Water Waves
https://doi.org/10.1007/s42286-022-00062-8

ORIGINAL ARTICLE

Well-Balanced Second-Order Convex Limiting Technique for Solving the Serre–Green–Naghdi Equations

Jean-Luc Guermond[2] · Chris Kees[3] · Bojan Popov[2] · Eric Tovar[1,2]

Received: 29 September 2021 / Accepted: 19 May 2022

Abstract
In this paper, we introduce a numerical method for approximating the dispersive Serre–Green–Naghdi equations with topography using continuous finite elements. The method is an extension of the hyperbolic relaxation technique introduced in Guermond et al. (J Comput Phys 450:110809, 2022). It is explicit, second-order accurate in space, third-order accurate in time, and is invariant-domain preserving. It is also well balanced and parameter free. Special attention is given to the convex limiting technique when physical source terms are added in the equations. The method is verified with academic benchmarks and validated by comparison with laboratory experimental data.

Keywords Shallow water · Serre · Serre–Green–Naghdi · Well-balanced approximation · Invariant domain · Second-order accuracy · Finite-element method · Positivity-preserving · Entropy viscosity · Convex limiting

Mathematics Subject Classification 65M60 · 65M12 · 35L50 · 35L65 · 76M10

1 Introduction

The objective of this paper is to present an approximation technique for the dispersive Serre–Green–Naghdi equations (also known as just the Serre equations, Green–Naghdi equations or fully non-linear Boussinesq equations; see [13, 14, 30, 38, 39]) with topography using continuous finite elements and explicit time stepping. In addition to the

✉ Eric Tovar
ejtovar1@tamu.edu

1 U.S. Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory (ERDC-CHL), Vicksburg, MS 39180, USA

2 Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843, USA

3 Department of Civil and Environmental Engineering, Louisiana State University, 3255 Patrick F. Taylor, Baton Rouge, LA 70803, USA

Published online: 27 June 2022

topography source terms, we consider external physical source terms in the equations relevant for applications in coastal hydrodynamics. The idea is to construct a method that is at least second-order accurate in space, third order accurate in time, well balanced and invariant-domain preserving (i.e., robust with respect to dry states). The starting point of this present paper is the hyperbolic relaxation technique introduced in [12] and further expanded in [23] for solving the dispersive Serre–Green–Naghdi equations with topography effects. The approach reformulates the Serre–Green–Naghdi equations as a first-order hyperbolic system which allows for explicit time stepping. The hyperbolic relaxed model is shown to converge to the original Serre–Green–Naghdi model with respect to a small relaxation parameter. The goal of this paper is to augment this hyperbolic relaxation technique with a finite-element approximation that is second-order accurate in space, invariant-domain preserving, and well balanced. We build on the work seen in [21] where a similar approximation technique is shown for a partial Serre model with incomplete topography effects. A major departure from [21] that we consider here is the construction of the numerical artificial viscosity. The way the artificial viscosity is defined in [21] makes the accuracy of the method limited to second-order at best with a loss of accuracy at extrema in the water height: it is second order on the water height in the L^1-norm but only first order in the L^∞-norm when approximating smooth solutions supported by the model such as a solitary waves and periodic waves. The goal of the present work is to go beyond the method described in [21] and include the full topography effects considered in [23]. More specifically, we make the proposed method higher order in space by adapting the commutator-based entropy-viscosity methodology introduced in [19]. In addition, the high-order method is made invariant-domain preserving (i.e., positivity-preserving) via a convex limiting process as seen [15, 20]. One question we address in this paper is how to handle source terms in the limiting process.

The paper is organized as follows. In Sect. 2, we introduce the Serre equations and the hyperbolic relaxation model as formulated in [23]. In this section, we discuss the properties of the hyperbolic system along with the mathematical treatment of the external source terms. The finite-element setting is introduced in Sect. 3. Then, in Sect. 4, we describe the low-order space/time approximation of the hyperbolic Serre model (2.3) which is an extension of the scheme introduced in [21] along with the numerical treatment of the external source terms. The key results regarding positivity-preserving and well-balanced properties of the low-order scheme are summarized in Proposition 4.4. In Sect. 5, we introduce a provisional higher order method using the entropy-viscosity technology that is possibly invariant-domain preserving violating. Then, in Sect. 6, we introduce the convex limiting technique that is used to make the provisional high-order method positivity-preserving with an emphasis on how to handle the source terms. The key results which show that the final limited update is invariant-domain preserving and well balanced are summarized in Theorem 6.6 and Proposition 6.8. The implementation details of the numerical method are discussed in Sect. 7.1. To verify reproducibility, the method is implemented with three different codes. The first code does not use any particular software and is written in Fortran 95/2003. The second code uses the `Proteus` toolkit (see [27]). Both codes use continuous $\mathbb{P}_1$ Lagrange elements. The third code, called `Ryujin`, is a high-performance finite-element solver based on the `deal.II` library and uses continuous $\mathbb{Q}_1$ elements

(see, e.g., [1, 33]). In Sect. 7.2, we verify the convergence rate of the numerical method using analytical solutions of the original Serre model (2.1). In Sect. 7.3, we verify that the method is numerically well balanced up to machine precision with a series of synthetic tests involving topography. Then, in Sect. 7.4, we consider an academic benchmark involving a friction source term and compare the results of the hyperbolic Serre model to the typical Saint-Venant shallow water equations. Finally, in Sect. 7.5, we validate the numerical results with data from several laboratory experiments.

2 Preliminaries

In this section, we recall the Serre equations with topography effects and the hyperbolic relaxation model introduced in [12, 23]. We also recall important properties of the hyperbolic system.

2.1 The Dispersive Serre–Green–Naghdi Model with Topography

Let D be a polygonal domain in $\mathbb{R}^d$, $d \in \{1, 2\}$, occupied by a body of water evolving in time under the action of gravity. Let $\boldsymbol{u} = (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ be the dependent variable, where h is the water height and $\boldsymbol{q}$ the momentum vector (also known as the flow discharge). The Serre model as first introduced in [39], and extended in [13, 38] to include topography effects, can be written as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q} + p(\boldsymbol{u})\mathbb{I}_d) = -r(\boldsymbol{u})\nabla z. \tag{2.1b}$$

Here, the given topography map (also known as the bathymetry) is represented by $z(\boldsymbol{x})$ and the pressure $p(\boldsymbol{u})$ and source $r(\boldsymbol{u})$ are defined by

$$p(\boldsymbol{u}) := \tfrac{1}{2}g\mathsf{h}^2 + \mathsf{h}^2\Big(\tfrac{1}{3}\ddot{\mathsf{h}} + \tfrac{1}{2}\dot{\mathsf{k}}\Big), \quad \dot{\mathsf{h}} := \partial_t \mathsf{h} + \boldsymbol{v}\cdot\nabla\mathsf{h}, \quad \ddot{\mathsf{h}} := \partial_t \dot{\mathsf{h}} + \boldsymbol{v}\cdot\nabla\dot{\mathsf{h}}, \tag{2.2a}$$

$$r(\boldsymbol{u}) = g\mathsf{h} + \mathsf{h}\Big(\tfrac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}}\Big), \qquad \dot{\mathsf{k}} := \partial_t(\boldsymbol{v}\cdot\nabla z) + \boldsymbol{v}\cdot\nabla(\boldsymbol{v}\cdot\nabla z). \tag{2.2b}$$

Here, $\boldsymbol{v}$ is the velocity vector field and is defined such that $\boldsymbol{q} := \mathsf{h}\boldsymbol{v}$. For brevity, we interchange the Serre–Green–Naghdi convention with just the Serre model.

Remark 2.1 (Saint-Venant shallow water equations) The Saint-Venant shallow water equations can be recovered from (2.1) to (2.2) by removing the dispersive effects from the definition of the pressure and from the topography source term; that is, when $p(\boldsymbol{u}) = \frac{1}{2}g\mathsf{h}^2$ and $r(\boldsymbol{u}) = g\mathsf{h}$. □

Remark 2.2 (Admissible set) The Serre model admits the important physical property that the water height h stays positive for $t \geq 0$. We define the set of admissible states for the Serre model by: $\mathcal{A} = \{\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T} \in \mathbb{R}^{d+1} \,|\, \mathsf{h} > 0\}$. Numerical methods that preserve such admissible sets are known to be positivity-preserving and can be classified as a subset of invariant-domain preserving schemes (see for example [17]).□

Remark 2.3 (Lake-at-rest property) When the fluid is at rest, i.e., when there is no flow discharge ($\boldsymbol{q} \equiv \boldsymbol{0}$), the Serre equations (2.1) and (2.2) reduce to the following partial differential equation: $g\mathsf{h}\nabla(\mathsf{h}+z) = \boldsymbol{0}$. This is the well-known lake-at-rest steady-state problem. Preserving solutions to this partial differential equation is essential for constructing numerical methods that are well balanced. □

It is well known that the non-hydrostatic pressure (2.2a) leads to a dispersive time step restriction when discretizing the equations in space in time. More specifically, any approximation technique that is explicit in time would require the time step τ to behave like $\mathcal{O}(\mathsf{h}^3)V^{-1}L^{-2}$, where h is the mesh size, V is a character wave speed scale and L a characteristic length scale. There are two popular methodologies in the literature for addressing this setback. The first method is based on splitting techniques (such as Strang's operator splitting) that combine both explicit and implicit time stepping (see [6, 10, 37]). The second method consists of reinterpreting the Serre equations as a constrained first-order system and then relaxing the constraints to obtain a hyperbolic system (see [4, 11, 12]). For the rest of the paper, we consider the hyperbolic relaxed system of the Serre model derived in [23].

2.2 The Hyperbolic Relaxation Model

Denoting $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^{\mathsf{T}}$ as the new conserved variable, we consider the hyperbolic relaxation of the Serre equations with topography effects introduced in [23]:

$$\partial_t \mathsf{h} + \nabla\cdot(\boldsymbol{v}\mathsf{h}) = 0, \tag{2.3a}$$
$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla p(\boldsymbol{u}) = -r(\boldsymbol{u})\nabla z, \tag{2.3b}$$
$$\partial_t q_1 + \nabla\cdot(\boldsymbol{v}q_1) = q_2 - \tfrac{3}{2}\boldsymbol{q}\cdot\nabla z, \tag{2.3c}$$
$$\partial_t q_2 + \nabla\cdot(\boldsymbol{v}q_2) = -s(\boldsymbol{u}), \tag{2.3d}$$
$$\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) = \widetilde{s}(\boldsymbol{u}), \tag{2.3e}$$
$$p(\boldsymbol{u}) := \tfrac{1}{2}g\mathsf{h}^2 + \widetilde{p}(\boldsymbol{u}), \quad \widetilde{p}(\boldsymbol{u}) := -\tfrac{1}{3}\tfrac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2\big(\eta\Gamma'(\tfrac{\eta}{\mathsf{h}}) - 2\mathsf{h}\Gamma(\tfrac{\eta}{\mathsf{h}})\big), \tag{2.3f}$$
$$r(\boldsymbol{u}) := g\mathsf{h} - \tfrac{1}{2}s(\boldsymbol{u}) + \tfrac{1}{4}\widetilde{s}(\boldsymbol{u}), \tag{2.3g}$$
$$s(\boldsymbol{u}) := \overline{\lambda}g\tfrac{\mathsf{h}^2}{\epsilon}\Gamma'(\tfrac{\eta}{\mathsf{h}}), \quad \widetilde{s}(\boldsymbol{u}) := \overline{\lambda}g\mathsf{h}_0\tfrac{\mathsf{h}}{\epsilon}\Phi(\tfrac{\boldsymbol{v}\cdot\nabla z-\beta}{\sqrt{g\mathsf{h}_0}}). \tag{2.3h}$$

for a.e. $(\boldsymbol{x}\in D), t\in\mathbb{R}_+$

$$\mathsf{h}(\boldsymbol{x},0) = \mathsf{h}_0(\boldsymbol{x}), \quad \boldsymbol{q}(\boldsymbol{x},0) = \boldsymbol{q}_0(\boldsymbol{x}), \quad q_1(\boldsymbol{x},0) = \mathsf{h}_0(\boldsymbol{x})^2, \tag{2.4a}$$
$$q_3(\boldsymbol{x},0) = \boldsymbol{q}_0(\boldsymbol{x})\cdot\nabla z, \quad q_2(\boldsymbol{x},0) = -\mathsf{h}_0^2(\boldsymbol{x})\nabla\cdot\boldsymbol{v}_0(\boldsymbol{x}) + \tfrac{3}{2}q_3(\boldsymbol{x},0). \tag{2.4b}$$

The dependent (or conserved) variables, are the water height h, the discharge $\boldsymbol{q}$, and the auxiliary variables q_1, q_2, q_3. We recall that q_1 is an ansatz for h^2, q_3 an ansatz for $\boldsymbol{q}\cdot\nabla z$, and q_2 an ansatz for $\mathsf{h}\dot{\mathsf{h}} + \frac{3}{2}q_3$ (or equivalently $-\mathsf{h}^2\nabla\cdot\boldsymbol{v} + \frac{3}{2}q_3$). Here, $\mathbb{I}_d$ is the $d\times d$ identity matrix, and the mapping $z : D \ni \boldsymbol{x} \mapsto z(\boldsymbol{x}) \in \mathbb{R}$ is the bottom

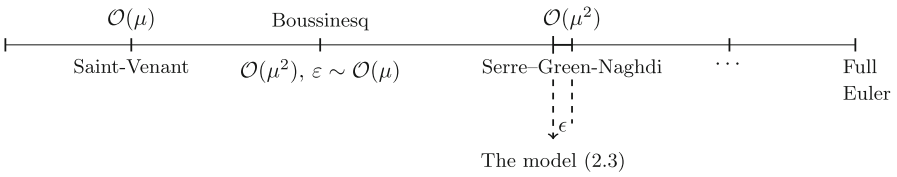


Fig. 1 A representation of the physical accuracy of the model (2.3) compared to the Saint-Venant, Boussinesq and Serre–Green–Naghdi model. Here ϵ is the relaxation parameter of the model

topography which we assume to be given. We adopt the same notation as in [23] and introduce the following primitive quantities $q_1 := \mathsf{h}\eta$, $q_2 := \mathsf{h}\omega$, $q_3 := \mathsf{h}\beta$ where η is thought of as ansatz for h, ω an ansatz for $\mathsf{h} + \frac{3}{2}\beta$ and β an ansatz for $\boldsymbol{v}\cdot\nabla z$.

The quantity ϵ is a small length scale and is the relaxation parameter introduced in [23]. It is shown in [23, Cor. (3.9)] that as $\epsilon \to 0$, the relaxed model (2.3) is a consistent approximation of Serre–Green–Naghdi model. This result is rigorously proved in [9] in the absence of topography; see Remark 1.6 therein. More precisely, it is established in [9] that, under reasonable assumptions, the difference between the solutions of the original system (2.1) and the perturbed system (2.3) goes to zero as fast as the non-dimensional quantity $\frac{1}{\overline{\lambda}}\frac{\epsilon}{\mathsf{h}_0}$. When the model is approximated in space in Sect. 3.1, the relaxation parameter ϵ will be replaced by the local mesh-size so that $\epsilon \to 0$ is analogous to the mesh-size decreasing. The symbol $\overline{\lambda}$ in (2.3f)–(2.3h) is a non-dimensional number of order one and is set to $\overline{\lambda} = 1$ for the rest of the paper. The function $\Gamma \in C^2(\mathbb{R}; [0, \infty))$ is a smooth non-negative function with the constraints $\Gamma(1) = 0$ and $\Gamma'(1) = 0$. Here, $\Phi \in C^0(\mathbb{R}; \mathbb{R})$ is a function such that $\xi\,\Phi(\xi) \geq 0$ for all $\xi \in \mathbb{R}$. In the applications reported at the end of the paper, we take $\Phi(\xi) = \xi$. In Fig. 1, we show a comparison of the physical accuracy of the relaxed model (2.3) with other common models. We now recall results for the system (2.3) established in [23] which will be used in Sect. 5.2 (this result is also established in [9, Sect. 2.3]).

