the shallow-water equations using the RBF-FD method on $n = 1600$ irregularly spaced nodes, integrated up to $t = 10$ with the Heun scheme. **Left:** Total water height at $t = 10$ and bottom topography. **Right:** Relative $l_\infty$-error in the water height.

Stabilization of the scheme was done by adding hyperviscosity of the form $-(1)^{k+1}\nu\Delta^k(uh)$ and $-(1)^{k+1}\nu\Delta^k(vh)$ to the momentum equations in the $x$- and $y$-directions, respectively. We chose $k = 2$ and experimentally set $\nu$ such that the scheme remains stable but does not become unnecessarily diffusive. Note that similar results were obtained when refining the grid (i.e., starting from a finer uniform mesh but performing the same random perturbations to both node location and bottom topography). In this case, we replace the fixed hyperviscosity parameter, $\nu$, by $\nu(\Delta r)^{2k}$, where $\Delta r$ is a measure of the average nodal distance.

### 3.3 Parabolic bowl

Having verified the numerical preservation of the lake at rest solution, it is instructive to compare the numerical solution of the balanced scheme to that of an unbalanced scheme for a more challenging test case. Here, we consider oscillatory flow in a parabolic bowl, in the same setting as presented in [26], which is based on the exact solution derived in [23]. In particular, we use the domain $\Omega = [-5000, 5000]$ with parabolic bottom topography $b = h_0(x/a)^2$, where $a = 3000$

and $h_0 = 10$. The initial conditions are such that the exact solution to this benchmark test is given by

$$h_a(t,x) = h_0 - \frac{B^2}{4g}(1+\cos 2\omega t) - \frac{Bx}{2a}\sqrt{\frac{8h_0}{g}}\cos(\omega t), \quad u_a(t,x) = \frac{Ba\omega}{\sqrt{2h_0 g}}\sin\omega t$$

where $\omega = \sqrt{2gh_0}/a$ and $B = 5$.

Using the three-point RBF-FD and Gaussian filter averaging rule described above in Section 3.1, we integrate the shallow-water equations numerically until $t = 2000$. We consider three measures of the error for varying numbers of nodal points, $n$:

1. the maximum error in conservation of the total mass, $\mathcal{M}$, over all time steps,
2. the error in water height $h$, measured by taking the relative error in $h$ (measured in the maximum norm) at each time step, and measuring the maximum of these values over all time steps, and
3. the error in momentum, $hu$, measured by taking the absolute error in $hu$ (measured in the maximum norm) at each time step, and measuring the maximum of these values over all time steps.

Note that this solution requires an inundation model, since the water surface hits the bowl and, thus, creates a moving boundary condition. Inundation is not considered here, and we use the analytical solution to prescribe the time-varying boundary condition. For the discussion of a possible inundation model for this case, consult [4]. The results of this convergence study are reported in Table 1, showing that, as expected, the balanced scheme is consistently better than the unbalanced scheme, both being of second order. The errors differ most dramatically (by about one order of magnitude) for mass conservation.

Table 1. Error in mass conservation $\mathcal{M}$, relative $l_\infty$-error for the total water height $h$, absolute $l_\infty$-error for the momentum $hu$ for the balanced and unbalanced schemes for the oscillatory flow in a parabolic bowl.

| | $\mathcal{M}$ error | | $h$ error | | $hu$ error | |
|---|---|---|---|---|---|---|
| $n$ | balanced | unbalanced | balanced | unbalanced | balanced | unbalanced |
| 128 | $1.64 \cdot 10^{-4}$ | $1.48 \cdot 10^{-3}$ | $9.75 \cdot 10^{-4}$ | $2.44 \cdot 10^{-3}$ | $8.43 \cdot 10^{-2}$ | $1.91 \cdot 10^{-1}$ |
| 256 | $7.06 \cdot 10^{-5}$ | $3.37 \cdot 10^{-4}$ | $2.68 \cdot 10^{-4}$ | $8.74 \cdot 10^{-4}$ | $2.20 \cdot 10^{-2}$ | $8.74 \cdot 10^{-2}$ |
| 512 | $8.21 \cdot 10^{-6}$ | $7.90 \cdot 10^{-5}$ | $7.20 \cdot 10^{-5}$ | $1.77 \cdot 10^{-4}$ | $5.38 \cdot 10^{-3}$ | $1.31 \cdot 10^{-2}$ |
| 1024 | $2.43 \cdot 10^{-6}$ | $1.77 \cdot 10^{-5}$ | $1.88 \cdot 10^{-5}$ | $6.26 \cdot 10^{-5}$ | $1.36 \cdot 10^{-3}$ | $3.3 \cdot 10^{-3}$ |

