WELL-BALANCED AND INVARIANT DOMAIN PRESERVING SCHEMES FOR DISPERSIVE SHALLOW WATER FLOWS.

A Dissertation

by

ERIC JOSEPH TOVAR

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jean-Luc Guermond
Committee Members,	Andrea Bonito
	Bojan Popov
	Jean Ragusa
Head of Department,	Sarah Witherspoon

December 2021

Major Subject: Mathematics

ABSTRACT

As urbanization encroaches more on flood prone regions and paved surfaces are ever expanding, more catastrophic flash floods occurring in urban environments are expected in the near future. These risks are compounded by global changes in the climate. Mathematics can help better predict and understand these situations through modeling and numerical simulations. The aim of this work is to discuss current mathematical and computational issues in modeling shallow water flows with applications in coastal hydraulics, large-scale oceanography and in-land flooding.

Our mathematical starting points are the systems of partial differential equations known as the (i) Saint-Venant shallow water equations and (ii) dispersive Serre–Green–Naghdi (SGN) equations. The goal of this work is to efficiently solve both mathematical models supplemented with external physical source terms for in-land flooding and large-scale coastal oceanography applications. In particular, the work focuses on introducing a novel technique for solving the Serre–Green–Naghdi equations. We introduce new analytical solutions of the SGN equations with topography that are used to verify the accuracy of numerical methods. Then, we propose a new relaxation technique for solving the SGN equations with topography effects that yields a hyperbolic formulation of the equations. This relaxation technique allows us to circumvent the dispersive time step restriction of the Serre Equations which is a major challenge when solving the equations. This method is then supplemented with a novel continuous finite element approximation that is second-order accurate in space, invariant domain preserving and well-balanced. The method is then verified with academic benchmarks and validated by comparison with laboratory experimental data.

DEDICATION

To my mother and my late grandparents, who raised me and shaped me into the person I am. To my wife, whose patience, advice, and unwavering support allowed me to complete this project. To Aleah, my sister at heart for whom I aim to be an adequate role model. To my uncle Javier for his support throughout the years. And to my Tia Rosie who taught me the importance of education and always pushed me to be the best I could be.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Jean-Luc Guermond, for his investment, support and advice throughout my time at Texas A&M University. I would also like to thank my committee members for their patience and guidance: Dr. Bojan Popov, Dr. Andrea Bonito and Dr. Jean Ragusa; your time and feedback was greatly appreciated. I would also like to thank Dr. Chris Kees, my former supervisor at ERDC, for his support, patience and trust throughout my time at ERDC. I would also like to thank Hwai-Ping (Pearce) Cheng, the branch chief at CHL/ERDC, for believing in me and supporting me through my internship. I would also like to thank Dr. Matthias Maier at TAMU for his help and support in the development of the Ryujin software. Lastly, I am also grateful for my friends and colleagues, Bennett C. and Weston B., for their input, discussions, and company throughout my last five years in this program.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Jean-Luc Guermond (advisor), Professor Bojan Popov, Professor Andrea Bonito of the Department of Mathematics at Texas A&M University and Professor Jean Ragusa of the Department of Nuclear Engineering at Texas A&M University. Part of this work was done in collaboration with Professor Chris Kees at the Department of Civil & Environmental Engineering at Louisiana State University for the completion of an internship program at the U.S. Army Engineer Research and Development Center. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was 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. Part of the graduate study was supported by the Texas Water Resources Institute (TWRI).

NOMENCLATURE

$\boldsymbol{x}$	Cartesian position vector
g	acceleration due to gravity $9.81\,\mathrm{ms}^{-2}$
h_0	reference water depth
$\boldsymbol{u}$	conserved solution variable
h	water depth
$\boldsymbol{q}$	momentum vector
$\boldsymbol{v}$	velocity vector
SV	Saint-Venant
SWE	Shallow Water Equations
SGN	Serre–Green–Naghdi
ODE	ordinary differential equation
PDE	partial differential equation
TAMU	Texas A&M University
ERDC	U.S. Army Engineer Research and Development Center

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Introduction

In this chapter, we provide an introduction to the thesis and discuss the background and motivation for the research. In particular, we introduce the concepts of hyperbolic systems of conservation laws and dispersive partial differential equations which are crucial for this work. We then introduce the notion of shallow water flows by deriving the mathematical models of interest from the free-surface Euler equations for water waves: (i) the Saint-Venant shallow water equations and (ii) the Serre–Green–Naghdi equations.

In collaboration with the Coastal and Hydraulics Laboratory at the U.S. Army Engineer Research and Development Center, the goal of this present research is to provide accurate and rapid modeling of shallow water flows with an emphasis on dispersive water waves for applications in coastal hydrodynamics. In particular, we are interested in the prediction and modeling of in-land and coastal flooding hazards due to events such as dam breaks, tsunami waves and hurricane surges. In this work, dispersive water waves are defined as waves whose phase speeds are dependent on their wavelength. This phenomena is paramount for modeling smooth periodic waves (and superposition of such waves) and their non-linear interactions in the near-shore region. Such interactions lead to physical processes such as wave shoaling, wave run-up and wave breaking. Dispersion is also an important property for the propagation of tsunami waves in the deep ocean and in the near-shore region when topography effects are important. The mathematical equations describing such physical phenomena are strongly non-linear and pose numerous challenges when one tries to solve them numerically.

Our starting point are the systems of partial differential equations known as the (i) Saint-Venant Shallow Water Equations (SWE) and (ii) dispersive Serre–Green–Naghdi (SGN) equations. The goal of this work is to efficiently solve both mathematical models supplemented with external physical source terms for in-land flooding and large-scale coastal oceanography applications. In

particular, the work focuses on introducing a novel technique for solving Serre–Green–Naghdi equations. That is to say, we propose a new relaxation technique for solving the SGN equations with topography effects that yields a hyperbolic formulation of the equations. This relaxation technique allows us to circumvent the dispersive time step restriction of the Serre Equations which is a major challenge when solving the equations. This method is then supplemented with a continuous finite element approximation that is second-order accurate in space, invariant domain preserving and well-balanced. The method is then verified with academic benchmarks and validated by comparison with laboratory experimental data.

1.2 Hyperbolic systems

The mathematical models of interest in this work are those with dominant *hyperbolic* features and are essential for applications in computational fluid dynamics. These mathematical models propagate "wave-like" profiles and exhibit special solutions such as shock and expansion waves. The theory of hyperbolic systems of conservation laws dates back to the 1950s in the seminal work of Lax, Hopf, Glimm and many others. For an overview on conservation laws and hyperbolic systems, we refer the reader to the books of Ern and Guermond [18, Chap. 80] and Godlewski and Raviart [24] and references therein. The following material in this section is an overview of [18, Chap. 80.1.2].

Let D be a polyhedral domain in $\mathbb{R}^d$ (where d is the spatial dimension). Let $\boldsymbol{x} \in \mathbb{R}^d$ denote the usual Cartesian position vector. Let $\mathcal{A}$ be a subset of the space $\mathbb{R}^m$ (where $m \in \mathbb{N} \setminus \{0\}$) henceforth referred to as the *admissible set of states*. Let $\boldsymbol{u}$ be a vector of conserved quantities (we call this the conserved variable) which takes values from the set $\mathcal{A}$. Let $\mathbb{f} \in \mathrm{Lip}(\mathcal{A}; \mathbb{R}^{m\times d})$ be a given flux. For a generic state $\boldsymbol{v} \in \mathcal{A}$, the flux $\mathbb{f}$ is a matrix with entries $\mathbb{f}_{il}(\boldsymbol{v})$ for all $i \in \{1 : m\}$ and $l \in \{1 : d\}$. In this thesis, we are interested in first-order systems of the form:

$$\partial_t \boldsymbol{u} + \nabla\cdot\mathbb{f}(\boldsymbol{u}) = \boldsymbol{0}, \tag{1.2.1}$$

where $\nabla \cdot \mathbb{f}(\boldsymbol{u})$ evaluated at $(\boldsymbol{x}, t)$ is a column vector in $\mathbb{R}^m$ with entries:

$$(\nabla \cdot \mathbb{f}(\boldsymbol{u}))_i := \sum_{l \in \{1:d\}} \partial_{x_l} \mathbb{f}_{il}(\boldsymbol{u}(\boldsymbol{x}, t)), \quad \forall i \in \{1 : m\}.$$

Remark 1.2.1. We say that a system of partial differential equations is a first-order system if it involves only first derivatives in space and time.

Definition 1.2.1 (Hyperbolicity)**.** We say that a system of the form (1.2.1) is hyperbolic if and only if the matrix $\mathbb{A}(\boldsymbol{v}, \boldsymbol{n}) \in \mathbb{R}^{m \times m}$ with entries:

$$(\mathbb{A}(\boldsymbol{v}, \boldsymbol{n}))_{ij} := \sum_{l \in \{1:d\}} n_l \partial_{v_j} \mathbb{f}_{il}(\boldsymbol{v}), \quad \forall i, m \in \{1 : m\}, \tag{1.2.2}$$

is diagonalizable over $\mathbb{R}$ for all $\boldsymbol{v} \in \mathcal{A}$ and all unit vectors $\boldsymbol{n} \in \mathbb{R}^d$.

This definition of hyperbolicity is needed in Chapter 4 when we introduce a novel hyperbolic relaxation technique for solving the Serre–Green–Naghdi equations. In this work, we will consider the *Cauchy problem* for hyperbolic (and similar) systems such as the Saint-Venant equations and Serre–Green–Naghdi equations. The Cauchy problem is defined as follows:

Definition 1.2.2 (Cauchy problem for hyperbolic system)**.** For $(\boldsymbol{x}, t) \in D \times \mathbb{R}_+$, the Cauchy problem for a hyperbolic system of conservation laws is given by:

$$\begin{cases} \partial_t \boldsymbol{u} + \nabla \cdot \mathbb{f}(\boldsymbol{u}) = 0, \\ \boldsymbol{u}(\boldsymbol{x}, 0) = \boldsymbol{u}_0(\boldsymbol{x}) \in \mathcal{A}, \end{cases} \tag{1.2.3}$$

where $\mathbb{R}_+ := [0, \infty)$.

A particular Cauchy problem of interest in the context of hyperbolic systems is called the Riemann problem. The Riemann problem is a Cauchy problem with discontinuous initial data consisting of two constant states. More precisely, it is defined as follows:

Definition 1.2.3 (Riemann Problem)**.** Let $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A} \times \mathcal{A}$ be a pair of constant admissible states. Let $\boldsymbol{n} \in \mathbb{R}^d$ be a unit vector. Then, we define the one-dimensional Riemann problem as follows:

$$\partial_t \boldsymbol{u} + \partial_x(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = 0, \qquad \boldsymbol{u}(\boldsymbol{x}, 0) = \begin{cases} \boldsymbol{u}_L, & \text{if } x \leq 0, \\ \boldsymbol{u}_R, & \text{if } x > 0, \end{cases} \tag{1.2.4}$$

Understanding the solution to (1.2.4) is essential for constructing robust numerical schemes. We refer the reader to Godlewski and Raviart [24, Chap. 1] and Ern and Guermond [18, Chap. 80] and for a more general discussion on the Riemann Problem for hyperbolic systems. We also refer the reader to Toro [61] for a detailed study on the Riemann Problem for the Saint-Venant shallow water equations.

1.3 Dispersive partial differential equations

In this work, we are also interested in dispersive partial differential equations. In particular, we are interested in dispersive PDEs because they admit special smooth solutions such as periodic waves and solitary waves which are of interest in the context of coastal hydrodynamics. The main mathematical model of interest is the Serre–Green–Naghdi equations which is a *dispersive* system. In this work, we follow Whitham [64, Chap. 1] and define a dispersive partial differential equation as follows:

Definition 1.3.1 (Linear dispersive partial differential equation)**.** Let $(\boldsymbol{x}, t) \in D \times \mathbb{R}_+$. Then, a linear partial differential equation of the form $\partial_t \boldsymbol{u}(\boldsymbol{x}, t) = \mathcal{L}\boldsymbol{u}(\boldsymbol{x}, t)$ is dispersive if it admits a solution of the form:

$$\boldsymbol{u}(\boldsymbol{x}, t) = a\cos(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t),$$

where the wave frequency ω is a definite real function of the wave number $k = |\boldsymbol{k}|$ and if $\omega''(k) \neq 0$.

Remark 1.3.1 (Non-linear dispersive systems). We say that a system of non-linear partial differential equations is dispersive if its linearized counterpart satisfies Definition 1.3.1.

1.4 Shallow water flows

In this section, we introduce the notion of shallow water flows. In particular, we derive the two mathematical models of interest from the free-surface Euler equations for water waves: (i) the Saint-Venant shallow water equations and (ii) the Serre–Green–Naghdi equations for dispersive water waves. The derivation of the models is presented in an Eulerian reference frame and follows an asymptotic expansion approach based on two non-dimensional quantities ε and μ (which are to be defined). The derivation presented in this section is not new and can be found throughout the literature, but is given for completeness. In particular, the derivation shown here loosely follows that of Lannes [44]. We refer the reader to [44] for a more thorough overview of the derivation of various shallow water models. We also refer the reader to Alvarez-Samaniego and Lannes [1] where a rigorous justification for such approach is given.

1.4.1 Free-surface Euler equations for water waves

Let $\boldsymbol{x} \in \mathbb{R}^d$ ($d = 1, 2$) be the horizontal Cartesian position vector and let z be the vertical Cartesian coordinate. Let $t \geq 0$ denote the time variable. Let $\nabla_{\boldsymbol{x},z}(\cdot)$ denote the gradient operator on the full Cartesian position vector $(\boldsymbol{x}, z)$ and let ∇ be the $\mathbb{R}^d$-dimensional gradient with respect to the horizontal position vector $\boldsymbol{x}$.