Proposition 2.4 (Hyperbolicity) *Let $\mathbb{k}(\boldsymbol{u})$ be the conservative flux of the system (2.3). For any unit vector $\boldsymbol{n} \in \mathbb{R}^d$, the $d + 4$ eigenvalues of the Jacobian matrix of the flux $\mathbb{k}(\boldsymbol{u})\boldsymbol{n}$ are $\mu_k = \boldsymbol{v}\cdot\boldsymbol{n}$, $k \in \{2{:}d + 3\}$ and*

$$\mu_1 = \boldsymbol{v}\cdot\boldsymbol{n} - \sqrt{g\mathsf{h} + \partial_{\mathsf{h}}\widetilde{p}(\mathsf{h}, \eta)}, \quad \mu_{d+4} = \boldsymbol{v}\cdot\boldsymbol{n} + \sqrt{g\mathsf{h} + \partial_{\mathsf{h}}\widetilde{p}(\mathsf{h}, \eta)}. \tag{2.5}$$

The system (2.3) is hyperbolic iff the following holds for all $\eta \in \mathbb{R}$ and all $\mathsf{h} \geq 0$:

$$g\mathsf{h}\left(1 + \frac{1}{3}\frac{\overline{\lambda}}{\epsilon}\eta\left(x^3\partial_{xx}(x^{-2}\Gamma(x))\right)_{|x=\eta\mathsf{h}^{-1}}\right) \geq 0. \tag{2.6}$$

Lemma 2.5 (Energy results) *Let $\boldsymbol{u}$ be a smooth solution to (2.3). Then, the following holds true: $\partial_t E(\boldsymbol{u}) + \nabla\cdot(\boldsymbol{F}(\boldsymbol{u})) = \frac{1}{4}\widetilde{s}(\boldsymbol{u})(\beta - \boldsymbol{v}\cdot\nabla z) \leq 0$, with*

$$E(\boldsymbol{u}) := \tfrac{1}{2}g\mathsf{h}^2 + gz\mathsf{h} + \tfrac{1}{2}\mathsf{h}\boldsymbol{v}^2 + \tfrac{1}{6}\mathsf{h}\omega^2 + \tfrac{1}{8}\mathsf{h}\beta^2 + \tfrac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}}), \tag{2.7a}$$

$$\boldsymbol{F}(\boldsymbol{u}) := \boldsymbol{v}(E(\boldsymbol{u}) + p(\boldsymbol{u})). \tag{2.7b}$$

If the topography is flat, the following holds true: $\partial_t E_{flat} + \nabla\cdot(\boldsymbol{F}_{flat}) = 0$, *with*

$$E_{flat}(\boldsymbol{u}) := \tfrac{1}{2}gh^2 + \tfrac{1}{2}h\boldsymbol{v}^2 + \tfrac{1}{6}h\omega^2 + \tfrac{\overline{\lambda}g}{3\epsilon}h^3\Gamma(\tfrac{\eta}{h}), \tag{2.8a}$$

$$\boldsymbol{F}_{flat}(\boldsymbol{u}) := \boldsymbol{v}(E_{flat}(\boldsymbol{u}) + p(\boldsymbol{u})). \tag{2.8b}$$

Remark 2.6 (Definition of $\Gamma(x)$ function) For the rest of the paper, we take the function $\Gamma(x)$ to be

$$\Gamma(x) := \begin{cases} 3(1-x)^2 & \text{if } x \le 1, \\ (1+2x)(1-x)^2 & \text{if } 1 \le x. \end{cases} \tag{2.9}$$

□

Remark 2.7 (Saint-Venant from (2.3)) Note that when $\overline{\lambda} = 0$, $\widetilde{p}(\boldsymbol{u}) = 0$ in (2.3f). This decouples the mass and momentum equations in (2.3) from the evolution equations for q_1, q_2, q_3. This yields the classical Saint-Venant model. □

2.3 Physical Sources

In this section, we describe the mathematical formulation of the external physical sources that will be considered in this paper. For notation purposes, consider the condensed form of the system (2.3):

$$\partial_t \boldsymbol{u} + \nabla\cdot\mathbb{f}(\boldsymbol{u}) = \boldsymbol{R}(\boldsymbol{u}, \nabla z) + \boldsymbol{S}(\boldsymbol{u}).$$

Here, $\boldsymbol{R}(\boldsymbol{u}, \nabla z) := \big(0, -r(\boldsymbol{u})\nabla z, q_2 - \frac{3}{2}\boldsymbol{q}\cdot\nabla z, -s, \tilde{s}\big)^{\mathsf{T}}$ is henceforth referred to as the PDE source. The quantity $\boldsymbol{S}(\boldsymbol{u})$ represents the accumulation of the external physical sources described below. This term is henceforth referred to as the external source.

2.3.1 Gauckler–Manning Friction

We account for loss of discharge due to friction effects by adopting the Gauckler–Manning's friction law. The friction source is defined as follows:

$$\boldsymbol{S}_{\mathrm{F}}(\boldsymbol{u}) := \Big(0, -gn^2\mathsf{h}^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2}, 0, 0, 0\Big)^{\mathsf{T}}. \tag{2.10}$$

The parameter n is the Gauckler–Manning's roughness coefficient and has units $(\mathrm{m}^{\frac{\gamma-2}{2}}\,\mathrm{s})$. We take $\gamma = \frac{4}{3}$ in the computations reported below in Sect. 7.

2.3.2 Wave Generation and Absorption

In applications that involve the propagation of periodic waves, a common technique in the literature is to introduce relaxation zones in a numerical wave tank to smoothly

generate and absorb waves (see: [32, 43] and references therein). These generation and absorption zones are introduced as source terms in the equations.

For simplicity, let us assume we have a rectangular computational domain D. Assume that we want to generate uni-directional waves perpendicular to the inflow boundary so that wave profiles only depend on the x-direction (by convention x is the first Cartesian coordinate of the position vector $\boldsymbol{x}$). Let $\boldsymbol{u}_{\text{wave}}(\boldsymbol{x}, t)$ denote the theoretical wave profiles for each conserved variable. Denoting by h_{wave} and $q_{1,\text{wave}}$ the h and q_1 components of $\boldsymbol{u}_{\text{wave}}$, we assume that $\mathsf{h}_{\text{wave}}(\boldsymbol{x}, t) \geq 0$ and $q_{1,\text{wave}}(\boldsymbol{x}, t) \geq 0$ for all $\boldsymbol{x} \in D$ and all $t > 0$. To generate periodic waves through a relaxation zone, we introduce the following source:

$$\boldsymbol{S}_{\text{G}}(\boldsymbol{u}) := -\frac{\sqrt{g\mathsf{H}_0}}{\epsilon}(\boldsymbol{u} - \boldsymbol{u}_{\text{wave}}(\boldsymbol{x}, t))G\left(\tfrac{x - x_{\min}}{L_{\text{gen}}}\right), \tag{2.11}$$

where H_0 is the still water depth and $G(\xi)$ is a non-dimensional relaxation function defined as follows:

$$G(\xi) := \begin{cases} \frac{\exp\left(-|\log(\alpha)|\xi^2\right) - \alpha}{1-\alpha} & \text{if } \xi < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, L_{gen} is the length of the generation zone. In this paper, we take $\alpha := 0.005$.

We follow a similar methodology as above to absorb waves in a relaxation zone at the outflow boundary. The absorption zone is enforced via the following source term:

$$\boldsymbol{S}_{\text{A}}(\boldsymbol{u}) := \left(0, -\frac{\sqrt{g\mathsf{H}_0}}{\epsilon}G\left(\tfrac{x_{\max} - x}{L_{\text{abs}}}\right)\boldsymbol{q}, 0, -\frac{\sqrt{g\mathsf{H}_0}}{\epsilon}G\left(\tfrac{x_{\max} - x}{L_{\text{abs}}}\right)q_2, 0\right)^{\mathsf{T}}. \tag{2.12}$$

Note that now, instead of enforcing a theoretical wave profile, we are enforcing the zero value on $\boldsymbol{q}$ and q_2 to dissipate the waves.

3 Finite-Element Setting

In this section, we introduce the continuous finite-element setting used for the approximation of the hyperbolic Serre model (2.3). Note that the techniques shown here can be also be adapted using discontinuous finite elements and finite volumes as discussed in [20].

Let $(\mathcal{T}_h)_{h>0}$ be a shape-regular family of matching meshes where h can be thought of as the typical mesh-size. Given some mesh $\mathcal{T}_h$, we consider a scalar-valued finite-element space $P(\mathcal{T}_h)$ with global shape functions $\{\varphi_i\}_{i\in\mathcal{V}}$ associated with the Lagrange nodes $\{\boldsymbol{a}_i\}_{i\in\mathcal{V}}$. Note that $\dim(P(\mathcal{T}_h)) := \text{card}(\mathcal{V})$. The approximation in space of the conserved variable $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^{\mathsf{T}}$ is done in the space of $\mathbb{R}^{d+4}$-valued finite elements $\boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^{d+4}$. The bottom topography z is approximated in $P(\mathcal{T}_h)$. For every $i \in \mathcal{V}$, we call the stencil of the shape function, φ_i, the index set

$$\mathcal{I}(i) := \{j \in \mathcal{V} \mid \text{supp}(\varphi_i) \cap \text{supp}(\varphi_j) \neq \emptyset\}.$$

We also define $\mathcal{I}^*(i) := \mathcal{I}(i)\backslash\{i\}$. The following mesh-dependent quantities play an important role for the space and time approximation:

$$m_{ij} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x, \quad m_i := \int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}x, \tag{3.1a}$$

$$\boldsymbol{c}_{ij} := \int_D \varphi_i \nabla\varphi_j\,\mathrm{d}x, \quad \boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}, \tag{3.1b}$$

where $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. Here, m_{ij} are the entries of the consistent mass matrix and m_i the entries of the lumped mass matrix. By the partition of unity property ($\sum_{i\in\mathcal{V}}\varphi_i = 1$), we have that $m_i = \sum_{j\in\mathcal{I}(i)} m_{ij}$. The following three properties are essential to establish conservation: (1) $\sum_{j\in\mathcal{V}} \boldsymbol{c}_{ij} = \boldsymbol{0}$ (partition of unity property); (2) $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ if either φ_i or φ_j is zero on ∂D (integration by parts); (3) $\sum_{i\in\mathcal{V}} \boldsymbol{c}_{ij} = \boldsymbol{0}$ if φ_j is zero on ∂D (partition of unity property).

3.1 Finite-Element Representations

The finite-element approximation of the conserved variable at time t is denoted $\boldsymbol{u}_h(t) := (\mathsf{h}_h, \boldsymbol{q}_h, q_{1,h}, q_{2,h}, q_{3,h})^{\mathsf{T}}$ and represented as follows $\boldsymbol{u}_h(t) = \sum_{i\in\mathcal{V}} \mathbf{U}_i(t)\varphi_i$ in $\boldsymbol{P}(\mathcal{T}_h)$, where

$$\mathbf{U}_i(t) := (\mathsf{H}_i(t), \mathbf{Q}_i(t), \mathsf{Q}_{1,i}(t), \mathsf{Q}_{2,i}(t), \mathsf{Q}_{3,i}(t))^{\mathsf{T}}.$$

Here, $\mathsf{h}_h := \sum_{i\in\mathcal{V}} \mathsf{H}_i\varphi_i$ is the approximation of the water height, $\boldsymbol{q}_h := \sum_{i\in\mathcal{V}} \mathbf{Q}_i\varphi_i$ is approximation of the discharge, and $q_{1,h} := \sum_{i\in\mathcal{V}} \mathsf{Q}_{1,i}\varphi_i$, $q_{2,h} := \sum_{i\in\mathcal{V}} \mathsf{Q}_{2,i}\varphi_i$, $q_{3,h} := \sum_{i\in\mathcal{V}} \mathsf{Q}_{3,i}\varphi_i$ are the approximations of the three auxiliary variables. We denote by $z_h := \sum_{i\in\mathcal{V}} \mathsf{Z}_i\varphi_i \in \mathcal{P}(\mathcal{T}_h)$ the approximate bottom topography.

Let $\mathsf{H}_{0,\max}$ be some reference scale for the water height. For instance we can take $\mathsf{H}_{0,\max} := \operatorname{ess\,sup}_{\boldsymbol{x}\in D} \mathsf{h}_0(\boldsymbol{x})$, where h_0 is the initial water height. The approximate velocity $\boldsymbol{v}_h$ and the approximate auxiliary quantities $\eta_h, \omega_h, \beta_h$ are defined by regularization as follows for all $i \in \mathcal{V}$:

$$\mathbf{V}_i := \frac{\mathbf{Q}_i}{\mathsf{H}_i^\delta}, \quad \mathsf{N}_i := \frac{\mathsf{Q}_{1,i}}{\mathsf{H}_i^\delta}, \quad \mathsf{W}_i := \frac{\mathsf{Q}_{2,i}}{\mathsf{H}_i^\delta}, \quad \mathsf{B}_i := \frac{\mathsf{Q}_{3,i}}{\mathsf{H}_i^\delta}, \tag{3.2}$$

with

$$\mathsf{H}_i^\delta := \left(\frac{2\mathsf{H}_i}{\mathsf{H}_i^2 + \max(\mathsf{H}_i, \delta\mathsf{H}_{0,\max})^2}\right)^{-1} \tag{3.3}$$

where δ is a small dimensionless parameter. We take $\delta = 10^{-5}$ in the simulations reported at the end of the paper. We note that it is possible to take δ to be smaller, but in our experience this requires the CFL number to be smaller. Notice that the regularization is active only when dry state occurs, for example: $\mathbf{V}_i := \frac{1}{\mathsf{H}_i}\mathbf{Q}_i$ if $\mathsf{H}_i \geq$

$\delta \mathsf{H}_{0,\max}$. The reader is referred to [29, Eq. (2.17)], [8, Eq. (3.10)], and [3, 19, §5.1], where this technique is also adopted.

The relaxation parameter ϵ is chosen to be proportional to the local mesh size. Recalling that $m_i := \int_D \varphi_i \, \mathrm{d}x$ is proportional to the volume of the support of the shape function φ_i, we set $\epsilon_h := \sum_{i\in\mathcal{V}} \mathcal{E}_i \varphi_i$ with $\mathcal{E}_i := m_i^{\frac{1}{d}}$ (recall that $d \in \{1, 2\}$ is the space dimension).

4 The Low-Order Method

We now describe the low-order approximation of the hyperbolic Serre model (2.3) using the finite-element setting shown above. The method is formally first-order accurate in space and is presented with forward Euler time stepping. Higher order accuracy in time is achieved using any explicit strong stability preserving Runge–Kutta method.

4.1 Numerical Flux, PDE Source, Star States

Let $\boldsymbol{u}_h^0 := \sum_{i\in\mathcal{V}} \mathbf{U}_i^0 \varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be some reasonable approximation of the initial data $\boldsymbol{u}_0$ (see (2.4)) at time t^0. Let t^n, $n \in \mathbb{N}$, be the current time, and let $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{V}} \mathbf{U}_i^n \varphi_i \in \boldsymbol{P}(\mathcal{T}_h)$ be the current admissible approximation of $\boldsymbol{u}$.

For all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$, we define the numerical flux as follows:

$$\mathbf{F}_{ij}^n := \mathbf{U}_j^n(\mathbf{V}_j^n \cdot \boldsymbol{c}_{ij}) + \Big(0, \big(\widetilde{\mathsf{P}}(\mathbf{U}_j^n) + g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\big)\boldsymbol{c}_{ij}, 0, 0, 0\Big)^{\mathsf{T}}, \tag{4.1a}$$

$$\widetilde{\mathsf{P}}(\mathbf{U}) := -\frac{\overline{\lambda} g}{3\mathcal{E}} \times \begin{cases} 6\mathsf{H}(\mathsf{Q}_1 - \mathsf{H}^2), & \text{if } \mathsf{Q}_1 \le \mathsf{H}^2 \\ 2\frac{(\mathsf{Q}_1 - \mathsf{H}^2)}{\mathsf{H}^\delta}(\mathsf{N}^2 + \mathsf{Q}_1 + \mathsf{H}^2), & \text{if } \mathsf{H}^2 < \mathsf{Q}_1. \end{cases} \tag{4.1b}$$

Note that the hydrostatic pressure and the Saint-Venant topography source term are discretized as: $\nabla \frac{1}{2} g\mathsf{h}^2 + g\mathsf{h}\nabla z = g\mathsf{h}\nabla(\mathsf{h} + z)$ to ensure well balancing.

To approximate the PDE source term $\boldsymbol{R}(\boldsymbol{u}, \nabla z)$ in (2.3), we introduce the following quantities:

$$\mathsf{R}_1(\mathbf{U}, \nabla Z) := \mathsf{Q}_2 - \tfrac{3}{2}\mathbf{Q} \cdot \nabla Z, \tag{4.2a}$$

$$\mathsf{R}_2(\mathbf{U}) := \frac{\overline{\lambda} g}{\mathcal{E}} \times \begin{cases} 6(\mathsf{Q}_1 - \mathsf{H}^2), & \text{if } \mathsf{Q}_1 \le \mathsf{H}^2 \\ 6\mathsf{N}\frac{(\mathsf{Q}_1 - \mathsf{H}^2)}{\mathsf{H}^\delta}, & \text{if } \mathsf{H}^2 < \mathsf{Q}_1, \end{cases} \tag{4.2b}$$

$$\mathsf{R}_3(\mathbf{U}, \nabla Z) := \frac{\overline{\lambda}}{\mathcal{E}} \sqrt{g\mathsf{H}_{0,\max}}(\mathbf{Q} \cdot \nabla Z - \mathsf{Q}_3). \tag{4.2c}$$

For all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$, let $(\nabla Z)_i := \sum_{j\in\mathcal{I}(i)} Z_j \boldsymbol{c}_{ij}$ denote the approximate gradient of the topography map. Then, the discrete PDE source is defined by

$$\boldsymbol{R}_i^n := \Big(0, \Big(\tfrac{1}{2}\mathsf{R}_2(\mathbf{U}_i^n) - \tfrac{1}{4}\mathsf{R}_3(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i)\Big)\tfrac{1}{m_i}(\nabla Z)_i,$$

$$\left.\mathsf{R}_1(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i), -\mathsf{R}_2(\mathbf{U}_i^n), \mathsf{R}_3(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i)\right)^{\mathsf{T}}. \tag{4.3}$$

We define the following hydrostatic reconstructed star states $\mathbf{U}_i^{*,j,n}$ and $\mathbf{U}_j^{*,i,n}$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$:

$$\mathbf{U}_i^{*,j,n} := \frac{\mathsf{H}_i^{*,j}}{\mathsf{H}_i^{\delta}}\left(\mathsf{H}_i, \mathbf{Q}_i, \frac{\mathsf{H}_i^{*,j}}{\mathsf{H}_i^{\delta}}\mathsf{Q}_{1,i}, \mathsf{Q}_{2,i}, \mathsf{Q}_{3,i}\right)^{\mathsf{T}}, \tag{4.4a}$$

$$\mathsf{H}_i^{*,j} := \max(0, \mathsf{H}_i + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j)), \tag{4.4b}$$

which are essential for well balancing (see [2, 21]). The star states $\mathbf{U}_i^{*,j,n}$ and $\mathbf{U}_j^{*,i,n}$ are used in the definition of the artificial viscosity terms (see (4.11)).