For a second experiment, we consider fourth-order schemes and extend the comparison to include both a standard FD scheme and an RBF-FD scheme as described above. Since we continue to use the Heun scheme for time stepping, we now decrease the time step by a factor of four each time the number of points is space is doubled, in order to balance the errors between the second-order time stepper and the fourth-order spatial discretizations. We use uniform grids in space, and integrate to $t = 1000$; for $n = 64$ points in space, we use 250 points in the time direction, which approximately balances the spatial and temporal discretization errors at this spatial mesh size. For the standard FD scheme, we use the fourth-order (five-point) central difference stencil for the source derivative terms, the identity operator for the averaging rule, and the well-balanced prescription in (5) for the flux derivative. For the RBF-FD scheme, we also use a five-point discretization for the source derivative, following the description in Section 2.2, but now including polynomials up to third order in the construction of the differencing

scheme. To achieve a nontrivial fourth-order averaging rule, we use a five-point averaging, with weights of 1.6 for the point itself, $-0.4$ for the two immediate neighbours, and 0.1 for the two distance-two neighbours. We note that the normalized Gaussian filter used above cannot yield a fourth-order averaging, as some negative weights must appear in the averaging rule to attain fourth order. We consider this scheme in both balanced form, following (5) to prescribe the flux derivative, and unbalanced form, directly using the fourth-order RBF-FD derivative for the flux term. Numerical results are given in Table 2. We note slight differences in the errors from the FD and balanced RBF-FD schemes, but these are small overall. In comparison with the unbalanced RBF-FD scheme, however, we see notably larger errors in height and momentum, by up to a factor of three over the balanced schemes. Most notably, comparing results for $n = 256$ and $n = 512$, we see reductions in both height and momentum error by factors of almost 15 for the two balanced schemes (consistent with fourth-order discretization), but only by a factor of 7 or 8 for the unbalanced scheme. Taken together with the results for the lake at rest solution, these results indicate the advantage of choosing a well-balanced scheme for the shallow-water equations over an unbalanced scheme.

Table 2. Error in mass conservation $\mathcal{M}$, relative $l_\infty$-error for the total water height $h$, absolute $l_\infty$-error for the momentum $hu$ for the balanced and unbalanced schemes for the oscillatory flow in a parabolic bowl using fourth-order schemes.

| | | $n = 64$ | $n = 128$ | $n = 256$ | $n = 512$ |
|---|---|---|---|---|---|
| $\mathcal{M}$ error | FD | $7.16 \cdot 10^{-4}$ | $2.18 \cdot 10^{-4}$ | $8.82 \cdot 10^{-5}$ | $3.22 \cdot 10^{-6}$ |
| | RBF-FD bal. | $7.33 \cdot 10^{-4}$ | $1.66 \cdot 10^{-4}$ | $5.57 \cdot 10^{-5}$ | $5.47 \cdot 10^{-6}$ |
| | RBF-FD unbal. | $7.70 \cdot 10^{-4}$ | $1.65 \cdot 10^{-4}$ | $5.57 \cdot 10^{-5}$ | $5.48 \cdot 10^{-6}$ |
| $h$ error | FD | $2.52 \cdot 10^{-4}$ | $1.73 \cdot 10^{-5}$ | $1.19 \cdot 10^{-6}$ | $7.97 \cdot 10^{-8}$ |
| | RBF-FD bal. | $2.87 \cdot 10^{-4}$ | $2.09 \cdot 10^{-5}$ | $1.45 \cdot 10^{-6}$ | $9.84 \cdot 10^{-8}$ |
| | RBF-FD unbal. | $2.72 \cdot 10^{-4}$ | $2.61 \cdot 10^{-5}$ | $1.68 \cdot 10^{-6}$ | $2.06 \cdot 10^{-7}$ |
| $hu$ error | FD | $1.69 \cdot 10^{-2}$ | $1.11 \cdot 10^{-3}$ | $7.24 \cdot 10^{-5}$ | $4.92 \cdot 10^{-6}$ |
| | RBF-FD bal. | $1.88 \cdot 10^{-2}$ | $1.24 \cdot 10^{-3}$ | $8.15 \cdot 10^{-5}$ | $5.56 \cdot 10^{-6}$ |
| | RBF-FD unbal. | $1.48 \cdot 10^{-2}$ | $2.35 \cdot 10^{-3}$ | $1.01 \cdot 10^{-4}$ | $1.52 \cdot 10^{-5}$ |