We consider the movement of a fluid with a free-surface under the action of gravity and assume that the fluid is incompressible (i.e., constant material density), inviscid (i.e., having negligible viscosity) and irrotational (i.e., no vorticity). Let $\rho \in \mathbb{R}$ denote the fluid density. Assume that the fluid is bounded above by a free-surface $z = \xi(\boldsymbol{x}, t)$ and bounded below by a bottom surface that is independent of time $z = z_b(\boldsymbol{x})$. We call this bottom surface the topography map (this nomenclature is often interchanged with bathymetry map). Assume that at rest, the free-surface is given by the rest state $z = \xi_0 \in \mathbb{R}$ where ξ_0 is a reference elevation. We define the deviation of the free-surface from its rest state by $\eta(\boldsymbol{x}, t) := \xi(\boldsymbol{x}, t) - \xi_0$. We introduce the water depth quantity $\mathsf{h}(\boldsymbol{x}, t) := \xi(\boldsymbol{x}, t) - z_b(\boldsymbol{x}) > 0$ so that $\xi(\boldsymbol{x}, t) = \mathsf{h}(\boldsymbol{x}, t) + z_b(\boldsymbol{x})$. Equivalently, we can write $\mathsf{h}(\boldsymbol{x}, t) = \eta(\boldsymbol{x}, t) + \xi_0 - z_b(\boldsymbol{x})$. Note that at rest, the reference water depth is given by

$\mathsf{h}_0 := \xi_0 - z_b(\boldsymbol{x})$ (since at rest, $\eta = 0$). The part of the domain that is occupied by the fluid is defined by:

$$D(t) := \{(\boldsymbol{x}, z) \in \mathbb{R}^d \times \mathbb{R} \mid z_b(\boldsymbol{x}) < z < \xi(\boldsymbol{x}, t)\}. \tag{1.4.1}$$

Note that the domain is time-dependent. We illustrate the fluid domain set-up in Figure 1.1. Let

$$\mathbb{R}^{d+1} \ni \boldsymbol{V}(\boldsymbol{x}, z, t) := (u(\boldsymbol{x}, z, t), v(\boldsymbol{x}, z, t), w(\boldsymbol{x}, z, t))^{\mathsf{T}},$$

denote the velocity of a fluid particle located at $(\boldsymbol{x}, z)$ at time t. Let $\boldsymbol{v} := (u, v)$ denote the horizontal component of the fluid velocity and w the vertical component. Then, the free-surface Euler equations for an inviscid, incompressible and irrotational fluid under the action of gravity are given as follows:

$$\partial_t \boldsymbol{V} + \boldsymbol{V} \cdot \nabla_{\boldsymbol{x},z} \boldsymbol{V} = -\frac{1}{\rho} \nabla_{\boldsymbol{x},z} p - g \mathbb{e}_z, \qquad \text{in } D(t), \tag{1.4.2a}$$

$$\nabla_{\boldsymbol{x},z} \cdot \boldsymbol{V} = 0, \qquad \text{in } D(t), \tag{1.4.2b}$$

$$\nabla_{\boldsymbol{x},z} \times \boldsymbol{V} = \boldsymbol{0}, \qquad \text{in } D(t), \tag{1.4.2c}$$

where $p = p(\boldsymbol{x}, z, t)$ denotes the pressure in the fluid domain $D(t)$, $g = 9.81\,\mathrm{ms}^{-2}$ is the gravitational constant and $\mathbb{e}_z$ denotes the unit normal vector in the z-direction. The equation (1.4.2a) describes the evolution of the fluid particle velocity under the action of gravity (and can be thought of as a balance of forces), (1.4.2b) represents the incompressibility condition of the fluid and (1.4.2c) represents the irrotational condition. Since the flow is incompressible (i.e., $\nabla_{\boldsymbol{x},z} \cdot \boldsymbol{V} = 0$), the equation (1.4.2a) can be equivalently re-written as

$$\partial_t \boldsymbol{V} + \nabla_{\boldsymbol{x},z} \cdot (\boldsymbol{V} \otimes \boldsymbol{V}) = -\frac{1}{\rho} \nabla_{\boldsymbol{x},z} p - g \mathbb{e}_z \tag{1.4.3}$$

where $\otimes$ denotes the outer product of two vectors. This can be shown with a simple expansion of the advection term.

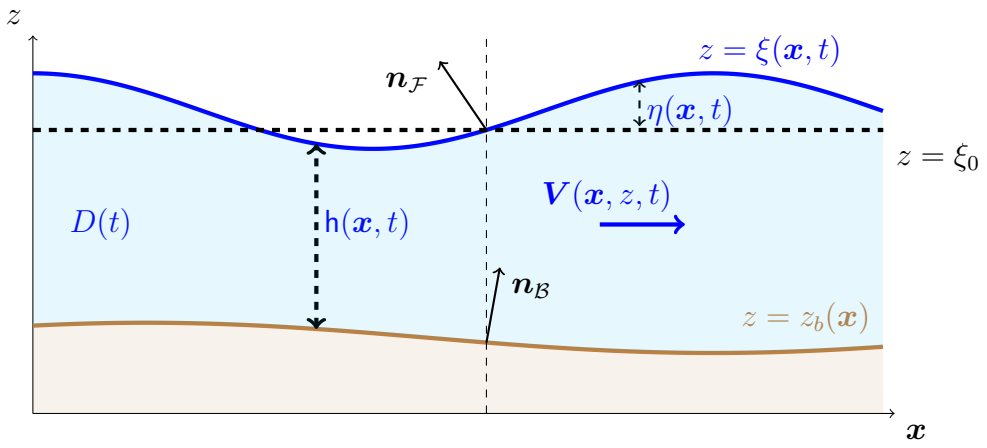


Figure 1.1: Two-dimensional schematic of the free-surface problem.

To complete the system (1.4.2), we need to consider boundary conditions on both the free-surface boundary and bottom boundary. The first boundary condition is the kinematic condition for the free-surface which says that the fluid particles can not cross the free-surface:

$$\boldsymbol{V} \cdot \boldsymbol{n}_{\mathcal{F}} = (\frac{d}{dt} \boldsymbol{x}_{\mathcal{F}}) \cdot \boldsymbol{n}_{\mathcal{F}}, \qquad z = \xi(\boldsymbol{x}, t). \tag{1.4.4}$$

Here, $\boldsymbol{x}_{\mathcal{F}} := (\boldsymbol{x}, z = \xi(\boldsymbol{x}, t))$ is the position coordinate evaluated on the free-surface. The quantity $\boldsymbol{n}_{\mathcal{F}}$ is the outward facing normal of the free-surface and is defined by:

$$\boldsymbol{n}_{\mathcal{F}} := (-\nabla \xi, 1)^{\mathsf{T}}. \tag{1.4.5}$$

Note that the kinematic condition (1.4.4) can be re-written as follows:

$$\partial_t \xi = \boldsymbol{V}_s \cdot \boldsymbol{n}_{\mathcal{F}}, \tag{1.4.6}$$

where $\boldsymbol{V}_s(\boldsymbol{x}, t) := \boldsymbol{V}(\boldsymbol{x}, z = \xi(\boldsymbol{x}, t), t)$ denotes the evaluation of the particle fluid velocity at the free-surface where the subscript "s" is meant to represent free-{s}urface. Note that we can expand

the kinematic condition even further:

$$\partial_t \mathsf{h} + \boldsymbol{v}_s \cdot \nabla(\mathsf{h} + z_b) = w_s, \tag{1.4.7}$$

where we used the definitions $\xi(\boldsymbol{x}, t) := \mathsf{h}(\boldsymbol{x}, t) + z_b(\boldsymbol{x})$ and $\boldsymbol{V} := (\boldsymbol{v}, w)^{\mathsf{T}}$. The second boundary condition we consider is the impermeability of the bottom boundary:

$$\boldsymbol{V} \cdot \boldsymbol{n}_b = 0, \qquad z = z_b(\boldsymbol{x}), \tag{1.4.8}$$

where $\boldsymbol{n}_b$ is the outward facing normal for the bottom boundary defined by:

$$\boldsymbol{n}_b := (-\nabla z_b, 1)^{\mathsf{T}}. \tag{1.4.9}$$

This impermeable condition can be expanded as follows:

$$-\boldsymbol{v}_b \cdot \nabla z_b + w_b = 0, \tag{1.4.10}$$

where the subscript "b" denotes the evaluation at the bottom boundary $z = z_b(\boldsymbol{x})$. The final boundary condition that we consider is the so called dynamic boundary condition (i.e., we neglect surface tension effects):

$$p(\boldsymbol{x}, z = \xi(\boldsymbol{x}, t), t) = p_{\text{atm}} \in \mathbb{R}, \tag{1.4.11}$$

where the constant p_{atm} is the atmospheric pressure. For simplicity, we assume that $p_{\text{atm}} = 0$.

Remark 1.4.1. When discussing the evaluation of quantities at the free-surface or bottom boundaries, we are assuming *a priori* that these operations are justified. That is to say, we do not make any rigorous statements in this derivation. We refer the reader to Alvarez-Samaniego and Lannes [1] where such operations are justified.

The system of partial differential equations (1.4.2) coupled with the boundary conditions (1.4.7)–(1.4.10)–(1.4.11) form a free-surface problem. That is to say, the domain of interest $D(t)$ is itself

an unknown since it is determined by the unknown free-surface quantity $\xi(\boldsymbol{x}, t)$. Thus, in order to solve this free-surface problem, it is necessary to recast it into an equivalent formulation in which the equations are solved in a fixed domain. In this work, we follow the derivation of [44] and recast the free-surface problem using a formulation based on the water depth and flow discharge.

1.4.2 The water depth and flow discharge formulation

In this section, we introduce the water depth and flow discharge formulation of the water waves problem. The goal of this formulation is to cast the free-surface problem in a fixed domain by removing the dependency of the vertical variable z.

We begin the derivation by introducing the horizontal flow discharge (also known as the momentum):

$$\boldsymbol{q}(\boldsymbol{x}, t) := \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}(\boldsymbol{x}, z, t)\, dz. \tag{1.4.12}$$

Recalling that $\xi(\boldsymbol{x}, t) := \mathsf{h}(\boldsymbol{x}, t) + z_b(\boldsymbol{x})$, we also define the average horizontal velocity:

$$\overline{\boldsymbol{v}}(\boldsymbol{x}, t) := \frac{1}{\mathsf{h}(\boldsymbol{x}, t)} \int_{z_b(\boldsymbol{x})}^{\mathsf{h}(\boldsymbol{x},t)+z_b(\boldsymbol{x})} \boldsymbol{v}(\boldsymbol{x}, z, t)\, dz. \tag{1.4.13}$$

We now integrate the incompressibility condition (1.4.2b) over the water column (that is, from $z = z_b(\boldsymbol{x})$ to $z = \xi(\boldsymbol{x}, t)$):

$$\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\nabla_{\boldsymbol{x},z} \cdot \boldsymbol{V})\, dz = 0, \tag{1.4.14}$$

$$\implies \tag{1.4.15}$$

$$\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\nabla \cdot \boldsymbol{v})\, dz + w_s - w_b = 0. \tag{1.4.16}$$

Note that an application of the Leibniz integral rule yields the following relation:

$$\nabla \cdot (\mathsf{h}\overline{\boldsymbol{v}}) = \nabla \xi \cdot \boldsymbol{v}_s - \nabla z_b \cdot \boldsymbol{v}_b + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\nabla \cdot \boldsymbol{v})\, dz. \tag{1.4.17}$$

Then using the kinematic condition (1.4.7), impermeable condition (1.4.10) and (1.4.16), the above

can be re-written as follows:

$$\partial_t \mathsf{h} + \nabla \cdot (\mathsf{h}\overline{\boldsymbol{v}}) = 0.$$

This equation gives an evolution equation for the water depth $\mathsf{h}(\boldsymbol{x}, t)$. This equation is also known as the mass conservation equation and can be equivalently written as:

$$\partial_t \mathsf{h} + \nabla \cdot \boldsymbol{q} = 0. \tag{1.4.18}$$

Notice that this equation does not depend on the vertical coordinate z.

We now want to find an evolution equation for the flow discharge $\boldsymbol{q} := \mathsf{h}\overline{\boldsymbol{v}}$. This can be done by integrating the horizontal part of the velocity evolution equations (1.4.2a) over the vertical water column. Recall that the horizontal components of (1.4.2a) can be expressed as follows:

$$\partial_t \boldsymbol{v} + \nabla \cdot (\boldsymbol{v} \otimes \boldsymbol{v}) + \partial_z (w\boldsymbol{v}) = -\frac{1}{\rho} \nabla p.$$

We will integrate each term individually and apply the Leibniz integral rule to simplify the time derivative term and advection term. We have that for the time derivative term:

$$\begin{aligned} \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_t \boldsymbol{v} \, dz &= \partial_t \left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v} \, dz \right) - \boldsymbol{v}_s \partial_t \xi \\ &= \partial_t \left(\mathsf{h}\overline{\boldsymbol{v}} \right) - \boldsymbol{v}_s \partial_t \mathsf{h}. \end{aligned}$$

The pressure term gives:

$$\begin{aligned} \frac{1}{\rho} \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \nabla p \, dz &= \frac{1}{\rho} \nabla \left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p \, dz \right) - \frac{1}{\rho} \underbrace{p_s \nabla (\mathsf{h} + z_b)}_{=0} + \frac{1}{\rho} p_b \nabla z_b \\ &= \frac{1}{\rho} \nabla \left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p \, dz \right) + \frac{1}{\rho} p_b \nabla z_b. \end{aligned}$$

Note that when the topography map is flat, the last term on the right hand side disappears. We now want to integrate the advection term. For simplicity, we expand the advection term into its

respective components and then integrate. Recalling that $\boldsymbol{v} := (u, v)$, note that:

$$\boldsymbol{v} \otimes \boldsymbol{v} = \begin{pmatrix} u^2 & uv \\ uv & v^2 \end{pmatrix},$$

and

$$\nabla\cdot(\boldsymbol{v} \otimes \boldsymbol{v}) = \Big(\partial_x(u^2) + \partial_y(uv), \partial_x(uv) + \partial_y(v^2)\Big)^{\mathsf{T}}.$$

Then integrating each component gives:

$$\begin{aligned}
\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \big(\partial_x(u^2) + \partial_y(uv)\big)\, dz &= \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_x(u^2)\, dz + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_y(uv)\, dz \\
&= \partial_x\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} u^2\, dz\right) - u_s^2\partial_x\xi + u_b^2\partial_x z_b \\
&\quad +\partial_y\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} uv\, dz\right) - u_s v_s\partial_y\xi + u_b v_b\partial_y z_b,
\end{aligned}$$

and

$$\begin{aligned}
\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \big(\partial_y(v^2) + \partial_x(uv)\big)\, dz &= \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_y(v^2)\, dz + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_x(uv)\, dz \\
&= \partial_y\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} v^2\, dz\right) - v_s^2\partial_y\xi + v_b^2\partial_y z_b \\
&\quad +\partial_x\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} uv\, dz\right) - u_s v_s\partial_x\xi + u_b v_b\partial_x z_b.
\end{aligned}$$

Then, integrating the final term yields:

$$\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \partial_z(w\boldsymbol{v})\, dz = w_s\boldsymbol{v}_s - w_b\boldsymbol{v}_b.$$

Note that all the boundary terms can be combined as follows:

$$-\boldsymbol{v}_s\underbrace{\left(\partial_t\mathsf{h} + \boldsymbol{v}_s\cdot\nabla\xi - w_s\right)}_{=0} + \boldsymbol{v}_b\underbrace{\left(\boldsymbol{v}_b\cdot\nabla z_b - w_b\right)}_{=0} + \frac{1}{\rho}p_b\nabla z_b = \frac{1}{\rho}p_b\nabla z_b, \tag{1.4.19}$$

where the kinematic condition (1.4.7) and (1.4.10) were used to simplify the terms above. Then, the final form of the water depth and flow discharge formulation is given as follows:

$$\partial_t \mathsf{h} + \nabla\cdot\boldsymbol{q} = 0, \tag{1.4.20a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\boldsymbol{v}\otimes\boldsymbol{v})\,dz\Big) + \frac{1}{\rho}\nabla\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p\,dz\right) = -\frac{1}{\rho}p_b\nabla z_b(\boldsymbol{x}). \tag{1.4.20b}$$

This can be equivalently written as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\overline{\boldsymbol{v}}) = 0, \tag{1.4.21a}$$

$$\partial_t (\mathsf{h}\overline{\boldsymbol{v}}) + \nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\boldsymbol{v}\otimes\boldsymbol{v})\,dz\Big) + \frac{1}{\rho}\nabla(\mathsf{h}\overline{p}) = -\frac{1}{\rho}p_b\nabla z_b(\boldsymbol{x}), \tag{1.4.21b}$$

where $\overline{p}$ is the average pressure defined by:

$$\overline{p}(\boldsymbol{x},t) = \frac{1}{\mathsf{h}}\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p\,dz.$$