4.2 External Source

We discretize the Gauckler–Manning friction source term (2.10) by setting

$$\boldsymbol{S}_{\mathrm{F}}(\mathbf{U}_i^n) := \frac{-2gn^2\mathbf{Q}_i^n\|\mathbf{V}_i\|_{\ell^2}}{(\mathsf{H}_i^n)^{\gamma} + \max((\mathsf{H}_i^n)^{\gamma}, 2gn^2\tau_n\|\mathbf{V}_i\|_{\ell^2})}. \tag{4.5}$$

Note that we introduced a regularization for the term $\mathsf{h}^{-\gamma}$ to avoid division by 0. This expression has been shown in [19] to be stable under the usual hyperbolic CFL time step restriction, i.e., no iterations or semi-implicit time stepping is needed to advance in time with this definition.

We discretize the wave generation source (2.11) as follows:

$$\boldsymbol{S}_{\mathrm{G}}(\mathbf{U}_i^n) := -\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}(\mathbf{U}_i^n - \boldsymbol{u}_{\mathrm{wave}}(\boldsymbol{a}_i, t^n))G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\mathrm{gen}}}), \tag{4.6}$$

The absorption zone source term (2.12) is approximated similarly.

4.3 Graph Viscosity and Time Step

We now define the low-order graph-viscosity coefficients that make the method positive. Just as in [21], we avoid solving the Riemann problem associated with the system (2.3) (since it is quite complicated) and set

$$\mu_{ij}^{\mathrm{L},n} := \max(|\boldsymbol{V}_i^n\cdot\boldsymbol{n}_{ij}|\|\boldsymbol{c}_{ij}\|_{\ell^2}, |\boldsymbol{V}_j^n\cdot\boldsymbol{n}_{ji}|\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{4.7}$$

$$d_{ij}^{\mathrm{L},n} := \max(\mu_{ij}^{\mathrm{L},n}, \max(\lambda_{ij}^n\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{ji}^n\|\boldsymbol{c}_{ji}\|_{\ell^2})), \tag{4.8}$$

where

$$\lambda_{ij}^n = \max(|\boldsymbol{V}_i^n\cdot\boldsymbol{n}_{ij} - (g\mathsf{H}_i^n + \theta_i^n)^{\frac{1}{2}}|, |\boldsymbol{V}_j^n\cdot\boldsymbol{n}_{ij} + (g\mathsf{H}_j^n + \theta_j^n)^{\frac{1}{2}}|) \tag{4.9}$$

with $\theta_i^n := \partial_{\mathsf{h}}\widetilde{p}(\mathsf{H}_i^n, \mathsf{N}_i^n)\big(\frac{\mathcal{E}_i}{\max(\mathcal{E}_i, \mathsf{H}_i^n)}\big)^2$. Note that by definition: $d_{ij}^{\mathrm{L},n} = d_{ji}^{\mathrm{L},n}$, $\mu_{ij}^{\mathrm{L},n} = \mu_{ji}^{\mathrm{L},n}$, $d_{ij}^{\mathrm{L},n} \geq \mu_{ij}^{\mathrm{L},n} \geq 0$, $i \neq j$.

We define the time step in the following way:

$$\tau_n := \mathrm{CFL} \times \max_{i\in\mathcal{V}} \left(\frac{m_i}{\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n}} \right), \tag{4.10}$$

where CFL is a user-defined positive constant.

4.4 Low-Order Update

Let us set $t^{n+1} := t^n + \tau_n$. Let $\boldsymbol{u}_h^{n+1} := \sum_{i\in\mathcal{V}} \mathbf{U}_i^{n+1}\varphi_i$ be the update of $\boldsymbol{u}$ at t^{n+1}. Let $\boldsymbol{S}_i^n$ denote the total contribution of the external sources at time t^n, i.e., $\boldsymbol{S}_i^n := \boldsymbol{S}_{\mathrm{F}}(\mathbf{U}_i^n) + \chi \boldsymbol{S}_{\mathrm{G}}(\mathbf{U}_i^n) + \boldsymbol{S}_{\mathrm{A}}(\mathbf{U}_i^n)$. Here, $\chi := 1$ if the wave generation source term is active and $\chi := 0$ otherwise. Then, the low-order update $\mathbf{U}_i^{n+1}$ for all $i \in \mathcal{V}$ is computed as follows:

$$\begin{aligned} \frac{m_i}{\tau_n}(\mathbf{U}_i^{L,n+1} - \mathbf{U}_i^n) &= m_i(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n) + \sum_{j\in\mathcal{I}(i)} -\mathbf{F}_{ij}^n \\ &+ \sum_{j\in\mathcal{I}^*(i)} \Big((d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n})\big(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\big) + \mu_{ij}^{\mathrm{L},n}\big(\mathbf{U}_j^n - \mathbf{U}_i^n\big) \Big). \end{aligned} \tag{4.11}$$

4.5 Well Balancing, Positivity, Conservation

We now show that the algorithm (4.11) is well balanced, positivity-preserving and conservative (up to the contribution of sources). We begin by recalling the precise definitions of the respective properties.

Definition 4.1 (*Exact rest*) A numerical state $(\mathsf{h}_h, \boldsymbol{q}_h, q_{1,h}, q_{2,h}, q_{3,h})^{\mathsf{T}}$ is said to be exactly at rest if $\boldsymbol{q}_h = \mathbf{0}$, $q_{2,h} = 0, q_{3,h} = 0$, $\mathsf{Q}_{1,i} = \mathsf{H}_i^2$, for all $i \in \mathcal{V}$, and if the approximate water height h_h and the approximate bathymetry map z_h satisfy the following alternative for all $i \in \mathcal{V}$: for all $j \in \mathcal{I}(i)$, either $\mathsf{H}_j = \mathsf{H}_i = 0$ or $\mathsf{H}_j + \mathsf{Z}_j = \mathsf{H}_i + \mathsf{Z}_i$.

Definition 4.2 (*Exactly well balanced*) A mapping $\boldsymbol{T} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ is said to be an exactly well-balanced scheme if $\boldsymbol{T}(\boldsymbol{u}_h) = \boldsymbol{u}_h$ when $\boldsymbol{u}_h$ is an exact rest state.

Definition 4.3 (*Positivity-preserving*) Let us denote $\mathsf{h}_h(\boldsymbol{u}_h) = \sum_{i\in\mathcal{V}} \mathsf{H}_i(\boldsymbol{u}_h)\varphi_i$ the water height of $\boldsymbol{u}_h$ for any $\boldsymbol{u}_h \in \boldsymbol{P}(\mathcal{T}_h)$. A mapping $\boldsymbol{T} : \boldsymbol{P}(\mathcal{T}_h) \to \boldsymbol{P}(\mathcal{T}_h)$ is said to be a positivity-preserving scheme if $\mathsf{H}_i(\boldsymbol{u}_h) \geq 0$, for all $i \in \mathcal{V}$, implies that $\mathsf{H}_i(\boldsymbol{T}(\boldsymbol{u}_h)) \geq 0$ for all $i \in \mathcal{V}$.

Proposition 4.4 *Let* $\boldsymbol{T} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ *be the scheme defined by* (4.11).

(i) *If* $\boldsymbol{S}_i^n = \boldsymbol{S}_F(\boldsymbol{U}_i^n)$ *is just the contribution of the Gauckler–Manning friction source, the scheme is exactly well balanced*;

(ii) *The scheme is positivity-preserving if the time step satisfies the following restriction* $\tau_n(\chi\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}+\frac{1}{m_i}\sum_{j\in\mathcal{I}^*(i)}d_{ij}^n)\le 1$.

Proof (i) Since $\boldsymbol{u}_h^n$ is exactly at rest, $\mu_{ij}^{\mathrm{L},n}=0$ for all $i\in\mathcal{V}$ and $j\in\mathcal{I}(i)$. We also have that $\mathbf{U}_i^{*,j,n}=\mathbf{U}_j^{*,i,n}$ as a consequence of the definition (4.4) (note that $\mathsf{Q}_{1,i}^{*,j}:=\left(\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\right)^2\mathsf{Q}_{1,i}$ needs to be defined as such for this to hold). Then, since $\boldsymbol{q}_h=0$ and $q_{2,h}=0$, we have that $\boldsymbol{S}_i^n=\mathbf{0}$, $\mathsf{R}_1(\mathbf{U}_i,(\nabla Z)_i)=\mathsf{R}_3(\mathbf{U}_i,(\nabla Z)_i)=0$ for all $i\in\mathcal{V}$. Note that $\mathsf{R}_2(\mathbf{U}_i)=0$ since $\mathbf{Q}_{1,i}^n=(\mathsf{H}_i^n)^2$. Thus, we have that $\boldsymbol{R}_i^n=\mathbf{0}$ for all $i\in\mathcal{V}$. The rest of the proof is exactly the same as [21, Prop. 4.4].

(ii) Referring to (2.11), we recall that $S_{\mathsf{h}}^n=-\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}(\mathsf{H}_i^n-\mathsf{h}_{\mathrm{wave}}(\boldsymbol{a}_i,t^n))G(\frac{\boldsymbol{a}_i-x_{\min}}{L_{\mathrm{gen}}})$ is the source in the mass balance equation, where $\mathsf{h}_{\mathrm{wave}}(\boldsymbol{x},t)\ge 0$ for all t and all $\boldsymbol{x}\in D$, and $G\in[0,1]$. Fixing $i\in\mathcal{V}$ and assuming $\mathsf{H}_j^n\ge 0$ for all $j\in\mathcal{I}(i)$, the water height update in (4.11) can be arranged as follows:

$$\mathsf{H}_i^{\mathrm{L},n+1}\ge\mathsf{H}_i^n-\chi\frac{\tau_n\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\mathsf{H}_i^n-\frac{1}{m_i}\sum_{j\in\mathcal{I}^*(i)}\Big(\mu_{ij}^{\mathrm{L},n}\mathsf{H}_i^n+(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\mathsf{H}_i^{*,j,n}\Big)$$
$$+\frac{1}{m_i}\sum_{j\in\mathcal{I}^*(i)}\Big((\mu_{ij}^{\mathrm{L},n}-\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij})\mathsf{H}_j^n+(d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n})\mathsf{H}_j^{*,i,n}\Big)$$

Since by definition $d_{ij}^{\mathrm{L},n}-\mu_{ij}^{\mathrm{L},n}\ge 0$, $\mu_{ij}^{\mathrm{L},n}\ge 0$, $\mathsf{H}_i^n\ge\mathsf{H}_i^{*,j,n}\ge 0$, $\mathsf{H}_j^{*,i,n}\ge 0$, we have the following inequality

$$\mathsf{H}_i^{\mathrm{L},n+1}\ge\mathsf{H}_i^n\Big(1-\chi\frac{\tau_n\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}-\frac{\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)}d_{ij}^{\mathrm{L},n}\Big)+\sum_{j\in\mathcal{I}^*(i)}\Big((\mu_{ij}^{\mathrm{L},n}-\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij})\mathsf{H}_j^n\Big).$$

The conclusion follows from the condition on τ_n and the definition (4.7). □

We now discuss the conservative properties of the scheme (4.11). Notice that when the topography is flat and there is no contribution of external sources (i.e., $\boldsymbol{S}_i^n\equiv 0$), there is still a contribution of the PDE source $\boldsymbol{R}_i^n$ in the update (4.11). More specifically, when the topography is flat we see that:

$$\boldsymbol{R}_i^n=\Big(0,\mathbf{0},\mathsf{R}_1(\mathbf{U}_i^n,0),-\mathsf{R}_2(\mathbf{U}_i^n),\mathsf{R}_3(\mathbf{U}_i^n,0)\Big).$$

This fact motivates the following definition.

Definition 4.5 (*Conservation with sources*) A mapping $\boldsymbol{T}:\boldsymbol{P}(\mathcal{T}_h)\to\boldsymbol{P}(\mathcal{T}_h)$ is said to be a conservative approximation of the system (2.3) if

$$\sum_{i\in\mathcal{V}}m_i\boldsymbol{T}(\mathbf{U})=\sum_{i\in\mathcal{V}}m_i\big(\mathbf{U}+\tau\boldsymbol{R}_i(\mathbf{U})\big).$$

Proposition 4.6 *Let* $\boldsymbol{T} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ *be the scheme defined by* (4.11). *Then, if the topography map is flat* (*i.e.,* $z(\boldsymbol{x}) \equiv z_0 \in \mathbb{R}$) *and there is no contribution of external forces* (*i.e.,* $\boldsymbol{S}_i^n \equiv 0$), *then* $\boldsymbol{T}$ *is conservative.*

Proof Recalling that $\sum_{j\in\mathcal{V}} \boldsymbol{c}_{ij} = \boldsymbol{0}$ by the partition of unity property, the low-order update (4.11) can be written as follows:

$$\frac{m_i\left(\mathbf{U}_i^{L,n+1} - \mathbf{U}_i^n\right)}{\tau_n} = m_i \boldsymbol{R}_i^n + \sum_{j\in\mathcal{I}(i)} \mathbb{F}_{ij}, \tag{4.12}$$

where

$$\begin{aligned}\mathbb{F}_{ij} = \mathbf{U}_j^n(\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}) + \mathbf{U}_i^n(\mathbf{V}_i^n\cdot\boldsymbol{c}_{ij}) + \Big(0, \big(\widetilde{\mathsf{P}}(\mathbf{U}_j^n) + \widetilde{\mathsf{P}}(\mathbf{U}_i^n) + g\mathsf{H}_i^n\mathsf{H}_j^n\big)\boldsymbol{c}_{ij}, 0, 0, 0\Big)^{\mathsf{T}}\\ + \Big((d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n})\big(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\big) + \mu_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)\Big).\end{aligned}$$

Since $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ij}$ if φ_i or φ_j are zero on ∂D and by definition $d_{ij}^{\mathrm{L},n} = d_{ji}^{\mathrm{L},n}$,$\mu_{ij}^{\mathrm{L},n} = \mu_{ji}^{\mathrm{L},n}$, then $\mathbb{F}_{ij} = -\mathbb{F}_{ji}$. Summing (4.12) over $i \in \mathcal{V}$ gives that the scheme (4.11) is conservative up to the contribution of sources. □

4.6 Local Auxiliary States and Bounds

We now define auxiliary states and extract exact local bounds that will be useful when limiting the yet to be defined high-order solution which might not be positivity-preserving.

The key idea behind defining the exact local bounds is noticing that the low-order update (4.11) can be rewritten as a convex combination of auxiliary states. This is summarized in the following lemma.

Lemma 4.7 (Convex combination) *Let* $\boldsymbol{W}_i^{\mathrm{L},n+1} := \boldsymbol{U}_i^{\mathrm{L},n+1} - \tau_n\widetilde{\boldsymbol{R}}_i^n$, *with the modified source given by*

$$\widetilde{\boldsymbol{R}}_i^n := \boldsymbol{R}_i^n + \boldsymbol{S}_i^n + \left(0, \sum_{j\in\mathcal{I}(i)} g\left(-H_i^n Z_j + \tfrac{1}{2}(H_j^n - H_i^n)^2\right)\boldsymbol{c}_{ij}, 0, 0, 0\right)^{T}. \tag{4.13}$$

Assume $1 - \frac{2\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n} \geq 0$. (i) *Then, the following convex combination holds true*:

$$\boldsymbol{W}_i^{\mathrm{L},n+1} = \boldsymbol{U}_i^n\left(1 - \frac{\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} 2d_{ij}^{\mathrm{L},n}\right) + \frac{\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} 2d_{ij}^{\mathrm{L},n}\left(\overline{\boldsymbol{U}_{ij}^n} + \widetilde{\boldsymbol{U}_{ij}^n}\right). \tag{4.14}$$

with the auxiliary states defined by

$$\overline{\boldsymbol{U}_{ij}^n} = -\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}\cdot(\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n)) + \frac{1}{2}(\boldsymbol{U}_j^n + \boldsymbol{U}_i^n), \tag{4.15a}$$

$$\widetilde{\boldsymbol{U}_{ij}^n} = \frac{d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n}}{2d_{ij}^{\mathrm{L},n}}(\boldsymbol{U}_j^{*,i,n} - \boldsymbol{U}_j^n - (\boldsymbol{U}_i^{*,j,n} - \boldsymbol{U}_i^n)). \tag{4.15b}$$

(ii) *Furthermore*, $\overline{H_{ij}^n} + \widetilde{H_{ij}^n} \geq 0$ *for all* $j \in \mathcal{I}(i)$.