## 4 Conclusion

In this paper, we have derived general criteria for obtaining well-balanced numerical schemes for the shallow-water equations that employ a nodal expansion for their spatial derivative approximation. We have shown analytically and numerically that the resulting discretization schemes for the shallow-water equations exactly maintain the lake at rest steady state, which is considered as the first important criterion for applying such schemes to real-world problems such as tsunami modeling. We have further proved consistency and order conditions for the discrete differential and averaging operators involved in these well-balanced schemes.

One particularly important feature of the derived schemes is that they do not require the nodes to lay on a uniform, orthogonal grid. Rather, any nodal distribution can be used and the resulting schemes will remain well-balanced. This feature is important as it guarantees that various adaptation strategies, such as $h$- and $r$-adaptivity (i.e. introducing or removing nodes, as well as dynamically redistributing them), can be used without complicating the design of the resulting numerical method. Due to the involved time and length scales in the shallow-water equations when used for tsunami modeling (e.g. open ocean wave propagation vs. coastal

inundation), adaptivity is usually a practical necessity [14].

The present work is also an important step in the development of mimetic methods for general meshless numerical schemes, in that we have derived criteria that derivative approximations have to mimic in order for the resulting numerical scheme to be well-balanced. More generally, mimetic discretization is an active field of research in which one aims to discretize differential equations is such a way that certain important identities from vector calculus will be preserved in a numerical scheme. This is important since these vector identities are typically associated with central conservation laws in the equations of hydrodynamics and electrodynamics. For an overview of mimetic discretization schemes and some examples, including the shallow-water equations, consult e.g. [3,5,16,24,25,27]. Most mimetic methods derived so far apply to the finite difference, finite element and finite volume methodologies only and thus exclude the important class of meshless integration schemes. As these schemes are getting increasingly popular in fields such as atmospheric sciences, ocean sciences and geophysics, whose governing equations all admit important conservation laws, finding mimetic schemes in the wider meshless framework is an important timely research field. For first results of this research perspective in geophysics, see [19, 20]. The present study for the shallow-water equations can thus be regarded as being amongst the first examples for mimetic methods within the meshless methodology.

## Acknowledgments

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and the NSERC Discovery Grant program. The authors thank Grady Wright for helpful discussions, and the two anonymous referees for their helpful and considerate remarks.

## References

[1] 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 (2004), pp. 2050–2065.

[2] V. Bayona, N. Flyer, B. Fornberg, and G. A. Barnett, *On the role of polynomials in RBF-FD approximations: II. Numerical solution of elliptic PDEs*, J. Comput. Phys., 332 (2017), pp. 257–273.

[3] P. B. Bochev and J. M. Hyman, *Principles of mimetic discretizations of differential operators*, in Compatible spatial discretizations, Springer, 2006, pp. 89–119.

[4] R. Brecht, A. Bihlo, S. MacLachlan, and J. Behrens, *A well-balanced meshless tsunami propagation and inundation model*. arXiv:1705.09831.

[5] F. Brezzi, K. Lipnikov, and M. Shashkov, *Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes*, SIAM J. Numer. Anal., 43 (2005), pp. 1872–1896.

[6] E. J. Caramana, D. E. Burton, M. J. Shashkov, and P. P. Whalen, *The construction of compatible hydrodynamics algorithms utilizing conservation of total energy*, J. Comput. Phys., 146 (1998), pp. 227–262.

[7] Svetlana Dubinkina and Jason Frank, *Statistical relevance of vorticity conservation in the Hamiltonian particle-mesh method*, J. Comput. Phys., 229 (2010), pp. 2634–2648.