1.4.3 Decomposition of the pressure

The next step in the derivation of the shallow water models of interest is the decomposition of the pressure term in (1.4.21). We first note that the Euler Equations (1.4.2) admit a particular solution when the flow is at rest: $\xi(\boldsymbol{x},t) = \xi_0$ and $\boldsymbol{V} = \boldsymbol{0}$. At rest, the system (1.4.2) simplifies to the following ordinary differential equation for the pressure:

$$-\frac{1}{\rho}\partial_z p - g = 0, \qquad p_{|\xi=\xi_0} = 0.$$

A simple integration gives the solution $p(\boldsymbol{x},z,t) = -\rho g z + \rho g \xi_0$. This quantity is the so-called hydrostatic pressure. When the fluid is not at rest, the solution to following ordinary differential equation:

$$-\frac{1}{\rho}\partial_z p - g = 0, \qquad p_{|\xi} = 0,$$

is also known as the hydrostatic pressure and is given by: $p_H = -\rho g(z-\xi)$. We will use this special solution to decompose the pressure in the depth/discharge formulation (1.4.21). That is to say, we decompose the pressure into its hydrostatic part (p_H) and non-hydrostatic part (p_{NH}):

$$p(\boldsymbol{x}, z, t) = p_H + p_{NH} = \rho g(\xi - z) + p_{NH}. \tag{1.4.22}$$

We can find an expression for the non-hydrostatic pressure by integrating the vertical component of (1.4.3) and using (1.4.22). Recall that the vertical component of (1.4.3) is given by:

$$\begin{aligned} \partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V}) &= -\frac{1}{\rho}\partial_z p - g \\ &= -\frac{1}{\rho}\partial_z\big(\rho g(\xi - z) + p_{NH}\big) - g \\ &= -\frac{1}{\rho}\partial_z(p_{NH}). \end{aligned}$$

Then integrating with respect to z yields:

$$p_{NH}(\boldsymbol{x}, z, t) = \rho \int_z^{\xi(\boldsymbol{x},t)} \Big(\partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\, dz. \tag{1.4.23}$$

Thus, we can write the full pressure as:

$$p(\boldsymbol{x}, z, t) = \rho g(\xi(\boldsymbol{x}, t) - z) + \rho \int_z^{\xi(\boldsymbol{x},t)} \Big(\partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\, dz. \tag{1.4.24}$$

Consequently, we have the following relations:

$$\begin{aligned} p_b &= \rho g(\xi(\boldsymbol{x}, t) - z_b) + \rho \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \Big(\partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\, dz \\ &= \rho g \mathsf{h} + \rho \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \Big(\partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\, dz \\ &= \rho g \mathsf{h} + (p_{NH})_b, \end{aligned}$$

and

$$\begin{aligned}
\mathsf{h}\overline{p} &= \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p(\boldsymbol{x},z,t)\,dz \\
&= \frac{1}{2}\rho g(\xi - z_b)^2 + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p_{NH}\,dz \\
&= \frac{1}{2}\rho g\mathsf{h}^2 + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p_{NH}\,dz.
\end{aligned}$$

With this decomposition of the pressure, we can now re-write the water depth/flow discharge formulation (1.4.21) as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\overline{\boldsymbol{v}}) = 0, \tag{1.4.25a}$$

$$\partial_t(\mathsf{h}\overline{\boldsymbol{v}}) + \nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\boldsymbol{v}\otimes\boldsymbol{v})\,dz\Big) + \nabla(\frac{1}{2}g\mathsf{h}) + \frac{1}{\rho}\nabla\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} p_{NH}\,dz\right) = -g\mathsf{h}\nabla z_b(\boldsymbol{x}) - \frac{1}{\rho}(p_{NH})_b\nabla z_b(\boldsymbol{x}). \tag{1.4.25b}$$

By the Leibniz integral rule, the momentum equation (1.4.25b) is equivalent to:

$$\partial_t(\mathsf{h}\overline{\boldsymbol{v}}) + \nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\boldsymbol{v}\otimes\boldsymbol{v})\,dz\Big) + \nabla(\frac{1}{2}g\mathsf{h}) + \frac{1}{\rho}\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \nabla p_{NH}\,dz\right) = -g\mathsf{h}\nabla z_b(\boldsymbol{x}).$$

1.4.4 Decomposition of velocity vector

The equation (1.4.25b) is an evolution equation for the quantity $\mathsf{h}\overline{\boldsymbol{v}}$ where $\overline{\boldsymbol{v}}(\boldsymbol{x},t)$ is the average horizontal velocity that does not depend on vertical Cartesian coordinate z. Note that the advection term $\nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)}(\boldsymbol{v}\otimes\boldsymbol{v})\,dz\Big)$ contributes to the system's dependence on the vertical coordinate z. To separate this dependence on the vertical coordinate, we introduce a decomposition of the horizontal velocity vector:

$$\boldsymbol{v}(\boldsymbol{x},z,t) = \overline{\boldsymbol{v}}(\boldsymbol{x},t) + \boldsymbol{v}^*(\boldsymbol{x},z,t), \tag{1.4.26}$$

where $\boldsymbol{v}^*(\boldsymbol{x}, z, t)$ measures the deviation of the average horizontal velocity $\overline{\boldsymbol{v}}(\boldsymbol{x}, t)$ from the horizontal velocity $\boldsymbol{v}$. Recalling that $\overline{\boldsymbol{v}} := \frac{1}{\mathsf{h}} \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}\, dz$, integrating the decomposition (1.4.26) over the water column yields:

$$\begin{aligned}
\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}\, dz &= \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \Big(\overline{\boldsymbol{v}}(\boldsymbol{x}, t) + \boldsymbol{v}^*(\boldsymbol{x}, z, t)\Big)\, dz, \\
&\Longrightarrow \\
\mathsf{h}\overline{\boldsymbol{v}}(\boldsymbol{x}, t) &= \mathsf{h}\overline{\boldsymbol{v}}(\boldsymbol{x}, t) + \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}^*(\boldsymbol{x}, z, t)\, dz, \\
&\Longrightarrow \\
\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}^*(\boldsymbol{x}, z, t)\, dz &= 0.
\end{aligned}$$

That is to say, the mean-value over the water column of the deviations $\boldsymbol{v}^*$ is zero. Using this property and recalling the definition of the momentum quantity, $\boldsymbol{q} := \mathsf{h}\overline{\boldsymbol{v}}$, we can express the advection term as follows:

$$\nabla\cdot\Big(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} (\boldsymbol{v} \otimes \boldsymbol{v})\, dz\Big) = \nabla\cdot(\overline{\boldsymbol{v}} \otimes \boldsymbol{q}) + \nabla\cdot\boldsymbol{R},$$

where

$$\boldsymbol{R} = \int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \boldsymbol{v}^* \otimes \boldsymbol{v}^*\, dz.$$

Since the quantity $\boldsymbol{v}^* = \boldsymbol{v} - \overline{\boldsymbol{v}}$ measures the deviation of the average horizontal velocity from the horizontal velocity, $\boldsymbol{R}$ can be thought of as the contribution of these fluctuations to the momentum equations.

Remark 1.4.2. As stated in [44], the quantity $\boldsymbol{R}$ can be thought of as analog to the Reynolds stress tensor in turbulence and can thus be used to measure the effects of "turbulence" in the model.

Proceeding, we can again re-write the depth/discharge formulation with this velocity decom-

position as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\overline{\boldsymbol{v}}) = 0, \tag{1.4.27a}$$

$$\partial_t(\mathsf{h}\overline{\boldsymbol{v}}) + \nabla\cdot(\overline{\boldsymbol{v}} \otimes \boldsymbol{q}) + \nabla(\frac{1}{2}g\mathsf{h}) + \nabla\cdot\boldsymbol{R} + \frac{1}{\rho}\left(\int_{z_b(\boldsymbol{x})}^{\xi(\boldsymbol{x},t)} \nabla p_{NH}\, dz\right) = -g\mathsf{h}\nabla z_b(\boldsymbol{x}), \tag{1.4.27b}$$

$$p_{NH}(\boldsymbol{x}, z, t) = \rho \int_z^{\xi(\boldsymbol{x},t)} \Big(\partial_t w + \nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\, dz. \tag{1.4.27c}$$

Remark 1.4.3. The non-hydrostatic pressure term contributes towards the model's dispersive effects and will be important for describing realistic water waves.

1.4.5 Non-dimensionalization of the depth and flow discharge formulation

At its current state, the water depth and flow discharge formulation (1.4.27) is too complicated to solve due to the non-hydrostatic pressure term and the contribution of the deviations from the average velocity (i.e., $\nabla\cdot\boldsymbol{R}$). To proceed, we would like to derive *simpler* shallow water models based on asymptotic expansions in terms of the h and $\boldsymbol{q} := \mathsf{h}\overline{\boldsymbol{v}}$ variables.

We now introduce non-dimensionalized quantities to study the asymptotic expansions of the equations. These quantities are based on the typical scales of problems concerning water waves. The quantities are as follows: h_0, the typical water depth; a_{surf}, the order (i.e., size) of variation of the free-surface deviation (i.e., $\eta(\boldsymbol{x}, t)$); a_{bott}, the order of the bottom boundary (i.e., topography map) variation; and typical horizontal length-scale L. We also define the typical reference speed $V_0 := \sqrt{g\mathsf{h}_0}$. The following non-dimensional parameters are paramount for the asymptotic expansion:

$$\begin{aligned}
\mu &= \frac{\mathsf{h}_0^2}{L^2} \quad \rightarrow \quad \textit{shallowness} \text{ parameter},\\
\varepsilon &= \frac{a_{surf}}{\mathsf{h}_0} \quad \rightarrow \quad \textit{amplitude} \text{ parameter},\\
\beta &= \frac{a_{bott}}{\mathsf{h}_0} \quad \rightarrow \quad \textit{topography} \text{ parameter}.
\end{aligned}$$

Remark 1.4.4 (Shallow water flows). The term *shallow water flows* refers to when the shallowness

parameter μ is assumed to be small. In particular, we are interested in scenarios when $\mu \in (0, 1)$.

Remark 1.4.5 (Depth-averaged models). The models of interest in this thesis are the Saint-Venant Shallow Water Equations and the Serre–Green–Naghdi Equations. These models are often referred to as "depth-averaged" models since they evolve (in space and time) the water depth, h, and momentum quantity, $\boldsymbol{q} := \mathsf{h}\overline{\boldsymbol{v}}$, where $\overline{\boldsymbol{v}}$ is the depth-averaged horizontal velocity.

Using the typical problem quantities defined above, we now define the dimensionless variables for the flow/discharge formulation:

$$\boldsymbol{x}^{'} = \frac{\boldsymbol{x}}{L}, \qquad z^{'} = \frac{z}{\mathsf{h}_0}, \qquad t^{'} = \frac{t}{T},$$

$$\eta^{'} = \frac{\eta}{a_{surf}}, \qquad \boldsymbol{q}^{'} = \frac{\boldsymbol{q}}{a_{surf} V_0}, \qquad z_b^{'} = \frac{z_b}{a_{bott}},$$

where $T = L/V_0$. We also set $\mathsf{h}^{'} = \mathsf{h}/\mathsf{h}_0$ and $\overline{\boldsymbol{v}}' = \overline{\boldsymbol{v}}/(\frac{a_{surf}}{\mathsf{h}_0} V_0) = \overline{\boldsymbol{v}}/(\varepsilon V_0)$. The vertical scale quantity W_0 is defined to be $W_0 = \varepsilon \frac{\mathsf{h}_0}{L} V_0$ and is found by comparing the scales in the incompressibility condition (1.4.2b). Note that since $\mathsf{h} := \eta + \xi_0 - z_b$, we have that $\mathsf{h}^{'} = \varepsilon \eta^{'} + \frac{\xi_0}{\mathsf{h}_0} - \beta z_b^{'}$. For simplicity, we assume that reference elevation at rest ξ_0 is of the scale h_0. That is to say, we replace ξ_0 by h_0 so that $\mathsf{h}^{'} = \varepsilon \eta^{'} + 1 - \beta z_b^{'}$.

We now want to substitute the non-dimensionalized quantities above into the depth/discharge formulation (1.4.27). For completeness, we illustrate the non-dimensionalization for a few terms

in the formulation (1.4.27):

$$
\begin{aligned}
\partial_t \mathsf{h} &\rightarrow \frac{\mathsf{h}_0}{T}\partial_{t'}\mathsf{h}' = \mathsf{h}_0\frac{V_0}{L}\partial_{t'}\mathsf{h}', \\
\nabla\cdot\boldsymbol{q} &\rightarrow \frac{a_{surf}V_0}{L}\nabla_{\boldsymbol{x}'}\cdot\boldsymbol{q}' = \varepsilon\mathsf{h}_0\frac{V_0}{L}\nabla_{\boldsymbol{x}'}\cdot\boldsymbol{q}', \\
\partial_t\boldsymbol{q} &\rightarrow \frac{1}{T}a_{surf}V_0\partial_{t'}\boldsymbol{q}' = \mathsf{h}_0\varepsilon\frac{V_0^2}{L}\partial_{t'}\boldsymbol{q}', \\
\nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) &\rightarrow \frac{1}{L}(\varepsilon V_0)a_{surf}V_0\nabla_{\boldsymbol{x}'}\cdot(\overline{\boldsymbol{v}}'\otimes\boldsymbol{q}') = \mathsf{h}_0\varepsilon^2\frac{V_0^2}{L}\nabla_{\boldsymbol{x}'}\cdot(\overline{\boldsymbol{v}}'\otimes\boldsymbol{q}'), \\
\nabla(\frac{1}{2}g\mathsf{h}^2) &\rightarrow \frac{1}{L}\frac{V_0^2}{\mathsf{h}_0}\mathsf{h}_0^2\nabla_{\boldsymbol{x}'}(\frac{1}{2}(\mathsf{h}')^2) = \mathsf{h}_0\frac{V_0^2}{L}\nabla_{\boldsymbol{x}'}(\frac{1}{2}(\mathsf{h}')^2), \\
\partial_t w &\rightarrow \frac{1}{T}\varepsilon\frac{\mathsf{h}_0}{L}V_0\partial_{t'}w' = \varepsilon\frac{\mathsf{h}_0}{L}\frac{V_0^2}{L}\partial_{t'}w'.
\end{aligned}
$$

Note that we used the fact that $g = V_0^2/\mathsf{h}_0$ in the second-to-last expansion. Now, substituting the non-dimensionalized quantities above into the full depth/discharge formulation (1.4.27) yields (dropping the prime symbol $'$ notation for simplicity):

$$\partial_t\mathsf{h} + \varepsilon\nabla\cdot\boldsymbol{q} = 0, \tag{1.4.28a}$$

$$\varepsilon\partial_t(\mathsf{h}\overline{\boldsymbol{v}}) + \varepsilon^2\nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) + \nabla(\frac{1}{2}\mathsf{h}^2) + \varepsilon^2\nabla\cdot\boldsymbol{R} + \mu\varepsilon\int_{\beta z_b}^{\varepsilon\eta+1}\nabla p_{NH}\,dz = -\mathsf{h}\nabla(\beta z_b), \tag{1.4.28b}$$

$$p_{NH} = \int_z^{\varepsilon\eta+1}\Big(\partial_t w + \varepsilon\nabla_{\boldsymbol{x},z}\cdot(w\boldsymbol{V})\Big)\,dz, \tag{1.4.28c}$$

where $\boldsymbol{R} = \int_{\beta z_b}^{\varepsilon\eta+1}(\boldsymbol{v}^*\otimes\boldsymbol{v}^*)\,dz$. All the computations up to this point have been exact. That is to say, no approximations of the equations have been made. The goal moving forward is to introduce approximations in terms of the ε and μ quantities to derive the shallow water models of interest. In particular, we derive the Saint-Venant Shallow Water Equations which we will show

is the first approximation to the water waves problem and then we derive the more complicated Serre–Green–Naghdi Equations.