Notice that the quantity $\mathbf{W}_i^{\mathrm{L},n+1}$ is an update corresponding to solving the hyperbolic system without sources. The "source removing" concept is used in Sect. 6 to perform the convex limiting.

We now define the bounds that we use to limit the provisional higher order solution. The strategy that we propose consists of enforcing bounds that are naturally satisfied by the low-order update (4.14). More precisely, let us set

$$\mathsf{h}_i^{n,\min} := \min_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}), \quad \mathsf{h}_i^{n,\max} := \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}), \tag{4.16a}$$

$$q_{1,i}^{n,\min} := \min_{j\in\mathcal{I}(i)}(\overline{\mathsf{Q}_{1,ij}^n} + \widetilde{\mathsf{Q}_{1,ij}^n}), \quad q_{1,i}^{n,\max} := \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{Q}_{1,ij}^n} + \widetilde{\mathsf{Q}_{1,ij}^n}), \tag{4.16b}$$

$$\mathsf{K}_i^{n,\max} := \max_{j\in\mathcal{I}(i)}\psi(\overline{\mathbf{U}_{ij}^n} + \widetilde{\mathbf{U}_{ij}^n}). \tag{4.16c}$$

Here, the functional $\psi(\boldsymbol{u}) := \frac{1}{2}\frac{1}{\mathsf{h}(\boldsymbol{u})}\|\boldsymbol{q}(\boldsymbol{u})\|_{\ell^2}^2$ is the kinetic energy. Notice that the bounds are defined to be local in space and time.

Let us expand on the relationship between the bounds (4.16) and the update (4.14) with an example. We denote the components of the low-order solutions without sources, $\mathbf{W}^{\mathrm{L}}$, as follows:

$$(\mathsf{H}(\mathbf{W}^{\mathrm{L}}), \mathbf{Q}(\mathbf{W}^{\mathrm{L}}), \mathsf{Q}_1(\mathbf{W}^{\mathrm{L}}), \mathsf{Q}_2(\mathbf{W}^{\mathrm{L}}), \mathsf{Q}_3(\mathbf{W}^{\mathrm{L}}))^{\mathsf{T}}.$$

We can extract the following inequality on the water height update $\mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})$ as a direct consequence of the convex combination (4.14):

$$\mathsf{h}_i^{n,\min} \leq \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) \leq \mathsf{h}_i^{n,\max},$$

or equivalently:

$$\min_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}) \leq \mathsf{H}_i^{\mathrm{L},n+1} - \tau_n S_{\mathsf{h}}^n \leq \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}).$$

More precise statements are made in Sect. 6.1.

5 Provisional High-Order Method

We introduce in this section a provisional higher order method (second-order accurate in space) that may violate the invariant-domain preserving property. The two key ideas are as follows: (i) we reduce numerical dispersive errors induced by the lumped mass matrix; (ii) we define higher order graph viscosities $d_{ij}^{\mathrm{H},n}$, $\mu_{ij}^{\mathrm{H},n}$ via the estimation of an entropy residual/commutator.

5.1 Wave Generation

If active, the wave generation mechanism must be tempered in the high-order method since this source can potentially create dry states if the amplitude of the wave is too large. We formalize this by setting $\mathsf{h}_{\mathrm{wave}}^{\min} := \min_{\boldsymbol{x}\in D, t>0} \mathsf{h}_{\mathrm{wave}}(\boldsymbol{x}, t)$ and by introducing the cutoff function $\chi \in C(\mathbb{R}; [0, 1])$ defined by $\chi(\xi) = 1$ if $\xi \leq \frac{1}{2}$, $\chi(\xi) := 4(\xi - 1)^2(4\xi - 1)$ if $\frac{1}{2} \leq \xi \leq 1$, and $\chi(\xi) = 0$ otherwise. We then redefine the source term $\boldsymbol{S}_i^n$ used in the low-order approximation (4.11) by setting

$$\boldsymbol{S}_i^n := \boldsymbol{S}_{\mathrm{F}}(\mathbf{U}_i^n) + \chi\left(\frac{\mathsf{h}_i^{i,\max} - \mathsf{h}_i^{i,\min}}{\mathsf{h}_{\mathrm{wave}}^{\min}}\right)\boldsymbol{S}_{\mathrm{G}}(\mathbf{U}_i^n) + \boldsymbol{S}_{\mathrm{A}}(\mathbf{U}_i^n). \tag{5.1}$$

We also use this definition for the high-order update (see (5.8)). For most realistic applications, the amplitude of the incoming waves is of reasonable size and $\mathsf{h}_i^{n,\max} - \mathsf{h}_i^{n,\min}$ is a priori small compared to $\mathsf{h}_{\mathrm{wave}}^{\min}$ and the cutoff is therefore inactive. In particular, it is never active in the simulations reported below. Notice though that the cutoff is necessary for theoretical purposes (see Theorem 6.6).

5.2 Commutator-Based Entropy Viscosity

We present the definition of the higher order artificial viscosity coefficients $d_{ij}^{\mathrm{H},n}$, $\mu_{ij}^{\mathrm{H},n}$ following the method introduced in [19]. The key idea consists of measuring the smoothness of an entropy by measuring how well the chain rule is satisfied by the discretization described above.

Let $(E(\boldsymbol{u}), \boldsymbol{F}(\boldsymbol{u}))$ be the pair defined in (2.7a). Recall that by definition this pair satisfies the following relation:

$$\nabla\cdot(\boldsymbol{F}(\boldsymbol{u})) = (\nabla E(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot(\mathbb{f}(\boldsymbol{u})), \tag{5.2}$$

where $\mathbb{f}(\boldsymbol{u})$ is the flux for the hyperbolic Serre model (2.3). We want to estimate the entropy production by inserting the approximate solution in (5.2). For all $i \in \mathcal{V}$ we define the entropy commutator as follows:

$$C_i^n := \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\Big(\boldsymbol{F}(\mathbf{U}_j^n) - (\nabla E(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n)\Big). \tag{5.3}$$

This quantity measures how well the finite-element approximation satisfies the chain rule (5.2). Notice that when the approximate solution $\boldsymbol{u}_h^n$ is smooth, the quantity C_i^n is as small as the truncation error provided by the finite-element setting. For piecewise linear elements on unstructured meshes C_i^n scales like $\mathcal{O}(h)$ where h is the mesh-size. We define the normalized entropy residual/commutator to be

$$R_i^n := \frac{|C_i^n|}{D_i^n}, \tag{5.4a}$$

$$D_i^n := \Big| \sum_{i\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\boldsymbol{F}(\mathbf{U}_j^n)\Big| + \Big| \sum_{i\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot((\nabla E(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\Big|, \tag{5.4b}$$

where D_i^n is the rescaling factor. We then define the higher order graph viscosity (or entropy viscosity) as follows:

$$d_{ij}^{\mathrm{H},n} = d_{ij}^{\mathrm{L},n}\max(R_i^n, R_j^n), \quad d_{ii}^{\mathrm{H},n} := -\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{H},n} \tag{5.5}$$

$$\mu_{ij}^{\mathrm{H},n} = \mu_{ij}^{\mathrm{L},n}\max(R_i^n, R_j^n), \quad \mu_{ii}^{\mathrm{H},n} := -\sum_{j\in\mathcal{I}^*(i)} \mu_{ij}^{\mathrm{H},n}. \tag{5.6}$$

Notice that $R_i^n \in [0,1]$. Denoting by $\mathrm{diam}(D)$ the diameter of D, it is argued in [15] that $R_i^n = \mathcal{O}(h/\mathrm{diam}(D))$ when the solution is smooth. Thus, by making the high-order graph viscosities proportional to the entropy production, (5.5) and (5.6), we have $d_{ij}^{\mathrm{H},n} \sim d_{ij}^{\mathrm{L},n}$ when the entropy production is large, for instance in shock regions, and $d_{ij}^{\mathrm{H},n} \sim \mathcal{O}\big(\frac{h}{\mathrm{diam}(D)}\big) d_{ij}^{\mathrm{L},n}$ in regions where the approximate solution is smooth.

Remark 5.1 (Alternative options for entropy pair) We note that the choice of entropy pair for the above process is not unique. Actually, any entropy pair that satisfies a chain rule relation like (5.2) suffices. For instance, we also use the entropy pair for the Saint-Venant shallow water equations in some of the numerical illustrations reported below. More precisely, letting $\boldsymbol{u}_{\mathrm{SV}} := (\mathsf{h}, \boldsymbol{q})^{\mathsf{T}}$, the following pair

$$E_{\mathrm{SV}}(\boldsymbol{u}) := \frac{1}{2}g\mathsf{h}^2 + \frac{1}{2}\mathsf{h}\boldsymbol{v}^2, \tag{5.7a}$$

$$\boldsymbol{F}_{\mathrm{SV}}(\boldsymbol{u}) := \boldsymbol{v}\left(E_{\mathrm{SV}}(\boldsymbol{u}) + \frac{1}{2}g\mathsf{h}^2\right), \tag{5.7b}$$

satisfies the chain rule $\nabla\cdot(\boldsymbol{F}_{\mathrm{SV}}(\boldsymbol{u}_{\mathrm{SV}})) = (\nabla E_{\mathrm{SV}}(\boldsymbol{u}_{\mathrm{SV}}))^{\mathsf{T}}\nabla\cdot(\mathbb{f}_{\mathrm{SV}}(\boldsymbol{u}_{\mathrm{SV}}))$ where $\mathbb{f}_{\mathrm{SV}}(\boldsymbol{u}_{\mathrm{SV}})$ is Saint-Venant flux in (5.2). In Sect. 7.2, we show that the convergence behavior of the numerical method with either entropy pair, (2.8) or (5.7), is similar. □

5.3 Consistent Mass Matrix

Numerical dispersion errors can be significantly reduced using the consistent mass matrix for the discretization of the time derivative (at least for piecewise linear approximation). For more details on this, we refer the reader to [16] and the references therein.

We now replace the lumped mass matrix in (4.11) with the consistent mass matrix defined in (3.1) and the low-order graph viscosities in (4.11) with the entropy-viscosity coefficients (5.5) and (5.6). Then, the provisional higher order update is given as follows:

$$\sum_{j\in\mathcal{I}(i)} m_{ij}\frac{\widetilde{\mathbf{U}}_j^{\mathrm{H},n+1}-\mathbf{U}_j^n}{\tau_n} = m_i(\boldsymbol{R}_i^n+\boldsymbol{S}_i^n)+\sum_{j\in\mathcal{I}(i)}-\mathbf{F}_{ij}^n + \sum_{j\in\mathcal{I}^*(i)}\Big((d_{ij}^{\mathrm{H},n}-\mu_{ij}^{\mathrm{H},n})(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n})+\mu_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\Big). \tag{5.8}$$

Finding $\widetilde{\mathbf{U}}^{\mathrm{H},n+1}$ in (5.8) requires the inversion of the consistent mass matrix at every time step. Since this may be computationally costly, we follow the ideas of [16] and [33, Sec. (3.4)], and approximate the inverse of the mass matrix with a Neumann series. We do this as follows. We denote by $\widetilde{\boldsymbol{S}}_i^n$ the right-hand side in (5.8) and rewrite (5.8) as

$$\sum_{j\in\mathcal{I}(i)}\frac{m_{ij}}{m_j}\frac{m_j}{\tau_n}(\widetilde{\mathbf{U}}_j^{\mathrm{H},n+1}-\mathbf{U}_j^n)=\widetilde{\boldsymbol{S}}_i^n. \tag{5.9}$$

We then approximate the inverse $(\frac{m_{ij}}{m_j})^{-1}$ with the first-order approximation of its Neumann series representation:

$$(\frac{m_{ij}}{m_j})^{-1}=\left(\delta_{ij}-(\delta_{ij}-\frac{m_{ij}}{m_j})\right)^{-1}\approx\delta_{ij}+(\delta_{ij}-\frac{m_{ij}}{m_j})=\delta_{ij}+b_{ij}.$$

Then, using that $\sum_{j\in\mathcal{I}(i)} b_{ji}=0$, we infer the following new expression for the provisional higher order update $\mathbf{U}^{\mathrm{H},n+1}$: $\frac{m_i}{\tau_n}(\mathbf{U}_i^{\mathrm{H},n+1}-\mathbf{U}_i^n)=\widetilde{\boldsymbol{S}}_i^n+\sum_{j\in\mathcal{I}(i)}(b_{ij}\widetilde{\boldsymbol{S}}_j^n-b_{ji}\widetilde{\boldsymbol{S}}_i^n)$. Replacing the definition of $\widetilde{\boldsymbol{S}}_i^n$ therein gives

$$\frac{m_i}{\tau_n}(\mathbf{U}_i^{\mathrm{H},n+1}-\mathbf{U}_i^n)=m_i(\boldsymbol{R}_i^n+\boldsymbol{S}_i^n)+\sum_{j\in\mathcal{I}(i)}-\mathbf{F}_{ij}^n + \sum_{j\in\mathcal{I}^*(i)}\Big(b_{ij}\widetilde{\boldsymbol{S}}_j^n-b_{ji}\widetilde{\boldsymbol{S}}_i^n+(d_{ij}^{\mathrm{H},n}-\mu_{ij}^{\mathrm{H},n})(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n})+\mu_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\Big). \tag{5.10}$$

5.4 Loss of Positivity

It is proved in [18, Theorm 3.2] that the presence of the consistent mass matrix in any scheme that uses continuous finite elements based on artificial viscosity and explicit time stepping violates the maximum principle for scalar conservation laws. A consequence of this result is that the scheme (5.8) is non-positivity-preserving regardless of the definition of the artificial viscosity coefficients. It is also observed numerically in [19] that the use of the higher order entropy-viscosity coefficients (5.5) and (5.6) in (4.11) can cause the scheme to be non-positivity-preserving as well. We correct the loss of positivity in the following section.

6 Convex Limiting with Sources

In this section, we describe the convex limiting technique that is used to make the higher order methods described above positivity-preserving. Building on the ideas presented in [15, 19, 20], we give an emphasis on how to apply the convex limiting methodology to a hyperbolic system with source terms since it not well documented in the literature.

6.1 Quasiconcave Functionals and Bounds

We now give some definitions and results that will illustrate the notion of *convex* limiting.

Definition 6.1 (*Quasiconcavity*) Given a convex set $\mathcal{B} \subset \mathbb{R}^{d+4}$, we say that a function $\Psi : \mathcal{B} \to \mathbb{R}$ is quasiconcave if every upper level set of Ψ is *convex*; that is, the set $L_\lambda(\Psi) := \{\boldsymbol{u} \in \mathcal{B} \mid \Psi(\boldsymbol{u}) \geq \lambda\}$ is convex for any $\lambda \in \mathbb{R}$.

Lemma 6.2 *Let $\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid h > 0\} \subset \mathbb{R}^{d+4}$. Let $\Psi : \mathcal{B} \to \mathbb{R}$ and assume that the product $h\Psi$ is concave. Then, the function Ψ is quasiconcave.*

Proof This is a special case of the result in [20, Lem. 7.4]. □

Recall that the conserved variable for the system (2.3) is $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T}$ where h is the water height, $\boldsymbol{q}$ the momentum, q_1, q_2, q_3 the auxiliary variables which are thought of ansatz to h^2, $\mathsf{h}\dot{\mathsf{h}} + \frac{3}{2}q_3$ and $\boldsymbol{q}\cdot\nabla z$, respectively. The functional $\Psi_1 : \mathbb{R}^{d+4} \ni (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T} \mapsto \mathsf{h} \in \mathbb{R}$ is linear, hence concave, hence quasiconcave; this functional is also well defined over $\mathbb{R}^{d+4}$. The functional $\Psi_2 : \mathbb{R}^{d+4} \ni (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T} \mapsto q_1 \in \mathbb{R}$ is also linear, hence quasiconcave. Let us set

$$\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\}. \tag{6.1}$$

Observe that $\mathcal{A}$ is convex. Then, another important example is the (negative) kinetic energy $\Psi_3 : \mathcal{A} \ni (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T} \mapsto -\frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2$. Since the function $\mathsf{h}\Psi_3 := -\frac{1}{2}\|\boldsymbol{q}\|_{\ell^2}^2$ is concave, using Lemma 6.2 we conclude the (negative) kinetic energy

is quasiconcave. We are going to use the above functionals and bounds defined in (4.16) to enforce positivity of the water height, positivity of the auxiliary variable q_1, and a local maximum principle on the kinetic energy.