[8] B. Fornberg and N. Flyer, *A Primer on Radial Basis Functions with Applications to the Geosciences*, vol. 3529, SIAM Press, Philadelphia, PA, 2015.

[9] ——, *Solving PDEs with radial basis functions*, Acta Numer., 24 (2015), pp. 215–258.

[10] B. Fornberg, E. Larsson, and N. Flyer, *Stable computations with gaussian radial basis functions*, SIAM J. Sci. Comput., 33 (2011), pp. 869–892.

[11] B. Fornberg and E. Lehto, *Stabilization of rbf-generated finite difference methods for convective pdes*, J. Comput. Phys., 230 (2011), pp. 2270–2285.

[12] Jason Frank and Sebastian Reich, *Conservation properties of smoothed particle hydrodynamics applied to the shallow water equation*, BIT, 43 (2003), pp. 41–55.

[13] T. Gallouët, J.-M. Hérard, and N. Seguin, *Some approximate Godunov schemes to compute shallow-water equations with topography*, Comput. & Fluids, 32 (2003), pp. 479–513.

[14] D. L. George and R. J. LeVeque, *Finite volume methods and adaptive refinement for global tsunami propagation and local inundation*, Sci. Tsunami Haz., 24 (2006), p. 319.

[15] Y.-C. Hon, K. F. Cheung, X.-Z. Mao, and E. J. Kansa, *Multiquadric solution for shallow water equations*, J. Hydraul. Eng., 125 (1999), pp. 524–533.

[16] J. M. Hyman and M. Shashkov, *Mimetic discretizations for Maxwell's equations*, J. Comput. Phys., 151 (1999), pp. 881–909.

[17] A. Kurganov and G. Petrova, *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system*, Commun. Math. Sci., 5 (2007), pp. 133–160.

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

[19] B. Martin and B. Fornberg, *Seismic modeling with radial basis function-generated finite differences (RBF-FD)–a simplified treatment of interfaces*, J. Comput. Phys., 335 (2017), pp. 828–845.

[20] B. Martin, B. Fornberg, and A. St-Cyr, *Seismic modeling with radial-basis-function-generated finite differences*, Geophysics, 80 (2015), pp. T137–T146.

[21] J. Pedlosky, *Geophysical fluid dynamics*, Springer, New York, 1987.

[22] Benjamin Seibold, *Minimal positive stencils in meshfree finite difference methods for the Poisson equation*, Comput. Methods Appl. Mech. Engrg., 198 (2008), pp. 592–601.

[23] W. C. Thacker, *Some exact solutions to the nonlinear shallow-water wave equations*, J. Fluid Mech., 107 (1981), pp. 499–508.

[24] M. Van Reeuwijk, *A mimetic mass, momentum and energy conserving discretization for the shallow water equations*, Comput. & Fluids, 46 (2011), pp. 411–416.

[25] B. Van't Hof and A. E. P. Veldman, *Mass, momentum and energy conserving (MaMEC) discretizations on general grids for the compressible Euler and shallow water equations*, J. Comput. Phys., 231 (2012), pp. 4723–4744.

[26] S. Vater, N. Beisiegel, and J. Behrens, *A limiter-based well-balanced discontinuous Galerkin method for shallow-water flows with wetting and drying: One-dimensional case*, Adv. Water Resour., 85 (2015), pp. 1–13.

[27] R. W. C. P. Verstappen and A. E. P. Veldman, *Symmetry-preserving discretization of turbulent flow*, J. Comput. Phys., 187 (2003), pp. 343–368.

[28] S. M. Wong, Y. C. Hon, and M. A. Golberg, *Compactly supported radial basis functions for shallow water equations*, Appl. Math. Comput., 127 (2002), pp. 79–101.

[29] X. Xia, Q. Liang, M. Pastor, W. Zou, and Y.-F. Zhuang, *Balancing the source terms in a SPH model for solving the shallow water equations*, Adv. Water Resour., 59 (2013), pp. 25–38.

[30] X. Zhou, Y. C. Hon, and K. F. Cheung, *A grid-free, nonlinear shallow-water model with moving boundary*, Eng. Anal. Bound. Elem., 28 (2004), pp. 967–973.