1.4.6 Inner structure of the velocity and pressure fields

We begin the approximation by first studying the structure of the velocity field inside the fluid domain. In particular, we will study the incompressibility (1.4.2b) and irrotational conditions (1.4.2c) (along with the impermeable bottom condition (1.4.10)) to find an expression for $\boldsymbol{v}^*$ in terms of $\overline{\boldsymbol{v}}$. These conditions (in dimensionless form) can be re-written as follows:

$$\nabla\cdot\boldsymbol{v} + \partial_z w = 0, \tag{1.4.29a}$$

$$\partial_z\boldsymbol{v} - \mu\nabla w = 0, \tag{1.4.29b}$$

$$\nabla^{\mathsf{T}}\cdot\boldsymbol{v} = 0, \tag{1.4.29c}$$

$$w_b - \nabla(\beta z_b)\cdot\boldsymbol{v}_b = 0, \tag{1.4.29d}$$

where $\nabla^{\mathsf{T}}\cdot\boldsymbol{v} = \partial_y v - \partial_x u$.

Remark 1.4.6 (Representation of the irrotational condition). Note that the irrotational condition is represented by (1.4.29b) and (1.4.29c). Respectively, the equation (1.4.29b) is the horizontal component of the irrotational condition and equation (1.4.29c) is the vertical component of the irrotational condition. For completeness, we show how the horizontal component (1.4.29b) is found from the irrotational condition. Expanding each component of the (1.4.2c) yields

$$\mu\partial_y w - \partial_z v = 0, \quad \partial_z u - \mu\partial_x w = 0, \quad \partial_x v - \partial_y u = 0.$$

Notice that the first and second equations can be "switched" and written as follows:

$$\partial_z u - \mu\partial_x w = 0, \quad \partial_z v - \mu\partial_y w = 0.$$

Careful observation shows that this is equivalent to $\partial_z\boldsymbol{v} - \mu\nabla w = 0$.

We continue the analysis on the inner structure of the velocity field by vertically integrating (1.4.29a) (from βz_b to z) and using the condition (1.4.29d) to see that:

$$\begin{aligned}
w - w_b &= -\int_{\beta z_b}^{z} \nabla\cdot(\overline{\boldsymbol{v}} + \boldsymbol{v}^*)\, dz \\
&= -\int_{\beta z_b}^{z} \nabla\cdot\overline{\boldsymbol{v}}\, dz - \int_{\beta z_b}^{z} \nabla\cdot\boldsymbol{v}^*\, dz, \\
&\Longrightarrow \\
w &= \beta\nabla z_b\cdot\boldsymbol{v}_b - (z-\beta z_b)\nabla\cdot\overline{\boldsymbol{v}} - \int_{\beta z_b}^{z} \nabla\cdot\boldsymbol{v}^*\, dz \\
&= \beta\nabla z_b\cdot(\overline{\boldsymbol{v}} + \boldsymbol{v}_b^*) - (z-\beta z_b)\nabla\cdot\overline{\boldsymbol{v}} - \int_{\beta z_b}^{z} \nabla\cdot\boldsymbol{v}^*\, dz \\
&= \nabla\cdot(\beta z_b\overline{\boldsymbol{v}}) - z\nabla\cdot\overline{\boldsymbol{v}} + \underbrace{\beta\nabla z_b\boldsymbol{v}_b^* - \int_{\beta z_b}^{z} \nabla\cdot\boldsymbol{v}^*\, dz} \\
&= -\nabla\cdot\Big((z-\beta z_b)\overline{\boldsymbol{v}}\Big) - \nabla\cdot\left(\int_{\beta z_b}^{z} \boldsymbol{v}^*\, dz\right).
\end{aligned}$$

Then, notice that the equation (1.4.29b) gives the following:

$$\begin{aligned}
\partial_z\boldsymbol{v} - \mu\nabla w &= 0, \\
&\Longrightarrow \\
\partial_z(\overline{\boldsymbol{v}}(\boldsymbol{x},t) + \boldsymbol{v}^*(\boldsymbol{x},z,t)) - \mu\nabla w &= 0, \\
&\Longrightarrow \\
\partial_z\boldsymbol{v}^* &= \mu\nabla w, \\
&\Longrightarrow \\
constant - \boldsymbol{v}^* &= \int_{z}^{\varepsilon\eta+1} \mu\nabla w\, dz.
\end{aligned}$$

Averaging the last equation over the water column gives $constant = \mu\int_{\beta z_b}^{\varepsilon\eta+1}\int_{z}^{\varepsilon\eta+1} \nabla w\, d\tilde{z}\, dz$ where we used the fact that the mean-value over the water column of $\boldsymbol{v}^*$ is zero. Thus, we can

equivalently write the previous equation as

$$\boldsymbol{v}^* = -\mu\left(\int_z^{\varepsilon\eta+1} w\,dz\right) + \mu\int_{\beta z_b}^{\varepsilon\eta+1}\int_z^{\varepsilon\eta+1} \nabla w\,d\tilde{z}\,dz = -\mu\left(\int_z^{\varepsilon\eta+1} w\,dz\right)^*.$$

This equation can be expanded even further by using the expression for w:

$$\begin{aligned}\boldsymbol{v}^* &= -\mu\left(\int_z^{\varepsilon\eta+1}\nabla w\,dz\right)^* \\ &= \mu\left(\int_z^{\varepsilon\eta+1}\nabla\{\nabla\cdot\Big((z-\beta z_b)\overline{\boldsymbol{v}}\Big)\}\,dz\right)^* + \mu\left(\int_z^{\varepsilon\eta+1}\nabla\{\nabla\cdot\left(\int_{\beta z_b}^{z}\boldsymbol{v}^*\,dz\right)\,dz\}\right)^*\end{aligned}$$

We now follow [44] and introduce the following two operators:

$$\mathbb{T}\boldsymbol{w} = \int_z^{\varepsilon\eta+1}\nabla\{\nabla\cdot\left(\int_{\beta z_b}^{z}\boldsymbol{w}\,dz\right)\,dz\}, \tag{1.4.30a}$$

$$\mathbb{T}^*\boldsymbol{w} = (\mathbb{T}\boldsymbol{w})^*, \tag{1.4.30b}$$

where $\boldsymbol{w}$ is a (sufficiently smooth) $\mathbb{R}^d$-valued function defined on the fluid domain. Note that since $\int_{\beta z_b}^{z}\overline{\boldsymbol{v}}\,dz = (z-\beta z_b)\overline{\boldsymbol{v}}$, we can use the above defined operators to rewrite the equation for $\boldsymbol{v}^*$ as follows:

$$(1-\mu\mathbb{T}^*)\boldsymbol{v}^* = \mu\mathbb{T}^*\overline{\boldsymbol{v}}.$$

Applying the operator $(1+\mu\mathbb{T}^*)$ to both sides of the above equation, gives

$$\boldsymbol{v}^* = \mu\mathbb{T}^*\overline{\boldsymbol{v}} + \mathcal{O}(\mu^2), \tag{1.4.31}$$

which gives that the total horizontal fluid velocity is given by:

$$\boldsymbol{v} = \overline{\boldsymbol{v}} + \mu\mathbb{T}^*\overline{\boldsymbol{v}} + \mathcal{O}(\mu^2). \tag{1.4.32}$$

We have now found an approximate expression for the horizontal velocity in the fluid domain in terms of only the average horizontal velocity $\overline{\boldsymbol{v}}$ up to the order of $\mathcal{O}(\mu^2)$. Since $\overline{\boldsymbol{v}}$ does not depend

on z, we can compute $\boldsymbol{v}$ explicitly. Note that

$$\begin{aligned}\mathbb{T}^*\overline{\boldsymbol{v}} &= \left(\int_z^{\varepsilon\eta+1} \nabla\{\nabla\cdot\Big((z-\beta z_b)\overline{\boldsymbol{v}}\Big)\}\,dz\right)^* \\ &= \int_z^{\varepsilon\eta+1} \nabla\{\nabla\cdot\Big((z-\beta z_b)\overline{\boldsymbol{v}}\Big)\}\,dz - \frac{1}{\mathsf{h}}\int_{\beta z_b}^{\beta z_b+\mathsf{h}}\int_z^{\varepsilon\eta+1} \nabla\{\nabla\cdot\Big((\tilde{z}-\beta z_b)\overline{\boldsymbol{v}}\Big)\}\,d\tilde{z}\,dz.\end{aligned}$$

For simplification purposes, we expand each term separately:

$$\begin{aligned}\nabla\cdot\Big((z-\beta z_b(\boldsymbol{x}))\overline{\boldsymbol{v}}\Big) &= (z-\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} - \nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}, \\ &\Longrightarrow \\ \nabla\{\nabla\cdot\Big((z-\beta z_b(\boldsymbol{x}))\overline{\boldsymbol{v}}\Big)\} &= (z-\beta z_b(\boldsymbol{x}))\nabla(\nabla\cdot\overline{\boldsymbol{v}}) - \nabla(\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} \\ &\qquad -\nabla\{\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}\}, \\ &\Longrightarrow \\ \int_z^{\varepsilon\eta+1}\Big(\nabla\{\nabla\cdot\Big((z-\beta z_b(\boldsymbol{x}))\overline{\boldsymbol{v}}\Big)\}\Big)\,dz &= \frac{1}{2}(\mathsf{h}^2-(z-z_b(\boldsymbol{x}))^2)\nabla(\nabla\cdot\overline{\boldsymbol{v}}) \\ &\qquad -(\mathsf{h}+z_b(\boldsymbol{x})-z)\nabla(\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} \\ &\qquad -(\mathsf{h}+z_b(\boldsymbol{x})-z)\nabla\{\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}\}.\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\frac{1}{\mathsf{h}}\int_{z_b(\boldsymbol{x})}^{\varepsilon\eta+1}\int_z^{\varepsilon\eta+1}\Big(\nabla\{\nabla\cdot\Big((z-\beta z_b(\boldsymbol{x}))\overline{\boldsymbol{v}}\Big)\}\Big)\,dz &= \frac{1}{3}\mathsf{h}^2\nabla(\nabla\cdot\overline{\boldsymbol{v}}) - \frac{1}{2}\mathsf{h}\nabla(\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} \\ &\qquad -\frac{1}{2}\mathsf{h}\nabla\{\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}\}.\end{aligned}$$

Then, continuing the expansion for $\mathbb{T}^*\overline{\boldsymbol{v}}$:

$$\begin{aligned}\mathbb{T}^*\overline{\boldsymbol{v}} &= -\frac{1}{2}\left((z-\beta z_b(\boldsymbol{x}))^2 - \frac{1}{3}\mathsf{h}^2\right)\nabla(\nabla\cdot\overline{\boldsymbol{v}}) \\ &\qquad -\big(\beta z_b(\boldsymbol{x}) - z + \frac{1}{2}\mathsf{h}\big)\Big(\nabla(\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} + \nabla\{\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}\}\Big)\end{aligned}$$

Substituting the above expression into the definition for the full horizontal fluid velocity yields:

$$\boldsymbol{v} = \overline{\boldsymbol{v}} - \mu\frac{1}{2}\left((z-\beta z_b(\boldsymbol{x}))^2 - \frac{1}{3}\mathsf{h}^2\right)\nabla(\nabla\cdot\overline{\boldsymbol{v}}) \\ - \mu\big(\beta z_b(\boldsymbol{x}) - z + \frac{1}{2}\mathsf{h}\big)\Big(\nabla(\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}} + \nabla\{\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}\}\Big) + \mathcal{O}(\mu^2). \quad (1.4.33)$$

Consequently, we have that the definition of the vertical component of the velocity is given by:

$$w = -\nabla\cdot\Big((z-\beta z_b)\overline{\boldsymbol{v}}\Big) - \mu\nabla\cdot\int_{\beta z_b(\boldsymbol{x})}^{\mathsf{h}+z_b(\boldsymbol{x})}(\mathbb{T}^*\overline{\boldsymbol{v}})\,dz + \mathcal{O}(\mu^2). \quad (1.4.34)$$

Note that the first term in the above equation can be expanded as $-(z-\beta z_b(\boldsymbol{x}))\nabla\cdot\overline{\boldsymbol{v}}+\nabla(\beta z_b(\boldsymbol{x}))\cdot\overline{\boldsymbol{v}}$.

We can now use the approximate expressions for $\boldsymbol{v}$ (1.4.33) and w (1.4.34) to study the structure of the pressure field in the fluid domain. Let us recall the full Euler pressure (1.4.24) in dimensionless variables:

$$p = (\varepsilon\eta + 1 - \beta z_b(\boldsymbol{x})) + \mu\varepsilon\int_{\beta z_b}^{\varepsilon\eta+1}\nabla p_{NH}\,dz,$$
$$p_{NH} = \int_z^{\varepsilon\eta+1}\Big(\partial_t w + \varepsilon\nabla\cdot(w\boldsymbol{v}) + \varepsilon\partial_z(w^2)\Big)\,dz.$$

Since the hydrostatic part of the pressure is simpler, we only look at the non-hydrostatic part of the pressure. The goal is to substitute $\boldsymbol{v}$ (1.4.33) and w (1.4.34) into the above non-hydrostatic pressure.

Remark 1.4.7 (Flat topography – A simpler case). Note that when the topography is flat (i.e., $z_b(\boldsymbol{x}) \equiv 0$), a direct computation (with the help of the `Mathematica` software [65]) yields that the non-hydrostatic pressure is given by:

$$p_{NH} = \frac{1}{2}(z^2 - \mathsf{h}^2)\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) + \mathcal{O}(\mu). \quad (1.4.36)$$

This gives

$$\int_{\beta z_b}^{\varepsilon\eta+1} \nabla p_{NH}\, dz = -\nabla\Big\{\frac{1}{3}\mathsf{h}^3\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big)\Big\} + \mathcal{O}(\mu).$$

We now consider the case when the topography is not flat. Another direct computation (using the `Mathematica` software [65]) gives that:

$$\begin{aligned} p_{NH} = \frac{1}{2}((z-\beta z_b(\boldsymbol{x}))^2 - \mathsf{h}^2)\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) \\ + (\mathsf{h}+\beta z_b(\boldsymbol{x}) - z)\Big(\partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}})\Big) + \mathcal{O}(\mu), \quad (1.4.37) \end{aligned}$$

consequently:

$$\begin{aligned} \int_{\beta z_b}^{\varepsilon\eta+1} \nabla p_{NH}\, dz = \nabla\Big\{ -\frac{1}{3}\mathsf{h}^3\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) \\ + \frac{1}{2}\mathsf{h}^2\Big(\partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}})\Big)\Big\} + \mathsf{h}\Bigg(-\frac{1}{2}\mathsf{h}\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) \\ + \Big(\partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}})\Big)\Bigg)\nabla(\beta z_b(\boldsymbol{x})) + \mathcal{O}(\mu). \quad (1.4.38) \end{aligned}$$

Remark 1.4.8 (The structure of the non-hydrostatic pressure contribution). Although the structure of the non-hydrostatic pressure contribution (1.4.38) looks complicated, it can be re-written with a "simpler" mathematical structure. We show how to do this in Section 2.3.3.