The key idea behind the convex limiting technique is to correct (i.e., limit) the high-order update so that it satisfies the same quasiconcave constraints as the low-order solution. Letting $\mathcal{L} := \{1{:}5\}$ and $\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\} \subset \mathbb{R}^{d+4}$, we are going to work with the family of quasiconcave functionals $\{\Psi_l^{i,n}\}_{l\in\mathcal{L}}$, $\Psi_l^{i,n} : \mathcal{A} \to \mathbb{R}$ defined as follows:

$$\Psi_1^{i,n}(\boldsymbol{u}) = \mathsf{h} - \mathsf{h}_i^{n,\min}, \quad \Psi_2^{i,n}(\boldsymbol{u}) = \mathsf{h}_i^{n,\max} - \mathsf{h}, \tag{6.2a}$$

$$\Psi_3^{i,n}(\boldsymbol{u}) = q_1 - q_{1,i}^{n,\min}, \quad \Psi_4^{i,n}(\boldsymbol{u}) = q_{1,i}^{\max} - q_1, \tag{6.2b}$$

$$\Psi_5^{i,n}(\boldsymbol{u}) = \mathsf{K}_i^{n,\max} - \frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2. \tag{6.2c}$$

The following result is essential for the rest of the convex limiting argumentation.

Lemma 6.3 *Let $n \geq 0$, $i \in \mathcal{V}$, and assume that $\frac{2\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n} \leq 1$. Then, the low-order update $\boldsymbol{W}_i^{\mathrm{L},n+1}$ computed by (4.14) is in $\mathcal{A}$ and satisfies the following constraints for all $l \in \mathcal{L}$:*

$$\Psi_l^{i,n}(\boldsymbol{W}_i^{\mathrm{L},n+1}) \geq 0. \tag{6.3}$$

Proof Under the assumption $1 - \frac{2\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n} \geq 0$, $\mathbf{W}_i^{\mathrm{L},n+1}$ is a convex combination of the auxiliary states (4.15a) and (4.15b); thus, by Lemma 4.7, the update $\mathbf{W}_i^{\mathrm{L},n+1}$ is in $\mathcal{A}$. The constraints $\Psi_l^{i,n}(\mathbf{W}_i^{\mathrm{L},n+1}) \geq 0$ are a consequence of the convex combination (4.14), the definitions of the bounds in (4.16), and quasiconcavity. □

The limiting is done sequentially: First we limit $\mathbf{W}_i^{\mathrm{H},n+1}$ with respect to $\Psi_1^{i,n}$ and construct a $\mathbf{W}_i^{1,n+1}$ so that $\Psi_1^{i,n}(\mathbf{W}_i^{1,n+1}) \geq 0$. This guarantees positivity of the water height and must be computed before limiting with respect to the other quantities $(\Psi_l^{i,n})_{l>1}$. This is explained in more detail below.

6.2 Limiting Process

We discuss in this section the proposed convex limiting methodology. The idea going forward is that we apply the limiting process to the solution without sources $\mathbf{W}^n$ and then "put back" the sources after enforcing the quasiconcave constraints.

Let $\mathbf{W}^{\mathrm{H},n+1} := \mathbf{U}^{\mathrm{H},n+1} - \tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n)$ be the provisional high-order update without sources. Our goal is to construct the final update $\mathbf{U}_i^{n+1}$ so that $\mathbf{W}_i^{n+1} := \mathbf{U}_i^{n+1} - \tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n)$ satisfies all the constraints $\Psi_l^{i,n}(\mathbf{W}_i^{n+1}) \geq 0$, $l \in \mathcal{L}$, defined in (6.2). For this purpose, we also define the low-order update without sources, $\mathbf{W}^{\mathrm{L},n+1} := \mathbf{U}^{\mathrm{L},n+1} - \tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n)$. Proceeding as in the Flux-Corrected-Transport methodology, we now compute the difference $\mathbf{W}^{\mathrm{H},n+1} - \mathbf{W}^{\mathrm{L},n+1}$ by subtracting (4.11) from (5.10).

This gives

$$m_i(\mathbf{W}_i^{\mathrm{H},n+1} - \mathbf{W}_i^{\mathrm{L},n+1}) = \sum_{j\in\mathcal{I}^*(i)} \mathbf{A}_{ij}^n, \tag{6.4}$$

with the $\mathbb{R}^{d+4}$-valued coefficients $\mathbf{A}_{ij}^n$ defined by

$$\begin{aligned}\mathbf{A}_{ij}^n := \tau_n \Big[& b_{ij}\widetilde{\mathbf{S}}_j^n - b_{ji}\widetilde{\mathbf{S}}_i^n + \big((d_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{H},n}) - (d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n})\big)(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) \\ & + (\mu_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n)\Big].\end{aligned} \tag{6.5}$$

Notice that $\mathbf{A}_{ij}^n = -\mathbf{A}_{ji}^n$, which implies global mass conservation $\sum_{i\in\mathcal{V}} m_i \mathbf{W}_i^{\mathrm{H},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{W}_i^{\mathrm{L},n+1}$ (i.e., $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{H},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{L},n+1}$); that is to say, the high-order solution and low-order solution have the same mass whether the source term is present or not.

Using (6.4), we introduce the final limited update as follows:

$$\mathbf{W}_i^{n+1} = \sum_{j\in\mathcal{I}^*(i)} \theta_j \Big(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n\Big), \quad \text{with} \quad \mathbf{P}_{ij}^n := \frac{1}{m_i\theta_j}\mathbf{A}_{ij}^n, \tag{6.6}$$

where $\{\theta_j\}_{j\in\mathcal{I}^*(i)}$ is any set of strictly positive coefficients adding up to 1. In the computations reported below, we take $\theta_j := \frac{1}{\mathrm{card}(\mathcal{I}(i))-1}$. The parameter $\ell_{ij} \in [0,1]$, which we call the limiter, is defined to be symmetric $\ell_{ij} = \ell_{ji}$ to preserve the mass conservation property mentioned above. Note that $\mathbf{W}_i^{n+1} = \mathbf{W}_i^{\mathrm{L},n+1}$ if $\ell_{ij} = 0$ (i.e., $\mathbf{U}_i^{n+1} = \mathbf{U}_i^{\mathrm{L},n+1}$) and $\mathbf{W}_i^{n+1} = \mathbf{W}_i^{\mathrm{H},n+1}$ if $\ell_{ij} = 1$. The key idea is to find a set of limiters $\ell_{ij} \in [0,1]$ as large as possible so that $\Psi_l^{l,n}(\mathbf{W}_i^{n+1}) \geq 0$ for all $l \in \mathcal{L}$. Notice that this optimization program is possible since $\ell_{ij} = 0$ is in the feasible set owing to Lemma 6.3. The following lemma proved in [15, Lem. 4.4] is paramount for the convex limiting technique and sums up how to efficiently find the limiting parameters ℓ_{ij}.

Lemma 6.4 *Let $\mathcal{A} \subset \mathbb{R}^{d+4}$ and $\Psi \in C^0(\mathcal{A};\mathbb{R})$ be such that $\{\boldsymbol{u} \in \mathcal{A} \mid \Psi(\boldsymbol{u}) \geq 0\}$ is convex. Let $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. Assume that $\boldsymbol{W}_i^{\mathrm{L},n+1} \in \mathcal{A}$, $\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1}) \geq 0$, and $\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \boldsymbol{P}_{ij}^n) < 0$ (otherwise there is nothing to limit), then*

(i) *There is a unique $\ell_j^i \in [0,1]$ such that*

$$\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell_j^i \boldsymbol{P}_{ij}^n) = 0, \tag{6.7}$$

$\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell \boldsymbol{P}_{ij}^n) \geq 0$ *for all* $\ell \in [0, \ell_j^i]$, *and* $\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell \boldsymbol{P}_{ij}^n) < 0$ *for all* $\ell \in (\ell_j^i, 1]$.

(ii) *Setting $\ell_{ij} = \min(\ell_j^i, \ell_i^j)$, we have $\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell_{ij}\boldsymbol{P}_{ij}^n) \geq 0$ and $\ell_{ij} = \ell_{ji}$.*

(iii) *Let* $\boldsymbol{W}_i^{n+1}$ *be defined by* (6.6), *then* $\Psi(\boldsymbol{W}_i^{n+1}) \geq 0$.

6.3 Application to the System (2.3)

We now illustrate Lemma 6.4 with $\Psi := \Psi_l^{i,n}$, $l \in \mathcal{L}$, defined in (6.2) and $\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\}$. The limiting is implemented by traversing $\mathcal{L}$ from the smallest index to the largest one.

We begin with the limiting of the water height. To avoid divisions by zero, we introduce the small parameter $\delta \mathsf{h}_i^{n,\max}$ where $\delta := 10^{-14}$ for all $i \in \mathcal{V}$. Let us denote the h-component of $\mathbf{P}_{ij}$ by $\mathsf{P}_{ij}^{\mathsf{h}}$. Then, we set:

$$\ell_j^{i,\mathsf{h}} = \begin{cases} \min\left(\dfrac{\left|\mathsf{h}_i^{n,\min} - \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{\mathsf{h}}\right| + \delta \mathsf{h}_i^{n,\max}}, 1\right), & \text{if } \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}} < \mathsf{h}_i^{n,\min}, \\ 1, & \mathsf{h}_i^{n,\min} \leq \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}} \leq \mathsf{h}_i^{n,\max}, \\ \min\left(\dfrac{\left|\mathsf{h}_i^{n,\max} - \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{\mathsf{h}}\right| + \delta \mathsf{h}_i^{n,\max}}, 1\right), & \text{if } \mathsf{h}_i^{n,\max} < \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}}. \end{cases} \tag{6.8}$$

This guarantees that $\Psi_1(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ and $\Psi_2(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,\mathsf{h}}]$. This enforces a local minimum principle and a local maximum principle on the water height. As a corollary this also enforces positivity of the water height H_i^{n+1}.

We proceed similarly to limit q_1 since the functionals Ψ_3 and Ψ_4 are linear. Denoting the q_1-component of $\mathbf{P}_{ij}$ by $\mathsf{P}_{ij}^{q_1}$, for $\ell_j^{i,q_1} \in [0, \ell_j^{i,\mathsf{h}}]$, we set

$$\ell_j^{i,q_1} = \begin{cases} \min\left(\dfrac{\left|q_{1,i}^{n,\min} - \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{q_1}\right| + \delta q_{1,i}^{n,\max}}, 1\right), & \text{if } \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1} < q_{1,i}^{n,\min}, \\ 1, & q_{1,i}^{n,\min} \leq \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1} \leq q_{1,i}^{n,\max}, \\ \min\left(\dfrac{\left|q_{1,i}^{n,\max} - \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathbf{P}_{ij}^{q_1}\right| + \delta q_{1,i}^{n,\max}}, 1\right), & \text{if } q_{1,i}^{n,\max} < \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1}. \end{cases} \tag{6.9}$$

This guarantees that $\Psi_3(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ and $\Psi_4(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,q_1}]$. This enforces a local minimum principle and a local maximum principle on q_1. As a corollary this also enforces positivity of $\mathsf{Q}_{1,i}^{n+1}$.

Remark 6.5 (FCT limiting on linear functionals) It is also possible to use the FCT methodology for limiting the linear functionals $\Psi_1, \ldots, \Psi_4$. We refer the reader to [19] where this is shown for the shallow water equations. □

We now move on to the kinetic energy functional $\Psi_5^{i,n}$. We seek an $\ell_j^{i,\mathsf{K}} \in [0, \ell_j^{i,q_1}]$ such that $\Psi_5^{i,n}(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,\mathsf{K}}]$. Let us define the functional: $\Phi(\mathbf{U}) := \mathsf{H}\Psi_5^{i,n}(\mathbf{U}) = \mathsf{H}\mathsf{K}_i^{n,\max} - \frac{1}{2}\|\mathbf{Q}\|_{\ell^2}^2$. Notice that $\Psi_5^{i,n}(\mathbf{U}) \geq 0$ iff $\Phi(\mathbf{U}) \geq 0$

provided $\mathsf{H} > 0$. Hence, assuming that $\Psi_5^{i,n}(\mathbf{W}_i^{\mathrm{L},n+1} + \mathbf{P}_{ij}) < 0$ (otherwise there is nothing to optimize), our optimization problem consists of finding the unique $\ell \in [0, 1)$ such that $\Phi(\mathbf{W}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}) = 0$. But $\Phi(\mathbf{W}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij})$ is a quadratic functional with respect to ℓ: $\Phi(\mathbf{W}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}) = a\ell^2 + b\ell + c$, where

$$a = -\frac{1}{2}\|\mathbf{P}_{ij}^{\boldsymbol{q}}\|_{\ell^2}^2, \tag{6.10a}$$

$$b = \mathsf{K}_i^{n,\max}\mathsf{P}_{ij}^{\mathsf{h}} - \mathbf{Q}(\mathbf{W}_i^{\mathrm{L},n+1}) \cdot \mathbf{P}_{ij}^{\boldsymbol{q}}, \tag{6.10b}$$

$$c = \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\mathsf{K}_i^{n,\max} - \frac{1}{2}\|\mathbf{Q}(\mathbf{W}_i^{\mathrm{L},n+1})\|_{\ell^2}^2. \tag{6.10c}$$

Let t_0 be the smallest positive root of the equation $at^2+bt+c = 0$, with the convention that $t_0 = 1$ if the equation has no positive root. Then, we choose $\ell_j^{i,\mathsf{K}}$ to be such that

$$\ell_j^{i,\mathsf{K}} = \min(t_0, \ell_j^{i,q_1}). \tag{6.11}$$

It is proved in [15] that the definition (6.11) guarantees that $\Psi_5^{i,n}(\mathbf{W}_i^{\mathrm{L},n+1} + \ell\mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,\mathsf{K}}]$. This enforces a local maximum principle on the kinetic energy.

Finally, we set

$$\ell_{ij} = \min(\ell_j^{i,\mathsf{K}}, \ell_i^{j,\mathsf{K}}). \tag{6.12}$$

Then, with the above definition and by Lemma 6.4, the update $\mathbf{W}_i^{n+1}$ computed by (6.6) satisfies the following constraints $\Psi_l^{i,n}(\mathbf{W}_i^{n+1}) \geq 0$ for all $l \in \mathcal{L}$. We now "put back" the sources to compute the final limited update $\mathbf{U}_i^{n+1}$

$$\mathbf{U}_i^{n+1} = \tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n) + \sum_{j\in\mathcal{I}^*(i)} \theta_j\Big(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n\Big). \tag{6.13}$$

Theorem 6.6 *Let $i \in \mathcal{V}$ and $n \geq 0$. Assume that $\boldsymbol{U}_j^n \in \mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid h > 0\}$ for all $j \in \mathcal{I}(i)$. Suppose that the time step τ_n is small enough so that $\tau_n \max\left(\frac{2}{m_i}\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n}, \frac{\sqrt{gH_0}}{\mathcal{E}_i}\right) \leq 1$. Let $\boldsymbol{W}_i^{n+1}$ be defined by (6.6) with the limiter ℓ_{ij} given by (6.12). Then, $\boldsymbol{W}_i^{n+1} \in \mathcal{A}$. Consequently, the full update $\boldsymbol{U}_i^{n+1}$ defined by (6.13) is in $\mathcal{A}$ as well.*

Proof By construction, the definition (6.12) along with Lemma 6.4 gives

$$\mathsf{H}(\mathbf{W}_i^{n+1}) := \mathsf{H}\Big(\sum_{j\in\mathcal{I}^*(i)} \theta_j\Big(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij}\mathbf{P}_{ij}^n\Big)\Big) \geq \mathsf{h}_i^{n,\min},$$

for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$. The goal is to show that the limited water height update H_i^{n+1} stays positive with the contribution of the source $\tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n)$. For $i \in \mathcal{V}$, consider the update for H_i^{n+1}:

$$\begin{aligned}\mathsf{H}_i^{n+1} &= \tau_n \chi\left(\frac{\mathsf{h}_i^{i,\max}-\mathsf{h}_i^{i,\min}}{\mathsf{h}_{\text{wave}}^{\min}}\right)\left(\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\big(\mathsf{h}_{\text{wave}}(\boldsymbol{a}_i, t^n) - \mathsf{H}_i^n\big)G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\text{gen}}})\right) + \mathsf{H}(\mathbf{W}_i^{n+1}),\\ &\geq \tau_n \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\chi\left(\frac{\mathsf{h}_i^{i,\max}-\mathsf{h}_i^{i,\min}}{\mathsf{h}_{\text{wave}}^{\min}}\right)G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\text{gen}}})\left(\mathsf{h}_{\text{wave}}^{\min} - \mathsf{h}_i^{n,\max}\right) + \mathsf{h}_i^{n,\min},\end{aligned}$$

where we used the fact that $\mathsf{h}_i^{n,\max} \geq \mathsf{H}_i^n$. If the cutoff function χ is active (i.e., when $\mathsf{h}_i^{n,\max} - \mathsf{h}_i^{n,\min} \geq \mathsf{h}_{\text{wave}}^{\min}$), then $\chi = 0$ and $\mathsf{H}_i^{n+1} \geq \mathsf{h}_i^{n,\min}$ and thus positive. If the cutoff function χ is not active (i.e., when $\mathsf{h}_i^{n,\max} - \mathsf{h}_i^{n,\min} < \mathsf{h}_{\text{wave}}^{\min}$), then

$$\begin{aligned}\mathsf{H}_i^{n+1} &\geq \tau_n \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\chi\left(\frac{\mathsf{h}_i^{i,\max}-\mathsf{h}_i^{i,\min}}{\mathsf{h}_{\text{wave}}^{\min}}\right)G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\text{gen}}})\left(\mathsf{h}_{\text{wave}}^{\min} - \mathsf{h}_i^{n,\max}\right) + \mathsf{h}_i^{n,\min}\\ &\geq \tau_n \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\chi\left(\frac{\mathsf{h}_i^{i,\max}-\mathsf{h}_i^{i,\min}}{\mathsf{h}_{\text{wave}}^{\min}}\right)G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\text{gen}}})\left(-\mathsf{h}_i^{n,\min}\right) + \mathsf{h}_i^{n,\min}\\ &= \mathsf{h}_i^{n,\min}\left(1 - \tau_n \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\chi\left(\frac{\mathsf{h}_i^{i,\max}-\mathsf{h}_i^{i,\min}}{\mathsf{h}_{\text{wave}}^{\min}}\right)G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\text{gen}}}))\right)\\ &\geq \mathsf{h}_i^{n,\min}\left(1 - \tau_n \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\right).\end{aligned}$$

Thus, H_i^{n+1} is positive under the CFL condition. □

Remark 6.7 (Iterative limiting) We note here that the limiting process described above can be iterated multiple times by observing from (6.4) that

$$\mathbf{W}_i^{\mathrm{H},n+1} = \mathbf{W}_i^{\mathrm{L},n+1} + \frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij}^n + \frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} (1-\ell_{ij})\mathbf{A}_{ij}^n.$$

□

Then, by setting $\mathbf{W}^{(0)} := \mathbf{W}_i^{\mathrm{L},n+1}$ and $\mathbf{A}_{ij}^{(0)} = \mathbf{A}_{ij}^n$, the iterative limiting process is shown in Algorithm 1. In the numerical simulations reported in Sect. 7, we take $k_{\max} = 2$.

Algorithm 1 Iterative limiting with sources

Input: $\mathbf{W}^{\mathrm{L},n+1}$, $\mathbf{A}_{ij}^n$, $k_{\max}$
Output: $\mathbf{U}^{n+1}$
  Set $\mathbf{W}^{(0)} := \mathbf{W}_i^{\mathrm{L},n+1}$ and $\mathbf{A}_{ij}^{(0)} = \mathbf{A}_{ij}^n$
  for $k = 0$ *to* $k_{\max} - 1$ **do**
    Compute limiter $\ell^{(k)}$
    Update $\mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} + \frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} \ell_{ij}^{(k)}\mathbf{A}_{ij}^{(k)}$
    Update $\mathbf{A}_{ij}^{(k+1)} = (1-\ell_{ij}^{(k)})\mathbf{A}_{ij}^{(k)}$
  end
  $\mathbf{U}^{n+1} = \mathbf{W}^{(k_{\max})} + \tau_n(\boldsymbol{R}^n + \boldsymbol{S}^n)$

Proposition 6.8 (Well balancing) *Let* $\boldsymbol{T} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ *be the high-order scheme defined by* (6.13). *This scheme is exactly well balanced if* $\boldsymbol{S}_G \equiv \boldsymbol{0}$.

Proof Assume that $\boldsymbol{u}_h^n$ is exactly at rest, then one can verify that $\mathbf{P}_{ij} = \mathbf{0}$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}^*(i)$. Hence, $\boldsymbol{u}_h^{n+1}$ is equal to the low-order update. We conclude by invoking Proposition 4.4. □

6.4 Relaxation of the Bounds

The methodology described above leads to second-order accuracy in the L^1-norm, but the bounds defined in (4.16) are too tight to make the method higher order or even second-order in the L^∞-norm in the presence of smooth extrema. This is very important for the system (2.1) and (2.2) since we want to model smooth solitary waves and periodic waves. To recover the full accuracy in the L^∞-norm, one should *relax* the bounds (4.16) for smooth solutions. This has been observed in [28] and explained in [15, Sec. 4.7]. We refer the reader to [20, Sec. 7.6] where this is discussed in detail. All the numerical results reported in Sect. 7 are done using the relaxation technique from [15, 20].

7 Numerical Illustrations

In this section, we illustrate the performance of the proposed method. We first verify the convergence of the scheme (6.13) and then verify that it is well balanced. We then reproduce several laboratory experiments that validate the proposed model.

7.1 Implementation Details

The simulations reported in the paper are done in $\mathbb{R}^d$ with $d = \{1, 2\}$ using continuous, linear finite elements. When $d = 1$, we use a uniform grid. Some two-dimensional tests are done with continuous $\mathbb{P}_1$ finite elements on unstructured Delaunay meshes, and some tests are done using continuous $\mathbb{Q}_1$ finite elements on quadrangular meshes. In all the tests, we set $\overline{\lambda} = 1$ and set the approximate relaxation parameter to $\mathcal{E}_i := m_i^{\frac{1}{d}} = (\int_D \varphi_i \, \mathrm{d}x)^{\frac{1}{d}}$ for all $i \in \mathcal{V}$.

For the numerical tests in $\mathbb{R}^2$, three different codes implementing the method described in the paper have been written to ensure reproducibility. The first code, henceforth referred to as `TAMU`, does not use any particular software and is written in Fortran 95/2003. The second code has been written at the US Army Engineer Research and Development Center using the `Proteus` toolkit (the reader is referred to [27]). Both codes use continuous $\mathbb{P}_1$ Lagrange elements on triangles and unstructured, non-nested, Delaunay meshes. The third code is `Ryujin` [22, 33], a high-performance finite-element solver based on the `deal.II` library [1] and uses continuous $\mathbb{Q}_1$ elements. The time stepping in all three codes is done with the third-order, three stage, strong stability preserving Runge–Kutta method, SSP RK(3,3).

Table 1 Convergence table using $\|\mathsf{h} - \mathsf{h}_h\|_{L^1}/\|\mathsf{h}\|_{L^1}$ for solitary wave solution of Serre model (2.1). $T = 50$ s, CFL$= 0.075$

	Galerkin		EV with (2.8)		EV with (5.7)	
100	2.80E−04	Rate	2.48E−04	Rate	2.53E−04	Rate
200	4.24E−05	2.72	5.54E−05	2.16	5.64E−05	2.17
400	3.02E−05	0.49	3.74E−05	0.57	3.74E−05	0.59
800	2.32E−05	0.38	2.48E−05	0.59	2.48E−05	0.59
1600	1.39E−05	0.74	1.43E−05	0.79	1.43E−05	0.79
3200	7.67E−06	0.85	7.89E−06	0.86	7.89E−06	0.86
6400	3.84E−06	1.00	4.08E−06	0.95	4.08E−06	0.95

Remark 7.1 (Choosing the CFL value) When the relaxation parameter is chosen to be proportional to the local mesh size h, the algorithm performs optimally when the CFL value is proportional to $\sqrt{\frac{h}{\mathsf{H}_{0,\max}}}$. For more details on this, we refer the reader to [21, Sec. 5.6]. □

7.2 Convergence Tests

We now verify the convergence rate of the method defined by (6.13). For the sake brevity, we only use the `TAMU` code for these tests.

7.2.1 Solitary Wave over a Flat Bottom

The Serre model (2.1) admits an exact solution in the form of a solitary wave propagating over a flat bottom. We note here that the hyperbolic relaxed model (2.3) does not support exact solitary waves; more on this is discussed in Sect. 7.2.2.

Let $\tilde{\mathsf{h}}(x,t)$ and $\tilde{u}(x,t)$ be the water height and velocity of an exact solitary wave:

$$\tilde{\mathsf{h}}(x,t) = \mathsf{h}_0 + \frac{\alpha}{(\cosh(r(x - x_0 - ct)))^2}, \quad \tilde{u}(x,t) = c\frac{\tilde{\mathsf{h}}(x,t) - \mathsf{h}_0}{\tilde{\mathsf{h}}(x,t)}, \tag{7.1}$$

with wave speed $c = \sqrt{g(\mathsf{h}_0 + \alpha)}$ and width $r = \sqrt{\frac{3\alpha}{4\mathsf{h}_0^2(\mathsf{h}_0+\alpha)}}$. We initialize the water height and discharge by setting

$$\mathsf{h}(x,0) = \max\{\tilde{\mathsf{h}}(x,0) - z(x), 0\}, \quad q(x,0) = \tilde{u}(x,0)\mathsf{h}(x,0). \tag{7.2}$$

We consider a 1D uniform grid on the domain $D = (0, 1000\,\text{m})$. We set $\mathsf{h}_0 = 10\,\text{m}$ and $\alpha = 1\,\text{m}$. The solitary wave is initiated at $x_0 = 200\,\text{m}$. The final time is $T = 50\,\text{s}$. In Table 1, we compare the quantity $\|\mathsf{h} - \mathsf{h}_h\|_{L^1}/\|\mathsf{h}\|_{L^1}$ using (i) the Galerkin method (i.e., no artificial viscosity); (ii) the method (6.13) using the Shallow Water Equations entropy pair (5.7); (iii) the method (6.13) using the hyperbolic Serre entropy pair (2.8). We observe that the method converges to the solution of the Serre model (2.1) with

Table 2 Convergence rates using manufactured solution. $T = 50\,\mathrm{s}$, $\mathrm{CFL} = 0.05$

	E_1		E_2		E_∞	
100	1.76E−04	Rate	4.50E−04	Rate	1.35E−03	Rate
200	4.98E−05	1.82	1.23E−04	1.86	5.30E−04	1.35
400	1.61E−05	1.63	3.92E−05	1.66	1.77E−04	1.58
800	4.30E−06	1.90	1.03E−05	1.93	4.68E−05	1.92
1600	1.11E−06	1.95	2.60E−06	1.99	1.19E−05	1.97
3200	3.10E−07	1.84	6.65E−07	1.97	3.05E−06	1.97

first-order rate with respect to the mesh-size. The first-order rate is a consequence of the relaxation parameter in the relaxed system (2.3) being proportional to the local mesh size.

7.2.2 Method of Manufactured Solutions Using Solitary Wave Profile

We now verify that the numerical method (6.13) is indeed second-order accurate when considering solutions to the hyperbolic relaxed system (2.3). Since finding exact solutions for (2.3) is a highly non-trivial task, we use here the method of manufactured solutions.

Let $\tilde{\mathsf{h}}(x,t)$ and $\tilde{u}(x,t)$ be the same profiles defined by (7.1). To find the respective source term needed for the method of manufactured solutions, we set $\mathsf{h}(x,t) = \tilde{\mathsf{h}}(x,t)$, $q(x,t) = \tilde{\mathsf{h}}(x,t)\tilde{u}(x,t)$, $q_1(x,t) = \tilde{\mathsf{h}}^2(x,t)$, $q_2(x,t) = -\tilde{\mathsf{h}}^2(x,t)\partial_x(\tilde{u}(x,t))$. Substituting these profiles into (2.3) yields a source term in the form of $\boldsymbol{S}_{\mathrm{man}}(x,t) = \big(0, q_{\mathrm{man}}(x,t), 0, q_{2,\mathrm{man}}(x,t), 0\big)^{\mathsf{T}}$ where $q_{\mathrm{man}}(x,t)$ and $q_{2,\mathrm{man}}(x,t)$ are the residual functions for the equations for $\boldsymbol{q}$ and q_2. We refer the reader to [41] where the exact expressions are shown.

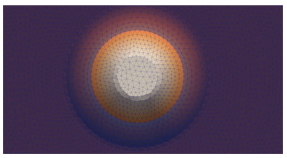
We use the same setup as in Sect. 7.2.1 for running the computations. We show in Table 2 the numerical results obtained at $T = 50\,\mathrm{s}$. The number of grid points is shown in the leftmost column. The relative errors on the water height measured in the L^1-norm, $E_1 := \|\mathsf{h} - \mathsf{h}_h\|_{L^1}/\|\mathsf{h}\|_{L^1}$, L^2-norm, $E_2 := \|\mathsf{h} - \mathsf{h}_h\|_{L^2}/\|\mathsf{h}\|_{L^2}$, and L^∞-norm, $E_\infty := \|\mathsf{h} - \mathsf{h}_h\|_{L^\infty}/\|\mathsf{h}\|_{L^\infty}$, are shown in the second, third and fourth columns. We observe that all the quantities converge with second-order rate with respect to the mesh size, thereby confirming that the proposed approximation technique is second-order accurate in space and time. (It is actually third-order accurate in time.)

7.3 Well-Balancing Tests

In this section, we verify that the scheme is well balanced. To quantify the concept of well balancing, we define the following error indicator:

I	$\delta_\infty(t)$
3587	2.5101E-13
14023	5.6512E-13

(a) $\mathsf{H}_0 = 1\,\text{m}.\ T = 50\,\text{s}.$



(b) Initial data exactly at rest. Top view of $\mathsf{h} + z$.

I	$\delta_\infty(t)$
3587	7.7165E-12
14023	1.0692E-11

(c) $\mathsf{H}_0 = 0.32\,\text{m}.\ T = 50\,\text{s}.$

Fig. 2 Tables and figure for well-balancing tests

$$\delta_\infty(t) := \frac{\|\mathsf{h}_h(t) - \mathsf{h}_0\|_{L^\infty(D)}}{\mathsf{H}_0} + \frac{\|\boldsymbol{q}_h(t) - \boldsymbol{q}_0\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}} + \frac{\|\mathsf{Q}_{1,h}(t) - \mathsf{Q}_{1,0}\|_{L^\infty(D)}}{\mathsf{H}_0^2} + \frac{\|\mathsf{Q}_{2,h}(t) - \mathsf{Q}_{2,0}\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}} + \frac{\|\mathsf{Q}_{3,h}(t) - \mathsf{Q}_{3,0}\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}}, \tag{7.3}$$

where H_0 is some reference water depth, $\mathsf{h}_0, \boldsymbol{q}_0, \mathsf{Q}_{1,0}, \mathsf{Q}_{2,0}, \mathsf{Q}_{3,0}$ are the initial states, and $\mathsf{h}_h(t), \boldsymbol{q}_h(t), \mathsf{Q}_{1,h}(t), \mathsf{Q}_{2,h}(t), \mathsf{Q}_{3,h}(t)$ are the finite-element approximations at time t for the respective conserved variables. We show that this quantity stays close to roundoff error for our numerical tests. All the computations are done with CFL $= 0.5$.

For these tests, we consider the set up of the 1995 experiments by [7] where the bathymetry is defined by a conical island. The experimental domain is given by $D = (0, 25\,\text{m}) \times (0, 30\,\text{m})$. The experimental bathymetry is defined by

$$z(\boldsymbol{x}) = \begin{cases} \min\left(0.625, 0.9 - r(\boldsymbol{x})/4\right), & r(\boldsymbol{x}) < 3.6 \\ 0, & \text{otherwise}, \end{cases}$$

where $r(\boldsymbol{x})$ is the radius from the center of the island located at (12.96 m, 13.80 m).

We first consider the case where the initial profile is a complete wet state. The reference water depth is set to $\mathsf{H}_0 = 1\,\text{m}$ (above the island) and we define the initial water height to be $\mathsf{h}_0(\boldsymbol{x}) = \mathsf{H}_0 - z(\boldsymbol{x})$ and initial flow rate $\boldsymbol{q}_0 = 0\,\text{m/s}$. The simulations are run until $T = 50\,\text{s}$. In Fig. 2a, we report the well-balancing quantity (7.3) for two different meshes and observe that indeed the values are near machine precision.

We now consider the case where the initial state is a wet-dry state. We set the reference depth to be $\mathsf{H}_0 = 0.32\,\text{m}$ so that the water elevation intersects the cone at $r(\boldsymbol{x}) = 2.32\,\text{m}$. To initiate this problem properly, we begin exactly at rest with respect to the mesh. That is to say, the mesh is *aligned* with the initial data in regions where $\mathsf{h} + z$ is constant. We show this refinement in Fig. 2b. In Fig. 2c, we report the well-balancing quantity (7.3) for two different meshes and see that it also stays close to machine precision.