Up to this point, we have been able to approximately describe the water waves problem in terms of the water depth h, averaged horizontal velocity $\overline{\boldsymbol{v}}$ and topography map $z_b(\boldsymbol{x})$. That is to say, we have explicitly removed the dependencies of the vertical quantities z (the vertical Cartesian coordinate) and w (the vertical fluid velocity). Although these quantities do not show up in the following models directly, there is still an indirect influence of the vertical fluid acceleration (in particular in the Boussinesq and Serre–Green–Naghdi models). We will now introduce further approximations in terms of the amplitude parameter ε, shallowness parameter μ and topography

parameter β to derive our models of interest.

1.4.7 First order approximation – Saint-Venant Shallow Water Equations

The first mathematical model of interest is the Saint-Venant Shallow Water Equations. This model is an approximation of order $\mathcal{O}(\mu)$ of the non-dimensional water waves problem described above. That is to say, all terms proportional to μ are dropped from the equations. Note that we make no assumption on the amplitude parameter ε. In particular, we have that:

$$\begin{aligned}
&\boldsymbol{v} = \overline{\boldsymbol{v}} + \mathcal{O}(\mu),\\
&\nabla\cdot\boldsymbol{R} = \mathcal{O}(\mu^2),\\
&\mu\int_{\beta z_b(\boldsymbol{x})}^{\varepsilon\eta+1}\nabla p_{NH} = \mathcal{O}(\mu).
\end{aligned}$$

Thus for $\boldsymbol{x}\in\mathbb{R}^d$ and $t\geq 0$, the Saint-Venant model in non-dimensional variables with topography effects is given as follows:

$$\partial_t \mathsf{h} + \varepsilon\nabla\cdot\boldsymbol{q} = 0, \tag{1.4.40a}$$

$$\partial_t \boldsymbol{q} + \varepsilon\nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) + \frac{1}{\varepsilon}\nabla(\frac{1}{2}\mathsf{h}^2) = -\frac{1}{\varepsilon}\mathsf{h}\nabla(\beta z_b(\boldsymbol{x})). \tag{1.4.40b}$$

where we used $\boldsymbol{q} := \mathsf{h}\overline{\boldsymbol{v}}$. Notice that by dropping all the terms proportional to μ, we have neglected all effects induced by the vertical fluid acceleration. In particular, the pressure in the Saint-Venant model (1.4.40) is only the hydrostatic pressure. Consequently, the Saint-Venant model can not model dispersive effects induced by the non-hydrostatic pressure. Since $\nabla\cdot\boldsymbol{R} = \mathcal{O}(\mu^2)$, the Saint-Venant model does not model "turbulent" effects.

Remark 1.4.9 (Dimensional form of the Saint-Venant model). Substituting the dimensional variables into (1.4.40) yields:

$$\partial_t \mathsf{h} + \nabla\cdot(\boldsymbol{q}) = 0, \tag{1.4.41a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) + \nabla(\frac{1}{2}g\mathsf{h}^2) = -g\mathsf{h}\nabla(z_b(\boldsymbol{x})). \tag{1.4.41b}$$

1.4.8 Weakly non-linear second order approximation – The Boussinesq model

Although the work in this thesis is not concerned with the Boussinesq model, we will briefly discuss it here for completeness. The Boussinesq model is a weakly non-linear second order approximation of the full water waves problem. That is to say, it is an order $\mathcal{O}(\mu^2)$ approximation with the additional assumption on the amplitude parameter $\varepsilon = \mathcal{O}(\mu)$ (i.e., size of the elevation variation). Another common assumption that is made is on the topography variations: $\beta = \mathcal{O}(\mu)$. Thus, we can drop any terms in the full water waves problem proportional to μ^2, $\varepsilon\mu$ and $\beta\mu$. Thus, we have the following approximate quantities:

$$
\begin{aligned}
&\boldsymbol{v} = \overline{\boldsymbol{v}} + \mu\mathbb{T}^*\overline{\boldsymbol{v}} + \mathcal{O}(\mu^2),\\
&\nabla\cdot\boldsymbol{R} = \mathcal{O}(\mu^2),\\
&\mu\int_{\beta z_b(\boldsymbol{x})}^{\varepsilon\eta+1}\nabla p_{NH} = \mu\nabla\{-\frac{1}{3}\mathsf{h}^3\partial_t(\nabla\cdot\overline{\boldsymbol{v}})\} + \mathcal{O}(\mu^2).
\end{aligned}
$$

Then, for $\boldsymbol{x} \in \mathbb{R}^d$ and $t \geq 0$, the Boussinesq model in non-dimensional variables with weak non-linear and weak topography effects is given as follows:

$$\partial_t\mathsf{h} + \varepsilon\nabla\cdot(\boldsymbol{q}) = 0, \tag{1.4.43a}$$

$$\partial_t\boldsymbol{q} + \varepsilon\nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) + \frac{1}{\varepsilon}\nabla(\frac{1}{2}\mathsf{h}^2) - \mu\nabla\Big(-\frac{1}{3}\mathsf{h}^3\partial_t(\nabla\cdot\overline{\boldsymbol{v}})\Big) = -\frac{1}{\varepsilon}\mathsf{h}\nabla(\beta z_b(\boldsymbol{x})). \tag{1.4.43b}$$

Remark 1.4.10 (Weakly dispersive). Mathematical models of order $\mathcal{O}(\mu^2)$ are often labeled as weakly dispersive in the literature.

Remark 1.4.11 (Dimensional form of the Boussinesq model). Substituting the dimensional variables into (1.4.43) yields:

$$\partial_t\mathsf{h} + \nabla\cdot(\boldsymbol{q}) = 0, \tag{1.4.44a}$$

$$\partial_t\boldsymbol{q} + \nabla\cdot(\overline{\boldsymbol{v}}\otimes\boldsymbol{q}) + \nabla(\frac{1}{2}g\mathsf{h}^2) + \nabla(-\frac{1}{3}\mathsf{h}^3\partial_t(\nabla\cdot\overline{\boldsymbol{v}})) = -g\mathsf{h}\nabla(z_b(\boldsymbol{x})). \tag{1.4.44b}$$

1.4.9 Strongly non-linear second order approximation – Serre–Green–Naghdi model

We now discuss how to derive the Serre–Green–Naghdi model which is the focus of this thesis. The SGN model is a strongly non-linear second order approximation of the water waves problem (1.4.28). In particular, it is an order of $\mathcal{O}(\mu^2)$ approximation of the free-surface Euler Equations (similar to the Boussinesq model), but without assumptions on the amplitude parameter ε and topography variation parameter β. Due to the lack of additional assumptions on these two parameters, the complexity of the model grows significantly (in particular in the expression for the non-hydrostatic pressure). We have the following approximate quantities for the SGN model:

$$\boldsymbol{v} = \overline{\boldsymbol{v}} + \mu \mathbb{T}^* \overline{\boldsymbol{v}} + \mathcal{O}(\mu^2),$$

$$\nabla \cdot \boldsymbol{R} = \mathcal{O}(\mu^2),$$

$$\begin{aligned}\mu \int_{\beta z_b}^{\varepsilon\eta+1} \nabla p_{NH}\, dz = \mu\nabla\Big\{ -\frac{1}{3}\mathsf{h}^3\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) \\ + \frac{1}{2}\mathsf{h}^2\Big(\partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}})\Big)\Big\} + \mu\mathsf{h}\Bigg(-\frac{1}{2}\mathsf{h}\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big) \\ + \Big(\partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}})\Big)\Bigg)\nabla(\beta z_b(\boldsymbol{x})) + \mathcal{O}(\mu^2).\end{aligned}$$

Notice that we have $\nabla\cdot\boldsymbol{R} = \mathcal{O}(\mu^2)$ which is also the case for the Saint-Venant and Boussinesq models. That is to say, none of these models exhibit "turbulent" effects. Before writing the full model in non-dimensional form, let us define the following two quantities:

$$\mathcal{P}_\varepsilon = -\mathsf{h}\Big(\partial_t(\nabla\cdot\overline{\boldsymbol{v}}) - \varepsilon(\nabla\cdot\overline{\boldsymbol{v}})^2 + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla\cdot\overline{\boldsymbol{v}})\Big), \tag{1.4.46a}$$

$$\mathcal{L}_\varepsilon = \partial_t(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}) + \varepsilon\overline{\boldsymbol{v}}\cdot\nabla(\nabla(\beta z_b)\cdot\overline{\boldsymbol{v}}). \tag{1.4.46b}$$

The expression $\mathcal{P}_\varepsilon + \mathcal{L}_\varepsilon$ can be thought of as the influence of vertical fluid acceleration at the free-surface and $\mathcal{L}_\varepsilon$ can be thought of as the vertical fluid acceleration at the bottom boundary (i.e.,

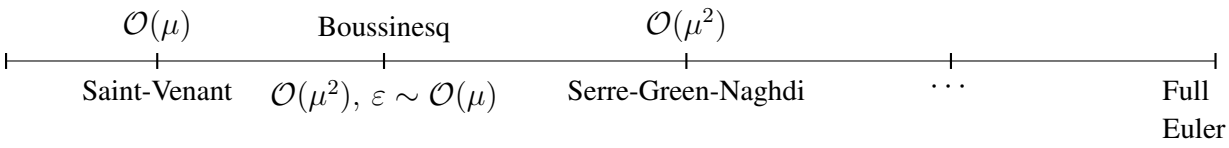


Figure 1.2: A representation of the physical accuracy of each model derived compared to the full free-surface Euler problem.

topography). Note that when the topography is flat, we have that $\mathcal{L}_\varepsilon \equiv 0$. For $\boldsymbol{x} \in \mathbb{R}^d$ and $t \geq 0$, the full non-dimensional Serre–Green–Naghdi model can be written as follows:

$$\partial_t \mathsf{h} + \varepsilon \nabla \cdot (\boldsymbol{q}) = 0, \tag{1.4.47a}$$

$$\partial_t \boldsymbol{q} + \varepsilon \nabla \cdot (\overline{\boldsymbol{v}} \otimes \boldsymbol{q}) + \frac{1}{\varepsilon} \nabla (\frac{1}{2} \mathsf{h}^2) - \mu \nabla \Big(\mathsf{h}^2 (\frac{1}{3} \mathcal{P}_\varepsilon + \frac{1}{2} \mathcal{L}_\varepsilon) \Big) = -\frac{1}{\varepsilon} \Big(\mathsf{h} + \varepsilon \mu \frac{1}{2} \mathcal{P}_\varepsilon + \varepsilon \mu \mathcal{L}_\varepsilon \Big) \nabla (\beta z_b(\boldsymbol{x})). \tag{1.4.47b}$$

In Figure 1.2, we give a representation of the physical accuracy of the models derived in this section. For an overview on open problems and discussion of boundary conditions regarding the models derived in this section, we refer the reader to [44].

2. SHALLOW WATER MODELS

2.1 Introduction

In this Chapter, we present the shallow water models of interest used for applications in inland flooding and coastal hydrodynamics: (i) the Saint-Venant shallow water equations and (ii) the Serre–Green–Naghdi model with topography effects for dispersive water waves. The goal of this chapter is to review the both mathematical models and discuss their mathematical and physical structures. In particular, we investigate steady-state solutions of the equations for developing robust numerical methods and verifying the accuracy of said methods. Although we review both mathematical models, the focus of this thesis is on dispersive water waves so we give particular attention to the Serre–Green–Naghdi equations. A fundamental question of this work is concerned with developing efficient methods for solving the SGN Equations and their use in applications in coastal hydrodynamics. Thus, it is important to understand the properties of the Serre–Green–Naghdi Equations and any potential drawbacks they exhibit. The discussion here lays the foundation for a robust and efficient hyperbolic relaxation technique shown in the following chapters. We then describe the mathematical formulation of the external physical sources that are considered in this work.

The chapter is organized as follows. In Section 2.2, we review the Saint-Venant model and give a brief discussion on its history and status in the literature. In Section 2.2.2, we describe some important properties of the Saint-Venant model and give a discussion on its drawbacks in Section 2.2.3. In Section 2.3, we give a general introduction and historical background to the Serre–Green–Naghdi equations. Then, in Section 2.3.1, we present the model problem for the Serre–Green–Naghdi equations along with a discussion of two important properties. Then in Section 2.3.3, we give two alternative representations of the equations. In Section 2.3.4, we discuss the challenges that are introduced when trying to solve the Serre Equations and its use for coastal hydrodynamics. The fundamental question of this thesis is presented in this section. In Section 2.3.5,

we derive important properties for the Serre model. In particular, a novel result concerning analytical solutions to the Serre–Green–Naghdi model is shown in Proposition 2.3.7. Finally, in Section 2.4 we discuss the external physical sources considered in this work.

2.2 The Saint-Venant model

The Saint-Venant equations are a hyperbolic system of conservation laws that were first derived in Saint-Venant [54] for one-dimensional flows in a channel. As shown in Section 1.4, the equations are an $\mathcal{O}(\mu)$ approximation of the Free-Surface Euler Equations and neglect the effects of vertical acceleration of the fluid. The Saint-Venant equations have a wide range of applications from in-land flooding to atmospheric flows. As stated in Vreugdenhil [62], some of the early applications of digital computations were simulations done using the Saint-Venant equations for atmospheric flows in the 1940's and for oceanographic flows in the 1950's. Due to the general simplicity of the equations' structure along with its mathematical properties (such as propagating shock waves), the Saint-Venant model has been studied and approximated extensively throughout the literature. For an overview of the Saint-Venant equations, we refer the reader to the books of Vreugdenhil [62] and Toro [61] and references therein.

2.2.1 The model problem

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). Let $z(\boldsymbol{x})$ be the given topography (or bathymetry) map. The Saint-Venant model with topography effects can be written as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.2.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q} + p(\boldsymbol{u})\mathbb{I}_d) = -g\mathsf{h}\nabla z \tag{2.2.1b}$$

where the hydrostatic pressure is defined by:

$$p(\boldsymbol{u}) = \tfrac{1}{2} g\mathsf{h}^2. \tag{2.2.2}$$

Here, $\boldsymbol{v}$ is the (depth-averaged) velocity vector and is defined such that $\boldsymbol{q} := \mathsf{h}\boldsymbol{v}$.

2.2.2 Saint-Venant properties

In this section, we briefly discuss a few important properties of the Saint-Venant model that will be used later in this work.

Proposition 2.2.1 (Saint-Venant Energy)**.** *Let $\boldsymbol{u}$ be a smooth solution to the Saint-Venant model* (2.2.1)*. Then, the Saint-Venant model admits a conservation equation for the total energy: $\partial_t E(\boldsymbol{u}) + \nabla\cdot\boldsymbol{F}(\boldsymbol{u}) = 0$ with the energy functional and energy flux defined bu:*

$$E(\boldsymbol{u}) := \frac{1}{2}g\mathsf{h}^2 + gzh + \frac{1}{2}\mathsf{h}\boldsymbol{v}^2, \tag{2.2.3a}$$

$$\boldsymbol{F}(\boldsymbol{u}) := \boldsymbol{v}(E(\boldsymbol{u}) + p(\boldsymbol{u})). \tag{2.2.3b}$$

We define the admissible set for the Saint-Venant model be:

$$\mathcal{A} = \{\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T} \mid \mathsf{h} > 0\}. \tag{2.2.4}$$

This (convex) set can be thought of as the set which contains physically relevant solutions to the Saint-Venant model.

Proposition 2.2.2 (Lake-at-rest for the Saint-Venant Equations)**.** *If the flow is at rest ($\boldsymbol{q} \equiv \boldsymbol{0}$ for $\boldsymbol{x} \in D$), the lake-at-rest steady-state problem for the Saint-Venant Equations* (2.2.1) *is given by:*

$$g\mathsf{h}\nabla(\mathsf{h} + z) = \boldsymbol{0}, \qquad \boldsymbol{x} \in D. \tag{2.2.5a}$$

Proof. Assume the flow is at rest, $\boldsymbol{q} := \mathsf{h}\boldsymbol{v} \equiv \boldsymbol{0}$ (and $\boldsymbol{v} \equiv \boldsymbol{0}$). Since we are interested in steady-state solutions (i.e., independent of time), the Saint-Venant system (2.2.1) reduces to the following

partial differential equation:

$$\nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0,$$
$$\nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla(\frac{1}{2}g\mathsf{h}^2) = -g\mathsf{h}\nabla z.$$

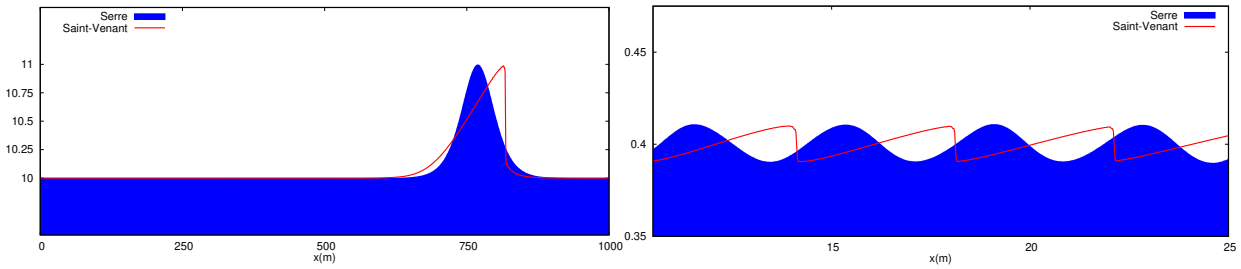
Using $\mathsf{h}\boldsymbol{v} \equiv \mathbf{0}$, the second equation above reduces to $\nabla(\frac{1}{2}g\mathsf{h}^2) + g\mathsf{h}\nabla z = \mathbf{0}$. Expanding the first gradient term yields the lake-at-rest steady-state equation:

$$g\mathsf{h}\nabla(\mathsf{h}+z) = \mathbf{0}.$$

□

2.2.3 Physical drawbacks of the Saint-Venant model

As discussed above, the Saint-Venant model has many applications. However, as shown in Chapter 1, the Saint-Venant model is only an $\mathcal{O}(\mu)$ approximation of the free-surface Euler Equations for water waves and neglects the vertical acceleration effects. Consequently, the Saint-Venant model lacks dispersive properties. More precisely, its linear phase speed is given by $c_p := \frac{\sigma}{k} = \sqrt{g\mathsf{h}_0}$. Here, σ is the wave frequency of a plane-wave solution and k is the wave number (or wave frequency in the spatial domain). Thus, the Saint-Venant model is a less accurate representation of the full water waves problem and can not properly propagate smooth solutions such as solitary waves or periodic waves. We illustrate this visually in Figure 2.1. It is clear from the figure that the Saint-Venant model produces a solution with shocks (i.e., breaking waves) instead of propagating the smooth waves. The lack of dispersive effects of the Saint-Venant model is a major drawback especially when studying wave dynamics in the near-shore region where phenomena such as wave shoaling and wave diffraction are important. The solution for correcting this drawback is by keeping the $\mathcal{O}(\mu^2)$ terms (as shown in Chapter 1) such as is done with the Serre–Green–Naghdi equations.



(a) Solitary wave comparison

(b) Periodic wave comparison

Figure 2.1: Comparison of smooth solutions with Hyperbolic Serre model and Saint-Venant model.

2.3 The Serre–Green–Naghdi Equations

The Serre–Green–Naghdi Equations are a non-linear system of partial differential equations used for modeling dispersive water waves and have a wide-range of applications in coastal hydrodynamics. The Serre Equations were first derived in one spatial-dimension in the work of Serre [57, Eq. (22), p. 860] as a generalization to the Saint-Venant model. The equations were then rediscovered in Su and Gardner [58, Eq. (A14)]. In both of these references, smooth exact solutions in the form of solitary waves were derived for the model (a key property for simulating tsunami waves). The equations were then derived in two spatial dimensions for a flat bottom in Green et al. [26, Eq. (4.12)–(4.15)] and then in Green and Naghdi [25, Eq. (4.27)–(4.30)] for variable topography. The equations were also derived with topography effects in Seabra-Santos et al. [56, Eq. (14)]. The authors of [56] supplemented their mathematical findings with physical experiments involving the propagation of solitary waves over variable topography. In the literature, the nomenclature of the Serre–Green–Naghdi Equations is often interchanged with the Serre equations, Green–Naghdi Equations, or the fully non-linear Boussinesq equations. The latter is a consequence of the Serre–Green–Naghdi model being a weakly dispersive model like the Boussinesq Equations (i.e., $\mathcal{O}(\mu^2)$ approximation to the free-surface Euler equations) but fully non-linear (as shown in §1.4). For the sake of brevity, we interchange the Serre–Green–Naghdi nomenclature with just the Serre model.

2.3.1 The model problem

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 of the system, where h is the water height and $\boldsymbol{q}$ the momentum vector (also known as the flow discharge). Then, the Serre model as first introduced in Serre [57] and extended in Green and Naghdi [25] and Seabra-Santos et al. [56] can be written as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.3.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q} + p(\boldsymbol{u})\mathbb{I}_d) = -r(\boldsymbol{u})\nabla z, \tag{2.3.1b}$$

with the pressure $p(\boldsymbol{u})$ and the source $r(\boldsymbol{u})$ defined by

$$p(\boldsymbol{u}) := \tfrac{1}{2}g\mathsf{h}^2 + \mathsf{h}^2\Big(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}}\Big), \tag{2.3.2a}$$

$$r(\boldsymbol{u}) = g\mathsf{h} + \mathsf{h}\Big(\frac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}}\Big). \tag{2.3.2b}$$

Here, the "dot" objects are defined by:

$$\dot{\mathsf{h}} := \partial_t \mathsf{h} + \boldsymbol{v}\cdot\nabla\mathsf{h}, \tag{2.3.3a}$$

$$\ddot{\mathsf{h}} := \partial_t \dot{\mathsf{h}} + \boldsymbol{v}\cdot\nabla\dot{\mathsf{h}}, \tag{2.3.3b}$$

$$\dot{\mathsf{k}} := \partial_t(\boldsymbol{v}\cdot\nabla z) + \boldsymbol{v}\cdot\nabla(\boldsymbol{v}\cdot\nabla z). \tag{2.3.3c}$$

The quantity $\boldsymbol{v}$ is the (depth-averaged) velocity vector field and is defined such that $\boldsymbol{q} := \mathsf{h}\boldsymbol{v}$. The quantity $\dot{\mathsf{k}}$ can be interpreted as the vertical acceleration of the fluid particles induced by the topography while $\ddot{\mathsf{h}} + \dot{\mathsf{k}}$ can be interpreted as the vertical acceleration of the fluid particles at the free surface ([56, 6]).

2.3.2 Admissible set and Lake-at-rest

The admissible set (or invariant domain) for the Serre Equations is defined by:

$$\mathcal{A} = \{\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^{\mathsf{T}} \mid \mathsf{h} > 0.\} \tag{2.3.4}$$

The admissible set is physically meaningful since the water depth h can be thought of as the distance between the free-surface elevation and topography map. A numerical method that preserves the invariant domain at the discrete level can be classified as invariant domain preserving. Since the admissible set contains only the positivity condition on the water depth$\mathsf{h}$, a numerical method that preserves this property can also be called *positivity-preserving*. These methods are robust and preserve the structure of their continuous counterparts without the requirement of tune-able parameters.

Proposition 2.3.1 (Lake-at-rest for the Serre Equations)**.** *If the flow is at rest ($\boldsymbol{q} \equiv \mathbf{0}$ for $\boldsymbol{x} \in D$), then the lake-at-rest steady-state problem for the Serre Equations* (2.3.1) *is given by:*

$$g\mathsf{h}\nabla(\mathsf{h} + z) = \mathbf{0}, \quad \boldsymbol{x} \in D. \tag{2.3.5}$$

Proof. Assume the flow is at rest $\boldsymbol{q} \equiv 0$ (and $\boldsymbol{v} \equiv \mathbf{0}$). Since we are interested in steady-state solutions (i.e., no dependence on the time variable t), we have that $\dot{\mathsf{h}} = \boldsymbol{v}\cdot\nabla\mathsf{h}$. However, by assumption, $\boldsymbol{v} = \mathbf{0}$ so $\dot{\mathsf{h}} = 0$. Consequently, $\ddot{\mathsf{h}} = 0$. The rest of the proof is exactly the same as Proof 2.2.2. □

Remark 2.3.2 (Lake-at-rest solution). The lake-at-rest problem (2.3.5) has one of two solutions: (i) $\mathsf{h} = 0$; or (ii) $\mathsf{h} + z =$ constant. The physical interpretation of the solution $\mathsf{h} + z =$ constant is as follows: If the flow is at rest, then the water elevation must remain constant.

We call numerical methods that preserve the lake-at-rest property at the discrete level *well-balanced*.

2.3.3 Mathematical reinterpretation of dispersive terms

In this section, we give different representation of the Serre model that are equivalent. Let D_t be the linear, total derivative operator defined by

$$\mathsf{D}_t\phi := (\partial_t + \boldsymbol{v}\cdot\nabla)\phi, \tag{2.3.6}$$

where ϕ is an arbitrary smooth function.

Proposition 2.3.3 (First alternative representation)**.** *A first alternative representation of the Serre Equations is given by:*

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.3.7a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla\left(\frac{1}{2}g\mathsf{h}^2 + \frac{1}{3}\mathsf{h}^2\mathsf{D}_t^2(\mathsf{h}+\frac{3}{2}z)\right) = -\left(g\mathsf{h} + \mathsf{D}_t^2(\frac{1}{2}\mathsf{h}+z)\right)\nabla z \tag{2.3.7b}$$

Proof. Assume $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ is a smooth solution to the Serre model. Let $z(\boldsymbol{x})$ be a given, smooth topography profile. Notice that only the non-hydrostatic pressure and source are represented differently from (2.3.1)–(2.3.2). Expanding the D_t^2 term in the pressure gives:

$$\begin{aligned}
\frac{1}{3}\mathsf{h}^2\mathsf{D}_t^2(\mathsf{h}+\frac{3}{2}z) &= \frac{1}{3}\mathsf{h}^2\mathsf{D}_t\left(\mathsf{D}_t(\mathsf{h}+\frac{3}{2}z)\right)\\
&= \frac{1}{3}\mathsf{h}^2\mathsf{D}_t\left(\dot{\mathsf{h}} + \mathsf{D}_t(\frac{3}{2}z)\right)\\
&= \frac{1}{3}\mathsf{h}^2\mathsf{D}_t\left(\dot{\mathsf{h}} + \frac{3}{2}\boldsymbol{v}\cdot\nabla z\right)\\
&= \frac{1}{3}\mathsf{h}^2\left(\ddot{\mathsf{h}} + \frac{3}{2}\dot{\mathsf{k}}\right)\\
&= \mathsf{h}^2\left(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}}\right)
\end{aligned}$$

which is the expression in (2.3.2). A similar expansion can be performed for the topography source term. □

The representation shown in Proposition 2.3.3 was first shown in Green and Naghdi [25] (after

some algebraic manipulation of equations (4.27)-–(4.31) therein) and can also be seen in Dellar and Salmon [14, Eq. (18)].

Proposition 2.3.4 (Second alternative representation)**.** *A second alternative representation of the Serre Equations is given by:*

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.3.8a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla\left(\frac{1}{2}g\mathsf{h}^2 + \bar{p}(\boldsymbol{u})\right) = -\left(g\mathsf{h} + \bar{r}(\boldsymbol{u})\right)\nabla z, \tag{2.3.8b}$$

$$\tag{2.3.8c}$$

where

$$\bar{p}(\boldsymbol{u}) = -\frac{1}{3}\mathsf{h}^2\left(\mathsf{h}\Big(\partial_t(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2 + \boldsymbol{v}\cdot\nabla(\nabla\cdot\boldsymbol{v})\Big) - \frac{3}{2}\dot{\mathsf{k}}\right), \tag{2.3.9a}$$

$$\bar{r}(\boldsymbol{u}) = -\frac{1}{2}\mathsf{h}\Big(\partial_t(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2 + \boldsymbol{v}\cdot\nabla(\nabla\cdot\boldsymbol{v})\Big) + \dot{\mathsf{k}}. \tag{2.3.9b}$$

Proof. Assume $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ is a smooth solution to the Serre model. Let $z(\boldsymbol{x})$ be a given, smooth topography profile. In the above formulation, we have expanded the quantity $\ddot{\mathsf{h}}$ using the mass

conservation equation (2.3.1a) as follows:

$$
\begin{aligned}
\ddot{\mathsf{h}} &= \mathsf{D}_t(\dot{\mathsf{h}}) \\
&= \mathsf{D}_t\left(\underbrace{\partial_t \mathsf{h} + \boldsymbol{v}\cdot\nabla\mathsf{h}}_{\text{M.C (2.3.1a)}}\right) \\
&= \mathsf{D}_t\left(-\mathsf{h}\nabla\cdot\boldsymbol{v}\right) \\
&= -\left(\partial_t(\mathsf{h}\nabla\cdot\boldsymbol{v}) + \boldsymbol{v}\cdot\nabla(\mathsf{h}\nabla\cdot\boldsymbol{v})\right) \\
&= -\left((\nabla\cdot\boldsymbol{v})\underbrace{\partial_t \mathsf{h}}_{\text{M.C (2.3.1a)}} + \mathsf{h}\partial_t(\nabla\cdot\boldsymbol{v}) + \boldsymbol{v}\cdot\left((\nabla\cdot\boldsymbol{v})\nabla\mathsf{h} + \mathsf{h}\nabla(\nabla\cdot\boldsymbol{v})\right)\right) \\
&= -\left((\nabla\cdot\boldsymbol{v})(-\boldsymbol{v}\cdot\nabla\mathsf{h} - \mathsf{h}(\nabla\cdot\boldsymbol{v})) + \mathsf{h}\partial_t(\nabla\cdot\boldsymbol{v}) + \boldsymbol{v}\cdot\left((\nabla\cdot\boldsymbol{v})\nabla\mathsf{h} + \mathsf{h}\nabla(\nabla\cdot\boldsymbol{v})\right)\right) \\
&= -\mathsf{h}\left(\partial_t(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2 + \boldsymbol{v}\cdot\nabla(\nabla\cdot\boldsymbol{v})\right)
\end{aligned}
$$

which is the expression in (2.3.2). A similar expansion can be performed for the topography source term. □