7.4 Dam Break with Friction

We now consider the test case of a dam break over a dry bottom with three conical obstacles introduced by [25] (and reproduced by others: [21, 24], etc.). This

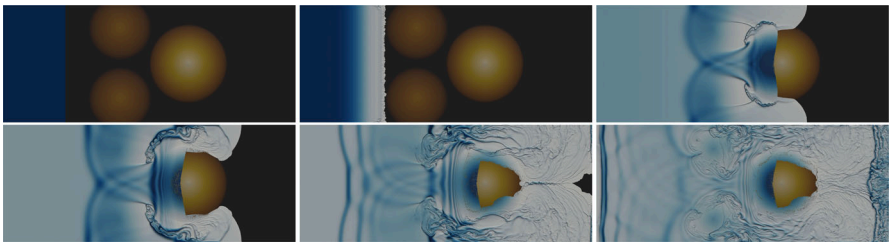


Fig. 3 Dam break with bumps—surface plot of the water elevation $\mathsf{h} + z$ at $t = \{0, 1, 7.8, 10, 15, 20\}$

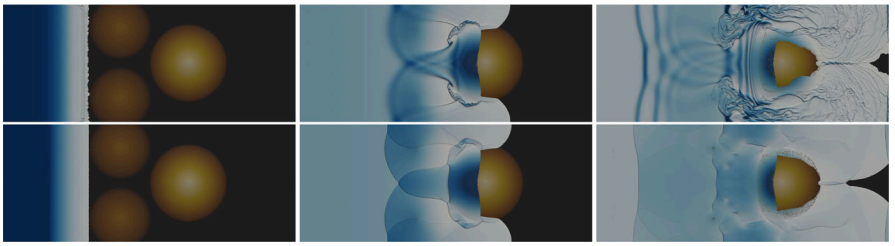


Fig. 4 Dam break with bumps—comparison with Serre's model (top) and SWEs (bottom) at $t = \{1, 7.8, 15\}$

benchmark tests the complex wetting/drying process and the method's ability to handle the Gauckler–Manning friction source. Here, the Mannings coefficient is set to $n = 0.02\,\mathrm{m}^{-1/3}$ s.

The domain is set to $D = [0, 75\,\mathrm{m}] \times [0, 30\,\mathrm{m}]$. The bathymetry consisting of three conical obstacles is defined by $z(\boldsymbol{x}) := \max\{0, z_1(\boldsymbol{x}), z_2(\boldsymbol{x}), z_3(\boldsymbol{x})\}$ where

$$z_1(\boldsymbol{x}) = 1 - \tfrac{1}{8}\sqrt{(x-30)^2 + (y-6)^2}, \tag{7.4a}$$

$$z_2(\boldsymbol{x}) = 1 - \tfrac{1}{8}\sqrt{(x-30)^2 + (y-24)^2}, \tag{7.4b}$$

$$z_3(\boldsymbol{x}) = 3 - \tfrac{3}{10}\sqrt{(x-47.5)^2 + (y-15)^2}. \tag{7.4c}$$

are the three obstacles. The initial state is set to

$$\mathsf{h}_0(\boldsymbol{x}) = \begin{cases} 1.875, & x \le 16 \\ 0, & \text{otherwise,} \end{cases} \qquad \boldsymbol{q}_0(\boldsymbol{x}) = \boldsymbol{0}.$$

For these computations, we use the `Ryujin` code described above. The mesh is composed of rectangular elements with 2, 307, 361 $\mathbb{Q}_1$ DOFS. The final time is set to $T = 20$ s with CFL 0.125. In Fig. 3, we show the computational free surface elevation $\mathsf{h} + z$ at several time snapshots. Then, in Fig. 4, we show the comparison of the computations with the hyperbolic Serre model (2.3) and the Saint-Venant shallow water equations. We observe that more realistic structures are produced by the dispersive Serre model. The computations for this benchmark were performed on the 32-node Whistler cluster at Texas A&M University. More specifically, 288 MPI ranks with

two threads per core were used for each model. The total computation time for the hyperbolic dispersive Serre model was approximately 74 min and the computation time for the Saint-Venant model was approximately 51 min.

7.5 Laboratory Experiments

We continue the numerical illustrations by reproducing several laboratory experiments that have been documented in the literature. We focus on two experiments involving the propagation of periodic waves over varying topographies and one involving the propagation of solitary waves.

7.5.1 Experiment 1: Propagation of Periodic Waves over an Elliptic Shoal

We consider the 1982 experiments of [5] conducted to study the propagation of monochromatic waves over an elliptic shoal. The goal of the experiments were to model the refraction and diffraction of waves when propagating over a varying bottom. These experiment have become a benchmark for validating dispersive wave models (see: [10, 34]).

The experimental basin is composed of a $\frac{1}{50}$ sloping bottom which forms a 20° angle with the y-axis and an elliptic-shaped shoal built on the ramp. We reproduce this bathymetry as follows. We first define the rotated coordinates $\boldsymbol{x} \mapsto \boldsymbol{x}_r(\boldsymbol{x})$:

$$x_r := x\cos(20^\circ) - y\sin(20^\circ), \quad y_r := x\sin(20^\circ) + y\cos(20^\circ).$$

Then, we define the sloping bottom and elliptic shoal profiles as

$$z_{\text{ramp}}(\boldsymbol{x}) := \begin{cases} \frac{1}{50}(x_r(\boldsymbol{x}) + 5.82), & -5.82 \leq x_r(\boldsymbol{x}) \leq 14 \\ 0.3964, & 14 \leq x_r(\boldsymbol{x}) \\ 0, & \text{otherwise}, \end{cases}$$

$$z_{\text{shoal}}(\boldsymbol{x}) := \begin{cases} -0.3 + \frac{1}{2}\sqrt{1 - (\frac{x_r(\boldsymbol{x})}{3.75})^2 - (\frac{y_r(\boldsymbol{x})}{5})^2}, & (\frac{x_r(\boldsymbol{x})}{3.75})^2 + (\frac{y_r(\boldsymbol{x})}{5})^2 \leq 1 \\ 0, & \text{otherwise}. \end{cases}$$

The full bathymetry is defined as $z(\boldsymbol{x}) := z_{\text{ramp}}(\boldsymbol{x}_r(\boldsymbol{x})) + z_{\text{shoal}}(\boldsymbol{x}_r(\boldsymbol{x}))$. Note that this bathymetry is slightly modified from that proposed in [5] to include a flat portion at the right-end of the basin.

The reference water depth is set to $\mathsf{H}_0 = 0.45\,\text{m}$. We initialize the water height with $\mathsf{h}_0(\boldsymbol{x}) = \mathsf{H}_0 - z(\boldsymbol{x})$ and discharge $\boldsymbol{q}_0(\boldsymbol{x}) = \boldsymbol{0}$. For the simulation of the experiments, we use the `Ryujin` code. The computational domain is set to be $D = (-14, 18\ \text{m}) \times (-10, 10\ \text{m})$. We generate the periodic waves via the generation zone methodology described in Sect. 2.3.2 with the profiles:

$$\mathsf{h}(x,t) = \mathsf{h}_0 + a\sin(kx - \sigma t), \quad u(x,t) = \frac{a}{\mathsf{h}_0}\frac{\sigma}{k}\sin(kx - \sigma t).$$

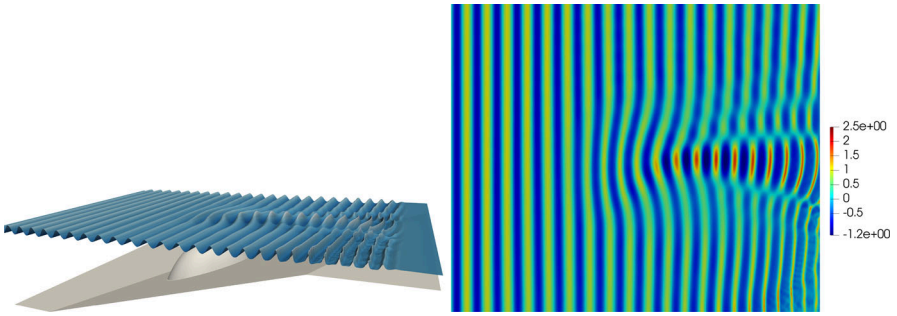


Fig. 5 Experiment 1—Elliptic shoal experiments. Left: Free surface elevation and topography; mesh refinement 1. Right: Normalized view of wave heights; mesh refinement 1

The amplitude is set to $a = 0.0232\,\mathrm{m}$ and the period to $T_p = 1\,\mathrm{s}$. The wave frequency is given by $\sigma = \frac{2\pi}{T_p}$ and k is found using the dispersion relation for the full Serre model: $k^2 = 3\sigma^2/(3g\mathsf{h}_0 - \mathsf{h}_0^2\sigma^2)$. The generation zone length and absorption zone length are set to 4 m, i.e., $L_{\mathrm{gen}} = L_{\mathrm{abs}} := 4\,\mathrm{m}$, and we set $x_{\min} := -14$ and $x_{\max} := 18$. We run the computations until the final time $T = 60\,\mathrm{s}$ to allow the waves to reach a steady state. To verify our results, we run the computations on three different meshes composed of 657,025, 2,624,769 and 10,492,417 $\mathbb{Q}_1$ nodes labeled refinement 1, 2, and 3 respectively. The finest simulation was done using 640 MPI ranks with two threads per core. The wall clock time was 17h. In Fig. 5, we show a snapshot of the free surface elevation and topography at time $T = 60\,\mathrm{s}$ and a normalized view of the generated wave heights (i.e., $\frac{\mathsf{h}+z(\boldsymbol{x})-\mathsf{H}_0}{a}$).

In the experiments, the water elevation is measured at eight sections throughout the basin. These sections are

$$\begin{aligned}
&\text{section } 1 : \{x = 1\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, && \text{section } 2 : \{x = 3\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\},\\
&\text{section } 3 : \{x = 5\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, && \text{section } 4 : \{x = 7\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\},\\
&\text{section } 5 : \{x = 9\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, && \text{section } 6 : \{y = \text{-2}\,\mathrm{m}, 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\},\\
&\text{section } 7 : \{y = 0\,\mathrm{m}, \ \ 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\}, && \text{section } 8 : \{y = 2\,\mathrm{m}, \ \ 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\}.
\end{aligned}$$

To properly compare our numerical results with the experimental data, we do the following: we extract the data over the temporal window $t \in [40\,\mathrm{s}, T]$ of the reference water elevation $\mathsf{h} + z - \mathsf{H}_0$; we then take the maximum of this data over every period in the temporal interval. We then normalize the wave heights with the incoming wave amplitude $a = 0.0232\,\mathrm{m}$. In Fig. 6, we show the comparison with the computational results for the three different meshes. We see that the approximate solutions converge and we observe that the computational results compare reasonably well with the experimental data.

7.5.2 Experiment 2: Propagation of Periodic Waves over Semi-circular Shoal

We now consider the 1971 experiments of [42] performed at the U.S. Army Engineer Waterways Experiment Station (now the U.S. Army Engineer Research and Develop-

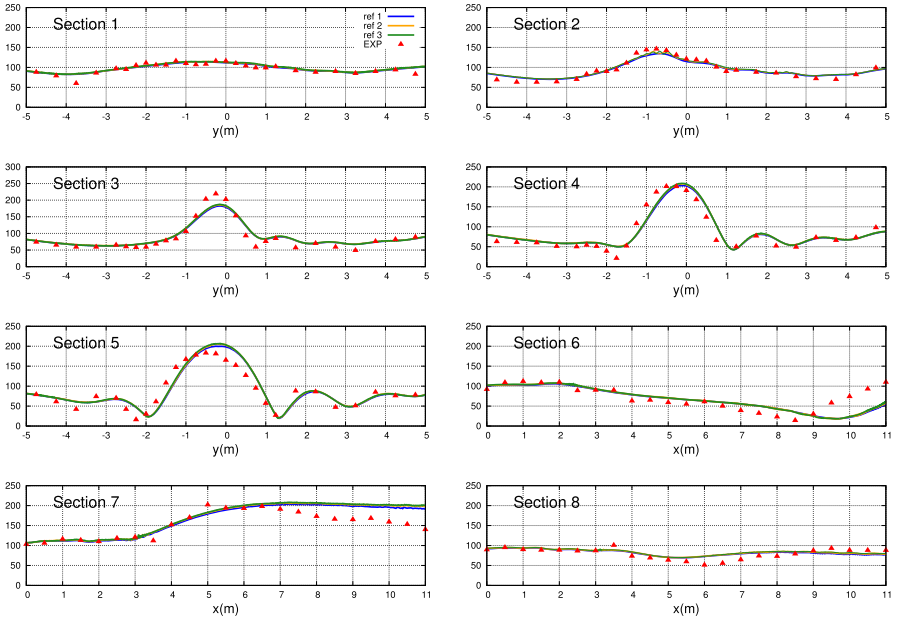


Fig. 6 Experiment 1—Comparison of numerical results (3 mesh refinements) with the experimental data along the eight sections. Experimental data: red triangles

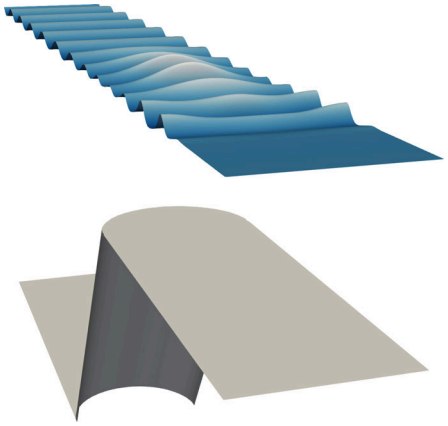
ment Center) in Vicksburg, Mississippi. The goal of the experiments is to study the refraction and diffraction of periodic waves propagating over a semi-circular shoal. In particular, we reproduce the experiments conducted where the wave period is $T = 2\,\mathrm{s}$ and amplitude $a = 0.0075\,\mathrm{m}$ (see [42, Fig. 68]).

The experimental basin is designed to be 25.603 m in length and 6.096 m wide and the still water elevation is set to 0.4572 m. We reproduce these experiments with the `Ryujin` code. We define the computational domain as $(-10, 33\,\mathrm{m}) \times (0, 6.096\,\mathrm{m})$. The lengths of the generation and relaxation zones are defined to be $L_{\mathrm{gen}} = L_{\mathrm{abs}} = 8\,\mathrm{m}$ (which is roughly 2 wave lengths) with $x_{\mathrm{min}} = -10$ and $x_{\mathrm{max}} = 33$. The bathymetry is reproduced as follows: Defining $G(y) := \sqrt{y(6.096 - y)}$, we set

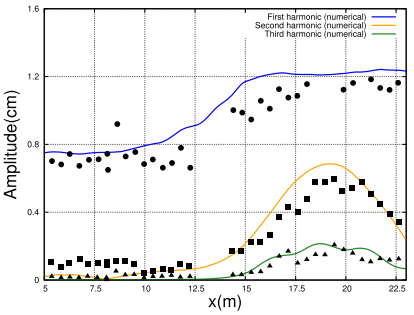
$$z(\boldsymbol{x}) = \begin{cases} 0, & 0 \le x \le 10.67 - G(y), \\ -0.04(10.67 - G(y) - x), & 10.67 - G(y) \le x \le 18.297 - G(y) \\ 0.3048, & 18.297 - G(y) \le x. \end{cases}$$

The computational domain is composed of a quadrilateral mesh with 265,761 $\mathbb{Q}_1$ dofs. We run the numerical simulations until the time $T = 60\,\mathrm{s}$ to allow the waves to reach a steady state. The CFL number is set to 0.125.

In [42], the authors perform the harmonic analysis of the wave elevation data at the centerline of the basin over one period. This is done to study the non-linear transfer of energy from lower to higher frequency components as the waves propagate and focus over the topography. We numerically reproduce this harmonic analysis as follows: We



(a) Free surface elevation and topography at $T = 60\,\mathrm{s}$



(b) Comparison of numerical and experimental amplitude spectrum

Fig. 7 Experiment 2—Whalin semi-circular shoal results

interpolate the centerline $y = 3.048\,\mathrm{m}$ with roughly 1400 points along the x-axis at every 0.001 s in the interval $t \in [58, 60\,\mathrm{s}]$. We then perform the discrete Fourier Transform of the time-series wave elevation data at each point along centerline. In Fig. 7, we show (a) the computational free surface elevation at $T = 60\,\mathrm{s}$; (b) the comparison of the amplitude spectrum with the numerical first, second and third harmonics (sold lines) and the experimental data of Whalin (black geometric shapes). The amplitude spectrum of the waves in the numerical simulations is very close to the experimental one. Note that the experimental data was extracted directly from [42, Fig. 68] using the software `WebPlotDigitzer` [36].

7.5.3 Experiment 3: Propagation over a Solitary Wave over a Triangular Shelf with Conical Island

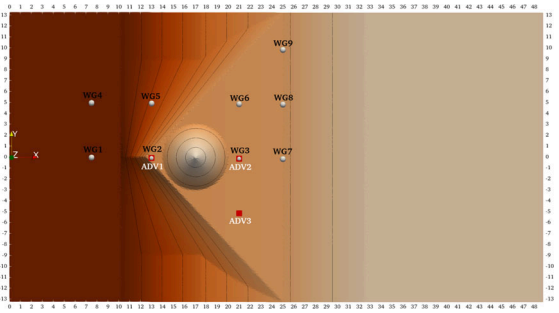
We now reproduce the experiments of [31, 40] performed at the O.H. Hinsdale Wave Research Laboratory of Oregon State University. The experiments were conducted to study specific phenomena that are known to occur when solitary waves propagate over irregular bathymetry such as shoaling, refraction, breaking, etc. Several others (see: [10, 26, 35]) have used these experiments for validation.