The representation shown in Proposition 2.3.4 was first presented in Serre [57] and Su and Gardner [58] for one spatial dimension and no topography effects. This representation shows up naturally in the in the derivation of the Serre–Green–Naghdi model as shown in Section 1.4 (see equation (1.4.38)). Notice that we can re-write the momentum equation (2.3.8b) as an evolution equation for the velocity $\boldsymbol{v}$ as follows:

$$
\partial_t\boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + \frac{1}{\mathsf{h}}\nabla(\frac{1}{2}g\mathsf{h}^2 + \bar{p}(\boldsymbol{u})) = -\frac{1}{\mathsf{h}}(g\mathsf{h} + \bar{r}(\boldsymbol{u}))\nabla z,
$$

with

$$
\begin{aligned}
\bar{p}(\boldsymbol{u}) &= -\frac{1}{3}\mathsf{h}^2\left(\mathsf{h}\Big(\partial_t(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2 + \boldsymbol{v}\cdot\nabla(\nabla\cdot\boldsymbol{v})\Big) - \frac{3}{2}(\partial_t\boldsymbol{v}\cdot\nabla z + \boldsymbol{v}\nabla(\boldsymbol{v}\cdot\nabla z))\right), \\
\bar{r}(\boldsymbol{u}) &= -\frac{1}{2}\mathsf{h}\Big(\partial_t(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2 + \boldsymbol{v}\cdot\nabla(\nabla\cdot\boldsymbol{v})\Big) + \partial_t\boldsymbol{v}\cdot\nabla z + \boldsymbol{v}\nabla(\boldsymbol{v}\cdot\nabla z).
\end{aligned}
$$

Proposition 2.3.5 (Third alternative representation)**.** *A third alternative representation of the Serre Equations is given by:*

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0, \tag{2.3.11a}$$

$$\mathsf{h}\left(1 + \frac{1}{\mathsf{h}}\mathbb{T}\right)\left[\partial_t \boldsymbol{v}\right] + \mathsf{h}(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + g\mathsf{h}\nabla(\mathsf{h}+z) \\ - \frac{1}{3}\nabla\{\mathsf{h}^3[(\boldsymbol{v}\cdot\nabla)(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2]\} + \mathsf{h}\mathbb{Q}[\boldsymbol{v}] = 0. \tag{2.3.11b}$$

Here, the operators $\frac{1}{\mathsf{h}}\mathbb{T}[\cdot]$ *and* $\mathbb{Q}[\cdot]$ *are defined by:*

$$\frac{1}{\mathsf{h}}\mathbb{T}[\boldsymbol{w}] = -\frac{1}{3\mathsf{h}}\nabla(\mathsf{h}^3(\nabla\cdot\boldsymbol{w})) - \frac{1}{2}\mathsf{h}(\nabla\cdot\boldsymbol{w})\nabla z + \frac{1}{2\mathsf{h}}\nabla(\mathsf{h}^2\nabla z\cdot\boldsymbol{w}) + (\nabla z\cdot\boldsymbol{w})\nabla z,$$
$$\mathbb{Q}[\boldsymbol{w}] = \frac{1}{2\mathsf{h}}\nabla(\mathsf{h}^2(\boldsymbol{w}\cdot\nabla)^2 z) - \frac{1}{2}\mathsf{h}\left((\boldsymbol{w}\cdot\nabla)(\nabla\cdot\boldsymbol{w}) - (\nabla\cdot\boldsymbol{w})^2\right)\nabla z + ((\boldsymbol{w}\cdot\nabla)^2 z)\nabla z.$$

with $\boldsymbol{w}$ *is a sufficiently smooth vector-valued function.*

Proof. Let $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ be a smooth solution to the system (2.3.11). Since there is no change to the mass conservation equation, we want to show that (2.3.11b) is equivalent to (2.3.1b)–(2.3.2). To simplify the computations, we will give certain terms an integer marker to signify that they will be combined. Let us recall two identities that will be useful in the following analysis. When the solution $\boldsymbol{u}$ and topography map $z(\boldsymbol{x})$ are smooth, we have that:

$$\dot{\mathsf{k}} = \partial_t(\boldsymbol{v}\cdot\nabla z) + (\boldsymbol{v}\cdot\nabla)^2 z,$$
$$\mathsf{h}^2\ddot{\mathsf{h}} = -\mathsf{h}^3\left(\partial_t(\nabla\cdot\boldsymbol{v}) + (\boldsymbol{v}\cdot\nabla)(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2\right),$$

where we used that $(\boldsymbol{v}\cdot\nabla)^2 z = \boldsymbol{v}\cdot\nabla(\boldsymbol{v}\cdot\nabla z)$. Using the second identity, we see that:

$$-\frac{1}{3}\nabla\{\mathsf{h}^3[(\boldsymbol{v}\cdot\nabla)(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2]\} \;=\; \nabla(\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) + \underbrace{\nabla(\frac{1}{3}\mathsf{h}^3\partial_t(\nabla\cdot\boldsymbol{v}))}_{(1)}.$$

Note that $g\mathsf{h}\nabla(\mathsf{h}+z) = \nabla(\frac{1}{2}g\mathsf{h}^2) + g\mathsf{h}\nabla z$. Let us now focus on the first two terms of (2.3.11b). We have that

$$\begin{aligned}\mathsf{h}\left(1+\frac{1}{\mathsf{h}}\mathbb{T}\right)[\partial_t \boldsymbol{v}] + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} &= \underbrace{\mathsf{h}\left(\partial_t \boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\right)}_{\text{Prop. 3.2.1}} + \mathsf{h}\frac{1}{\mathsf{h}}\mathbb{T}[\partial_t \boldsymbol{v}] \\ &= \partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{q}\otimes\boldsymbol{v}) + \mathsf{h}\frac{1}{\mathsf{h}}\mathbb{T}[\partial_t \boldsymbol{v}].\end{aligned}$$

Expanding the last term separately yields:

$$\mathsf{h}\frac{1}{\mathsf{h}}\mathbb{T}[\partial_t \boldsymbol{v}] = \underbrace{-\frac{1}{3}\nabla(\mathsf{h}^3(\nabla\cdot\partial_t\boldsymbol{v}))}_{(1)}\underbrace{-\frac{1}{2}\mathsf{h}^2(\nabla\cdot\partial_t\boldsymbol{v})\nabla z}_{(2)} + \underbrace{\frac{1}{2}\nabla(\mathsf{h}^2\nabla z\cdot\partial_t\boldsymbol{v})}_{(3)} + \underbrace{\mathsf{h}(\nabla z\cdot\partial_t\boldsymbol{v})\nabla z}_{(4)},$$

We now expand each term the topography term $\mathsf{h}\mathbb{Q}[\boldsymbol{v}]$:

$$\begin{aligned}\frac{1}{2}\nabla(\mathsf{h}^2(\boldsymbol{v}\cdot\nabla)^2 z) &= \nabla(\mathsf{h}^2\frac{1}{2}\dot{\mathsf{k}}) - \nabla(\frac{1}{2}\mathsf{h}^2\partial_t(\boldsymbol{v}\cdot\nabla z)), \\ &= \nabla(\mathsf{h}^2\frac{1}{2}\dot{\mathsf{k}}) - \underbrace{\nabla(\frac{1}{2}\mathsf{h}^2\partial_t\boldsymbol{v}\cdot\nabla z)}_{(3)},\end{aligned}$$

$$-\frac{1}{2}\mathsf{h}^2\left((\boldsymbol{v}\cdot\nabla)(\nabla\cdot\boldsymbol{v}) - (\nabla\cdot\boldsymbol{v})^2\right)\nabla z = \frac{1}{2}\mathsf{h}\ddot{\mathsf{h}}\nabla z + \underbrace{\frac{1}{2}\mathsf{h}^2(\nabla\cdot\partial_t\boldsymbol{v})\nabla z}_{(2)},$$

$$\begin{aligned}\mathsf{h}((\boldsymbol{v}\cdot\nabla)^2 z)\nabla z &= \mathsf{h}\dot{\mathsf{k}}\nabla z - \mathsf{h}\partial_t(\boldsymbol{v}\cdot\nabla z)\nabla z \\ &= \mathsf{h}\dot{\mathsf{k}}\nabla z \underbrace{-\mathsf{h}(\partial_t\boldsymbol{v}\cdot\nabla z)\nabla z}_{(4)}.\end{aligned}$$

We now combine all the expanded terms. Note that the terms with the integer markers will all cancel. Thus, we have that:

$$\partial_t\boldsymbol{q} + \nabla\cdot(\boldsymbol{q}\otimes\boldsymbol{v}) + \nabla(\frac{1}{2}g\mathsf{h}^2 + \mathsf{h}^2(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}})) = -(g\mathsf{h} + \mathsf{h}(\frac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}}))\nabla z,$$

which is (2.3.1b)–(2.3.2). □

This representation was first introduced in Lannes and Bonneton [46, Eq. (26)] for the non-dimensional Serre–Green–Naghdi Equations and can be seen in Bonneton et al. [10, Eq. (13)] and Marche [50, Eq. (3)].

2.3.4 Challenges

In this section, we briefly discuss some of the challenges involved in the general scientific process for solving the Serre–Green–Naghdi equations. In particular, we discuss the time step restriction that arises in the numerical approximation of the equations, the process of verification and validation and the status of boundary conditions. The key scientific question of this thesis is presented in this section. For an overview on open problems regarding the Serre–Green–Naghdi equations, we refer the reader to Lannes [44].

2.3.4.1 Dispersive time step restriction

As shown in Chapter 1, the Serre–Green–Naghdi equations are a better approximation of the free-surface Euler Equations than the Saint-Venant model and are able to accurately represent wave transformations in near-shore dynamics when dispersive and non-linear effects are strong. However, this physical accuracy comes at a price. A major drawback of the dispersive Serre model from a numerical perspective is that it involves third-order derivatives in space (this can be seen directly in (2.3.8b)–(2.3.9a)). More specifically, due to the presence of the $\mathcal{O}(\mu^2)$ dispersive terms, the pressure mapping $\boldsymbol{u} \mapsto p(\boldsymbol{u})$ defined in (2.3.2) is not a function, but a second-order differential operator in space and time. This second-order differential operator produces third-order spatial derivatives in the momentum equations (2.3.1b) and rule out any approximation technique that is explicit in time, since this would require that the time step τ satisfy:

$$\tau \sim \mathcal{O}(h^3)V^{-1}L^{-2},$$

where h is the mesh-size, V is a characteristic wave speed scale, and L is a characteristic length scale. We call this condition the dispersive time step restriction. There are currently two popular classes of techniques for addressing this difficulty. The first one is based on Strang's operator splitting and combines explicit and implicit time stepping, see for instance Bonneton et al. [10], Samii and Dawson [55], Duran and Marche [15]. Another approach consists of reinterpreting the dispersive system as a constrained first-order system and then relaxing the constraints to allow for using the typical hyperbolic CFL condition (see Chapters 3 and 4).

The goal of this work is concerned with answering the following scientific question:

> *Is it possible to construct an explicit, well-balanced numerical method for solving the Serre–Green–Naghdi equations that is at least second-order in space and time and invariant-domain preserving under the hyperbolic CFL condition?*

2.3.4.2 Verification and validation

Verification and validation are paramount processes in scientific computing. Analytical solutions are needed to verify the accuracy of numerical codes and can be useful for understanding the structure of mathematical models. At this time, the availability of analytical solutions for the Serre model with non-trivial topography is non-existent. The process of model validation is useful for investigating the physical limitations of the mathematical model of interest (such as the Serre Equations) and is often done by comparing numerical results with experimental data. However, the availability of reliable experimental data can be scarce.

2.3.4.3 Boundary conditions

Most simulations involving water waves have large-time scales especially those modeling large domains such as the ocean. Since we can only perform computations on finite domains, these large-time scales can be an issue if we want to avoid wave reflections at the boundary. One can try to limit these reflections through wave absorption zones introduced in the domain, but these techniques are often limited and require more computational resources. Another approach is to carefully enforce the "correct" boundary conditions so that information flows out of the boundary.

However, this approach is highly non-trivial and is PDE dependent.

In the following chapters, we work towards answering the scientific equation 2.3.4.1 along with addressing the above the difficulties.

2.3.5 Properties

In this section, we derive important properties of the Serre–Green–Naghdi model (2.3.1)–(2.3.2). In particular, we derive a conservation equation for the total energy of the system. We then derive analytical solutions for the Serre model with topography effects which are used for verifying the accuracy of numerical methods. Finally, we derive the dispersion relation for the linearized model. The techniques and results shown in this section will be useful when studying the hyperbolic relaxation technique introduced in Chapter 4.

2.3.5.1 Conservation of energy

We first derive the conservation of energy equation for the Serre system (2.3.1)–(2.3.2). This is done as follows. We first derive a conservation equation for the kinetic energy (see (2.3.13)). Then, we derive a conservation equation for the potential energy (see (2.3.14)). Finally, we combine the two equations and obtain a single conservation equation for the total energy of the system. This technique is a standard approach for deriving energy results for similar mathematical models such as the Saint-Venant model. The main result of this section is given in Proposition 2.3.6.