We reproduce the bathymetry of the experiments as follows: Let $r = 3$, $h_{\mathrm{cone}} = 0.45$, $d(y) := 1 - \min(1, |y|/13.25)$, $a_x(y) := 12.5 + 12.4999(1 - d(y))$, $a_z(y) := 0.7 + 0.05(1 - d(y))$. We define separately the cone, base and triangular shelf portions of the bathymetry:

$$\mathrm{cone}(x, y) =:= \max\left(h_{\mathrm{cone}} - \sqrt{\frac{(x-17)^2 + y^2}{(\frac{3}{0.45})^2}}, 0\right),$$

Gauge	x(m)	y(m)
WG1	7.5	0.0
WG2	13.0	0.0
WG3	21.0	0.0
WG4	7.5	5.0
WG5	13.0	5.0
WG6	21.0	5.0
WG7	25.0	0.0
WG8	25.0	5.0
WG9	25.0	10.0
ADV1	13.0	0.0
ADV2	21.0	0.0
ADV3	21.0	-5.0

(a) WGs and ADVs coordinates.



(b) Overview of bathymetry with WGs (white spheres) and ADVs locations (red boxes).

Fig. 8 Experiment 3—**a** Coordinates of the wave gauges and ADVs in meters; **b** overview of their respective locations on the bathymetry

$$\text{base}(x,y) := \begin{cases} 0, & x < 10.2, \\ \frac{0.5-0.0}{17.5-10.2}(x-10.2), & 10.2 \le x \le 17.5, \\ 1+\frac{1-0.5}{32.5-17.5}(x-32.5), & 17.5 \le x \le 32.5, \\ 1, & \text{otherwise}, \end{cases}$$

$$\text{shelf}(x,y) := \begin{cases} 0, & x < 10.2, \\ \frac{a_z(y)}{a_x(y)-10.2}(x-10.2), & 10.2 \le x \le a_x(y), \\ 0.75+\frac{a_z(y)-0.75}{a_x(y)-25}(x-25), & a_x(y) \le x \le 25, \\ 1+\frac{1-0.5}{32.5-17.5}(x-32.5), & 25 \le x \le 32.5, \\ 1, & \text{otherwise}. \end{cases}$$

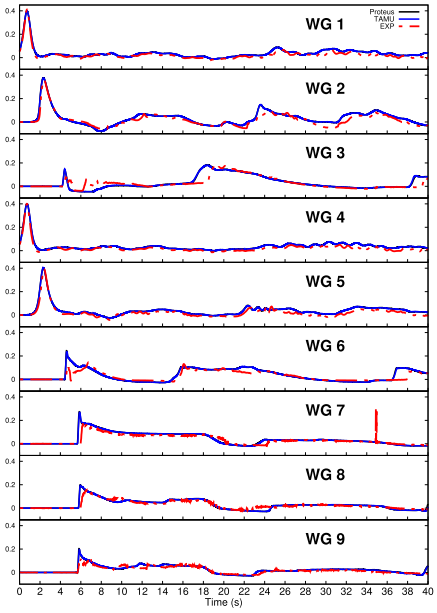
Then, the full bathymetry is defined by

$$z(x,y) := \text{cone}(x,y) + \max(\text{base}(x,y), \text{shelf}(x,y)).$$

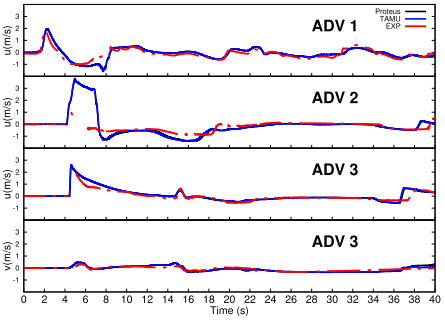
The setup of this complex bathymetry can be seen in Fig. 8a.

The computations are done in the domain $(0, 48.8\,\text{m}) \times (-13.25, 13.25\,\text{m})$. The solitary wave is initiated at $x_0 = 5\,\text{m}$ with reference water depth $\mathsf{h}_0 = 0.78\,\text{m}$ and amplitude $\alpha = 0.39\,\text{m}$ using (7.1). We run the computations until $T = 40\,\text{s}$ with a CFL number of 0.25. We note that for this particular problem, it is our experience that no friction is needed to reproduce correctly the experiment. In Fig. 11, we show the surface plots of the free surface elevation $\mathsf{h} + z$ on a mesh composed of 57,854 $\mathbb{P}_1$ nodes at various times using the `TAMU` code.

In the experiments, nine wave gauges (WGs) are placed along the basin to capture the free surface elevation along with three Acoustic Doppler Velocimeters (ADVs) that



(a) Wave gauges



(b) Acoustic Doppler Velocimeters

Fig. 9 Experiment 3—**a** Temporal series over the period $t \in [0, 40\text{ s}]$ of the free surface elevation $\mathsf{h} + z$ compared to the experimental data (red dashed). The `TAMU` code results are in blue (solid) and `Proteus` code results in black (solid). **b** Temporal series over the period $t \in [0, 40\text{ s}]$ of velocity $\boldsymbol{v}$ (blue solid) and experimental ADVs (red dashed)

measure the velocity. In Fig. 8, we show on the left panel the coordinates of the wave gauges and ADVs, and their respective locations on the bathymetry in the right panel of the figure. We show in Fig. 9a, the comparison between the free surface elevation values of the numerical simulation and the experimental data over the temporal period $t \in [0, 40\,\text{s}]$ using both codes. In Fig. 9b, we show the comparison between the numerical velocities and the experimental data from the ADVs. For both the free surface and velocities, our results compare exceptionally well with the experimental data. We also see that the results of the `TAMU` and `Proteus` codes agree very closely and are almost indistinguishable. The `Proteus` computations were done on a mesh composed of 57,188 $\mathbb{P}_1$ nodes and CFL number of 0.25. Notice that one observes wave breaking in this experiment. This phenomenon is naturally accounted for by the model (2.3) which is hyperbolic and therefore permits shocks and energy dissipation. For completeness, we show the results for WG1 and ADV1 up to the final time $t = 80\,\text{s}$ using the `TAMU` code (Figs. 10 and 11).

8 Conclusion

In this work, we propose a new numerical method for solving the dispersive Serre–Green–Naghdi equations with full topography effects and sources using continuous finite elements. The method is based on a hyperbolic relaxation technique introduced

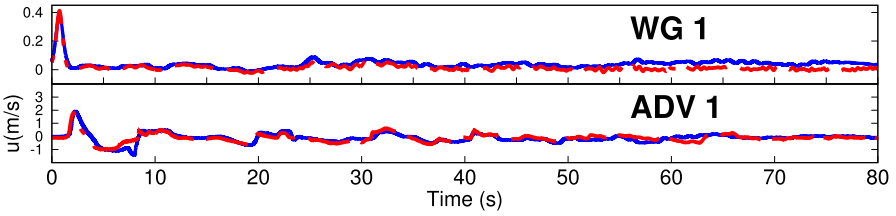


Fig. 10 Experiment 3—Results for WG1 and ADV1 for final time $T = 80\,\mathrm{s}$

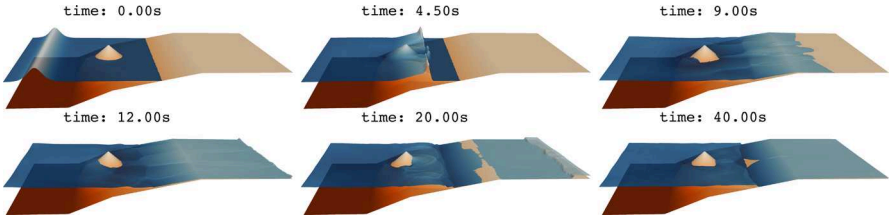


Fig. 11 Experiment 3—Surface plot of the water elevation $\mathsf{h} + z$ at several times

in [23]. The method is well balanced, positivity-preserving and explicit in time. The method extends the work in [21] by introducing physical source terms which are treated explicitly. The novelties of the method are the introduction of high-order artificial viscosity coefficients based on the entropy commutator and a local convex limiting technique that preserves positivity of the water height. The convex limiting follows the general framework from [20] but the treatment of the source terms is radically different. The method is numerically illustrated with various benchmarks seen in the literature and compared to experimental results. The robustness of the method is also illustrated by comparing three separate implementations.

Funding This material is based upon work supported in part by the National Science Foundation grants DMS-1619892 and DMS-1620058, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397, and by the Army Research Office under grant/contract number W911NF-19-1-0431. Permission was granted by the Chief of Engineers to publish this information.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

1. Arndt, D., Bangerth, W., Blais, B., Fehling, M., Gassmöller, R., Heister, T., Heltai, L., Köcher, U., Kronbichler, M., Maier, M., Munch, P., Pelteret, J.-P., Proell, S., Simon, K., Turcksin, B., Wells, D., Zhang, J.: The `deal.II` library, version 9.3. J. Numer. Math. (2021, accepted for publication)

2. Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**(6), 2050–2065 (2004)
3. Azerad, P., Guermond, J.-L., Popov, B.: Well-balanced second-order approximation of the shallow water equation with continuous finite elements. SIAM J. Numer. Anal. **55**(6), 3203–3224 (2017)
4. Bassi, C., Bonaventura, L., Busto, S., Dumbser, M.: A hyperbolic reformulation of the Serre–Green–Naghdi model for general bottom topographies. Comput. Fluids **212**, 104716 (2020)
5. Berkhoff, J., Booy, N., Radder, A.: Verification of numerical wave propagation models for simple harmonic linear water waves. Coast. Eng. **6**(3), 255–279 (1982)
6. Bonneton, P., Chazel, F., Lannes, D., Marche, F., Tissier, M.: A splitting approach for the fully nonlinear and weakly dispersive Green–Naghdi model. J. Comput. Phys. **230**(4), 1479–1498 (2011)
7. Briggs, M.J., Synolakis, C.E., Harkins, G.S., Green, D.R.: Laboratory experiments of tsunami runup on a circular island. Pure Appl. Geophys. **144**(3), 569–593 (1995)
8. Chertock, A., Cui, S., Kurganov, A., Wu, T.: Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. Int. J. Numer. Methods Fluids **78**(6), 355–383 (2015)
9. Duchêne, V.: Rigorous justification of the Favrie–Gavrilyuk approximation to the Serre–Green–Naghdi model. Nonlinearity **32**(10), 3772–3797 (2019)
10. Duran, A., Marche, F.: A discontinuous Galerkin method for a new class of Green–Naghdi equations on simplicial unstructured meshes. Appl. Math. Model. **45**, 840–864 (2017)
11. Escalante, C., Dumbser, M., Castro, M.J.: An efficient hyperbolic relaxation system for dispersive non-hydrostatic water waves and its solution with high order discontinuous Galerkin schemes. J. Comput. Phys. **394**, 385–416 (2019)
12. Favrie, N., Gavrilyuk, S.: A rapid numerical method for solving Serre–Green–Naghdi equations describing long free surface gravity waves. Nonlinearity **30**(7), 2718–2736 (2017)
13. Green, A.E., Naghdi, P.M.: A derivation of equations for wave propagation in water of variable depth. J. Fluid Mech. **78**(2), 237–246 (1976). https://doi.org/10.1017/S0022112076002425
14. Green, A.E., Laws, N., Naghdi, P.M.: On the theory of water waves. Proc. R. Soc. (Lond.) Ser. A **338**, 43–55 (1974)
15. Guermond, J., Nazarov, M., Popov, B., Tomas, I.: Second-order invariant domain preserving approximation of the Euler equations using convex limiting. SIAM J. Sci. Comput. **40**(5), A3211–A3239 (2018)
16. Guermond, J.-L., Pasquetti, R.: A correction technique for the dispersive effects of mass lumping for transport problems. Comput. Methods Appl. Mech. Eng. **253**, 186–198 (2013)
17. Guermond, J.-L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016)
18. Guermond, J.-L., Popov, B., Yang, Y.: The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations. J. Sci. Comput. **70**(3), 1358–1366 (2017)
19. Guermond, J.-L., Quezada de Luna, M., Popov, B., Kees, C., Farthing, M.: Well-balanced second-order finite element approximation of the shallow water equations with friction. SIAM J. Sci. Comput. **40**(6), A3873–A3901 (2018)
20. Guermond, J.-L., Popov, B., Tomas, I.: Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. Comput. Methods Appl. Mech. Eng. **347**, 143–175 (2019). ISSN:0045-7825
21. Guermond, J.-L., Popov, B., Tovar, E., Kees, C.: Robust explicit relaxation technique for solving the Green–Naghdi equations. J. Comput. Phys. **399**, 108917, 17 (2019)
22. Guermond, J.-L., Maier, M., Popov, B., Tomas, I.: Second-order invariant domain preserving approximation of the compressible Navier–Stokes equations. Comput. Methods Appl. Mech. Eng. **375**(1), 113608 (2021)
23. Guermond, J.-L., Kees, C., Popov, B., Tovar, E.: Hyperbolic relaxation technique for solving the dispersive Serre–Green–Naghdi equations with topography. J. Comput. Phys. **450**, 110809 (2022)
24. Huang, Y., Zhang, N., Pei, Y.: Well-balanced finite volume scheme for shallow water flooding and drying over arbitrary topography. Eng. Appl. Comput. Fluid Mech. **7**(1), 40–54 (2013)
25. Kawahara, M., Umetsu, T.: Finite element method for moving boundary problems in river flow. Int. J. Numer. Methods Fluids **6**(6), 365–386 (1986)

26. Kazolea, M., Delis, A., Synolakis, C.: Numerical treatment of wave breaking on unstructured finite volume approximations for extended Boussinesq-type equations. J. Comput. Phys. **271**, 281–305 (2014). Frontiers in Computational Physics
27. Kees, C.E., Farthing, M.W.: Parallel computational methods and simulation for coastal and hydraulic applications using the Proteus toolkit. In: Supercomputing 11: Proceedings of the PyHPC11 Workshop (2011)
28. Khobalatte, B., Perthame, B.: Maximum principle on the entropy and second-order kinetic schemes. Math. Comput. **62**(205), 119–131 (1994)
29. Kurganov, A., Petrova, G.: A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. Commun. Math. Sci. **5**(1), 133–160 (2007)
30. Lannes, D.: Modeling shallow water waves. Nonlinearity **33**(5), R1–R57 (2020)
31. Lynett, P., Swigler, D., El Safty, H., Motoya, L., Keen, A., Son, S., Higuera, P.: Study of the three-dimensional hydrodynamics associated with a solitary wave traveling over an alongshore-variable, shallow shelf. J. Waterw. Port Coast. Ocean Eng. (ASCE) (2019)
32. Madsen, P.A., Bingham, H.B., Schäffer, H.A.: Boussinesq-type formulations for fully nonlinear and extremely dispersive water waves: derivation and analysis. Proc. R. Soc. (Lond.) Ser. A **459**, 1075–1104 (2003)
33. Maier, M., Kronbichler, M.: Efficient parallel 3d computation of the compressible Euler equations with an invariant-domain preserving second-order finite-element scheme. ACM Trans. Parallel Comput. (2021, accepted). arXiv:2007.00094
34. Ricchiuto, M., Filippini, A.G.: Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. J. Comput. Phys. **271**, 306–341 (2014)
35. Roeber, V., Cheung, K.F.: Boussinesq-type model for energetic breaking waves in fringing reef environments. Coast. Eng. **70**, 1–20 (2012). ISSN:0378-3839
36. Rohatgi, A.: Webplotdigitizer: Version 4.5, 2021. https://automeris.io/WebPlotDigitizer
37. Samii, A., Dawson, C.: An explicit hybridized discontinuous Galerkin method for Serre–Green–Naghdi wave model. Comput. Methods Appl. Mech. Eng. **330**, 447–470 (2018)
38. Seabra-Santos, F.J., Renouard, D.P., Temperville, A.M.: Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacle. J. Fluid Mech. **176**, 117–134 (1987)
39. Serre, F.: Contribution à l'étude des écoulements permanents et variables dans les canaux. La Houille Blanche **6**, 830–872 (1953). https://doi.org/10.1051/lhb/1953058
40. Swigler, D.T.: Laboratory study investigating the three-dimensional turbulence and kinematic properties associated with a breaking solitary wave. Master's thesis, Texas A&M University, College Station, Texas (2009)
41. Tovar, E.: Well-balanced and invariant domain schemes for dispersive shallow water flows. PhD thesis, Texas A&M (2021, in preparation)
42. Whalin, R.: The limit of applicability of linear wave refraction theory in a convergence zone. PhD thesis, Texas A&M University (1971)
43. Zhang, Y., Kennedy, A.B., Panda, N., Dawson, C., Westerink, J.J.: Generating-absorbing sponge layers for phase-resolving wave models. Coast. Eng. **84**, 1–9 (2014)