Let $\boldsymbol{u}(\boldsymbol{x},t) := (\mathsf{h}(\boldsymbol{x},t), \boldsymbol{q}(\boldsymbol{x},t))^{\mathsf{T}}$ be a smooth solution to the Serre model (2.3.1)–(2.3.2). We first derive the conservation equation for the kinetic energy by applying the operator $\boldsymbol{v}\cdot()$ to (2.3.1b) and combining the results. For completeness and clarity, we simplify each term individually and then combine the terms at the end (note that when convenient, we interchange the notation $\dot{\mathsf{h}} =$

$\mathsf{D}_t\mathsf{h}$). Expanding $\boldsymbol{v}\cdot(\partial_t\boldsymbol{q}+\nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}))$ yields:

$$\begin{aligned}
\boldsymbol{v}\cdot\partial_t\boldsymbol{q} &= \boldsymbol{v}\cdot\partial_t(\mathsf{h}\boldsymbol{v}) \\
&= \|\boldsymbol{v}\|^2\partial_t\mathsf{h}+\mathsf{h}(\frac{1}{2}\partial_t\|\boldsymbol{v}\|^2) \\
&= \|\boldsymbol{v}\|^2\partial_t\mathsf{h}+\frac{1}{2}\partial_t(\mathsf{h}\|\boldsymbol{v}\|^2)-\frac{1}{2}\|\boldsymbol{v}\|^2\partial_t\mathsf{h} \\
&= \partial_t(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)-\frac{1}{2}\|\boldsymbol{v}\|^2\nabla\cdot(\mathsf{h}\boldsymbol{v}) \\
&= \partial_t(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)-\nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)+\mathsf{h}\boldsymbol{v}\cdot\nabla(\frac{1}{2}\|\boldsymbol{v}\|^2),
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{v}\cdot(\nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q})) &= \boldsymbol{v}\cdot(\nabla\cdot(\boldsymbol{v}\otimes\mathsf{h}\boldsymbol{v})), \\
&= \boldsymbol{v}\cdot\left(\mathsf{h}\boldsymbol{v}(\nabla\cdot\boldsymbol{v})+(\boldsymbol{v}\cdot\nabla\mathsf{h})\boldsymbol{v}+\mathsf{h}\nabla(\frac{1}{2}\|\boldsymbol{v}\|^2)\right) \\
&= \|\boldsymbol{v}\|^2(\mathsf{h}\nabla\cdot\boldsymbol{v})+\|\boldsymbol{v}\|^2(\boldsymbol{v}\cdot\nabla\mathsf{h})+\mathsf{h}\boldsymbol{v}\cdot(\nabla\frac{1}{2}\|\boldsymbol{v}\|^2) \\
&= \|\boldsymbol{v}\|^2\nabla\cdot(\mathsf{h}\boldsymbol{v})+\mathsf{h}\boldsymbol{v}\cdot(\nabla\frac{1}{2}\|\boldsymbol{v}\|^2).
\end{aligned}$$

Note that:

$$\begin{aligned}
-\nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)+\mathsf{h}\boldsymbol{v}\cdot\nabla(\|\boldsymbol{v}\|^2)+\|\boldsymbol{v}\|^2\nabla\cdot(\mathsf{h}\boldsymbol{v}) &= -\nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)+\nabla\cdot(\mathsf{h}\boldsymbol{v}\|\boldsymbol{v}\|^2) \\
&= \nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2).
\end{aligned}$$

Expanding $\boldsymbol{v}\cdot\nabla(p(\boldsymbol{u}))=\boldsymbol{v}\cdot\nabla(\frac{1}{2}g\mathsf{h}^2+\mathsf{h}^2(\frac{1}{3}\ddot{\mathsf{h}}+\frac{1}{2}\dot{\mathsf{k}}))$ yields:

$$\boldsymbol{v}\cdot\nabla(\frac{1}{2}g\mathsf{h}^2) = \boldsymbol{v}\cdot\nabla(\frac{1}{2}g\mathsf{h}^2),$$

$$
\begin{aligned}
\boldsymbol{v}\cdot\nabla(\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) &= \nabla\cdot(\boldsymbol{v}\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) - \frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}\underbrace{(\nabla\cdot\boldsymbol{v})}_{=-\dot{\mathsf{h}}/\mathsf{h}} \\
&= \nabla\cdot(\boldsymbol{v}\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) + \frac{1}{3}\mathsf{h}\dot{\mathsf{h}}\ddot{\mathsf{h}} \\
&= \nabla\cdot(\boldsymbol{v}\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) + \underbrace{\frac{1}{3}\mathsf{h}\mathsf{D}_t(\frac{1}{2}(\dot{\mathsf{h}})^2)}_{\text{Prop. 3.2.1}} \\
&= \nabla\cdot(\boldsymbol{v}\frac{1}{3}\mathsf{h}^2\ddot{\mathsf{h}}) + \partial_t(\frac{1}{6}\mathsf{h}(\dot{\mathsf{h}})^2) + \nabla\cdot(\boldsymbol{v}\frac{1}{6}\mathsf{h}(\dot{\mathsf{h}})^2),
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{v}\cdot\nabla(\frac{1}{2}\mathsf{h}^2\dot{\mathsf{k}}) &= \nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}^2\dot{\mathsf{k}}) - \frac{1}{2}\mathsf{h}^2\dot{\mathsf{k}}(\nabla\cdot\boldsymbol{v}) \\
&= \nabla\cdot(\boldsymbol{v}\frac{1}{2}\mathsf{h}^2\dot{\mathsf{k}}) + \underbrace{\frac{1}{2}\mathsf{h}\dot{\mathsf{h}}\dot{\mathsf{k}}}_{*}.
\end{aligned}
$$

Expanding $\boldsymbol{v}\cdot(r(\boldsymbol{u})\nabla z) = \boldsymbol{v}\cdot((g\mathsf{h} + \mathsf{h}(\frac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}})\nabla z))$ yields:

$$
\boldsymbol{v}\cdot(g\mathsf{h}\nabla z) = g\mathsf{h}\boldsymbol{v}\cdot\nabla z,
$$

$$
\begin{aligned}
\boldsymbol{v}\cdot\left(\frac{1}{2}\mathsf{h}\ddot{\mathsf{h}}\nabla z + \mathsf{h}\dot{\mathsf{k}}\nabla z\right) &= \frac{1}{2}\mathsf{h}\ddot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z) + \mathsf{h}(\boldsymbol{v}\cdot\nabla z)\dot{\mathsf{k}} \\
&= \underbrace{\frac{1}{2}\mathsf{h}\ddot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z)}_{*} + \underbrace{\mathsf{h}\mathsf{D}_t\left(\frac{1}{2}(\boldsymbol{v}\cdot\nabla z)^2\right)}_{*}.
\end{aligned}
$$

The terms with the '$*$' label can be simplified as follows:

$$
\begin{aligned}
\frac{1}{2}\mathsf{h}\dot{\mathsf{h}}\dot{\mathsf{k}} + \frac{1}{2}\mathsf{h}\ddot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z) + \mathsf{h}\mathsf{D}_t(\frac{1}{2}(\boldsymbol{v}\cdot\nabla z)^2) &= \underbrace{\mathsf{h}\mathsf{D}_t(\frac{1}{2}(\boldsymbol{v}\cdot\nabla z)^2)}_{\text{Prop. 3.2.1}} + \underbrace{\mathsf{h}\mathsf{D}_t\big(\frac{1}{2}\dot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z)\big)}_{\text{Prop. 3.2.1}} \\
&= \partial_t\big(\frac{1}{2}\mathsf{h}(\boldsymbol{v}\cdot\nabla z)^2\big) + \nabla\cdot\big(\boldsymbol{v}\frac{1}{2}\mathsf{h}(\boldsymbol{v}\cdot\nabla z)^2\big) \\
&\qquad + \partial_t\big(\frac{1}{2}\mathsf{h}\dot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z)\big) + \nabla\cdot\big(\boldsymbol{v}\frac{1}{2}\mathsf{h}\dot{\mathsf{h}}(\boldsymbol{v}\cdot\nabla z)\big).
\end{aligned}
$$

Then, the conservation equation for the kinetic energy can be written as follows:

$$\partial_t \left(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2 + \frac{1}{6}\mathsf{h}\big(\big(\dot{\mathsf{h}} + \frac{3}{2}(\boldsymbol{v}\cdot\nabla z)\big)^2 + \frac{3}{4}(\boldsymbol{v}\cdot\nabla z)^2\big)\right) \\ + \nabla\cdot\left(\boldsymbol{v}\left(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2 + \frac{1}{6}\mathsf{h}\big(\big(\dot{\mathsf{h}} + \frac{3}{2}(\boldsymbol{v}\cdot\nabla z)\big)^2 + \frac{3}{4}(\boldsymbol{v}\cdot\nabla z)^2\big)\right)\right) + \boldsymbol{v}\cdot\nabla(\frac{1}{2}g\mathsf{h}^2) + \mathsf{h}\boldsymbol{v}\cdot\nabla(gz) = 0. \tag{2.3.13}$$

To find a conservation equation for the potential energy, we multiply the mass conversation (2.3.1a) by $g(\mathsf{h}+z)$ and combine the results. Similarly as above, we expand each term individually. Expanding $g(\mathsf{h}+z)(\partial_t\mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}))$ yields

$$\begin{aligned} g(\mathsf{h}+z)\partial_t\mathsf{h} &= g\mathsf{h}\partial_t\mathsf{h} + gz\partial_t\mathsf{h} \\ &= \partial_t(\frac{1}{2}g\mathsf{h}^2 + g\mathsf{h}z), \end{aligned}$$

$$\begin{aligned} g(\mathsf{h}+z)\nabla\cdot(\mathsf{h}\boldsymbol{v}) &= g\mathsf{h}\left(\boldsymbol{v}\cdot\nabla\mathsf{h} + \mathsf{h}\nabla\cdot\boldsymbol{v}\right) + gz\nabla\cdot(\mathsf{h}\boldsymbol{v}) \\ &= g\boldsymbol{v}\cdot\mathsf{h}\nabla\mathsf{h} + g\mathsf{h}^2\nabla\cdot\boldsymbol{v} + gz\nabla\cdot(\mathsf{h}\boldsymbol{v}) \\ &= \boldsymbol{v}\cdot\nabla(\frac{1}{2}g\mathsf{h}^2) + \frac{1}{2}g\mathsf{h}^2\nabla\cdot\boldsymbol{v} + \frac{1}{2}g\mathsf{h}^2\nabla\cdot\boldsymbol{v} + gz\nabla\cdot(\mathsf{h}\boldsymbol{v}) \\ &= \nabla\cdot(\frac{1}{2}g\mathsf{h}^2\boldsymbol{v}) + \frac{1}{2}g\mathsf{h}^2(\nabla\cdot\boldsymbol{v}) + gz\nabla\cdot(\mathsf{h}\boldsymbol{v}). \end{aligned}$$

These terms are combined as follows to form the conservation equation for the potential energy:

$$\partial_t\big(\frac{1}{2}g\mathsf{h}^2 + g\mathsf{h}z\big) + \nabla\cdot\big(\boldsymbol{v}\frac{1}{2}g\mathsf{h}^2\big) + \frac{1}{2}g\mathsf{h}^2(\nabla\cdot\boldsymbol{v}) + gz\nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0. \tag{2.3.14}$$

Summing the equations for the kinetic energy (2.3.13) and potential energy (2.3.14) yields the following result.

Proposition 2.3.6. *Let $\boldsymbol{u}$ be a smooth solution to* (2.3.1)–(2.3.2)*, then the following holds true:*

$\partial_t \mathcal{E}(\boldsymbol{u}) + \nabla\cdot(\boldsymbol{\mathcal{F}}(\boldsymbol{u})) = 0$, *with*

$$\mathcal{E}(\boldsymbol{u}) := \tfrac{1}{2} g \mathsf{h}^2 + g z \mathsf{h} + \tfrac{1}{2}\mathsf{h}\boldsymbol{v}^2 + \tfrac{1}{6}\mathsf{h}\Big(\big(\dot{\mathsf{h}} + \tfrac{3}{2}(\boldsymbol{v}\cdot\nabla z)\big)^2 + \tfrac{3}{4}(\boldsymbol{v}\cdot\nabla z)^2\Big), \tag{2.3.15a}$$

$$\boldsymbol{\mathcal{F}}(\boldsymbol{u}) := \boldsymbol{v}(\mathcal{E}(\boldsymbol{u}) + p(\boldsymbol{u})). \tag{2.3.15b}$$

2.3.5.2 Analytical steady-state solution with topography

In this section, we derive a family of analytical steady-state solutions for the one-dimensional dispersive Serre equations. The solutions are used to validate the accuracy of the proposed relaxed model. The main interest of this exact solution is to help verify the accuracy numerical codes for the approximation of the dispersive Serre model with topography. These solutions are used in Chapter 6 to verify the accuracy of the proposed hyperbolic relaxation technique (introduced in Chapter 4) and associated numerical method (introduced in Chapter 5).

We restrict ourselves to one spatial dimension and assume that the solution to (2.3.1)–(2.3.2) is time-independent and smooth. Then, the steady state problem for the Serre equations with topography for $\boldsymbol{u}(x) := (\mathsf{h}(x), q(x))^\mathsf{T}$ is given by:

$$\partial_x(\mathsf{h}v) = 0, \tag{2.3.16a}$$

$$\partial_x\left(\mathsf{h}v^2 + \frac{1}{2}g\mathsf{h}^2 + \mathsf{h}^2\Big(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}}\Big)\right) = -\left(g\mathsf{h} + \mathsf{h}(\frac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}})\right)\partial_x z, \tag{2.3.16b}$$

By (2.3.16a), we see that the flow discharge $q := \mathsf{h}v$ is constant so we let $q \equiv q_0 \in \mathbb{R}$. Since the solution variable $\boldsymbol{u}$ is time-independent and $v = \frac{q_0}{\mathsf{h}}$, we have that

$$\dot{\mathsf{h}} = v\partial_x\mathsf{h} = \frac{q_0}{\mathsf{h}}\partial_x\mathsf{h},$$

$$\ddot{\mathsf{h}} = v\partial_x(v\partial_x\mathsf{h}) = \frac{q_0^2}{\mathsf{h}}\partial_x(\frac{1}{\mathsf{h}}\partial_x\mathsf{h}),$$

$$\dot{\mathsf{k}} = v\partial_x(v\partial_x z) = \frac{q_0^2}{\mathsf{h}}\partial_x(\frac{1}{\mathsf{h}}\partial_x z).$$

We now re-write (2.3.16b) as follows:

$$\partial_x\Big(\frac{q_0^2}{\mathsf{h}}+\frac{1}{2}g\mathsf{h}^2+\mathsf{h}\Big(\frac{1}{3}q_0^2\partial_x(\frac{1}{\mathsf{h}}\partial_x\mathsf{h})+\frac{1}{2}q_0^2\partial_x(\frac{1}{\mathsf{h}}\partial_x z)\Big)\Big) \;=\; -\Big(g\mathsf{h}+\frac{1}{2}q_0^2\partial_x(\frac{1}{\mathsf{h}}\partial_x\mathsf{h})+q_0^2\partial_x(\frac{1}{\mathsf{h}}\partial_x z)\Big)\partial_x z.$$

Then, dividing the above equation by $g\mathsf{h}$ and re-arranging the terms using the `Mathemetica` software [65] yields:

$$\partial_x\left(\frac{1}{g\mathsf{h}}\big(g\mathsf{h}^2+g\mathsf{h}z+\tfrac{1}{2}\tfrac{1}{\mathsf{h}}q_0^2+\tfrac{1}{3}q_0^2\partial_{xx}\mathsf{h}+\tfrac{1}{2}\partial_{xx}z-\tfrac{1}{6}\tfrac{q_0^2}{\mathsf{h}}(\partial_x\mathsf{h})^2+\tfrac{1}{2}\tfrac{q_0^2}{\mathsf{h}}(\partial_x z)^2\big)\right)=0. \qquad (2.3.17)$$

The above can be re-written even further by noticing that since $q\equiv q_0$, we have that:

$$\mathsf{h}^2(\tfrac{1}{3}\ddot{\mathsf{h}}+\tfrac{1}{2}\dot{\mathsf{k}})+\tfrac{1}{6}\mathsf{h}((\dot{\mathsf{h}}+\tfrac{3}{2}(v\partial_x z)^2+\tfrac{3}{4}(v\partial_x z)^2)=\tfrac{1}{3}q_0^2\partial_{xx}\mathsf{h}+\tfrac{1}{2}\partial_{xx}z-\tfrac{1}{6}\tfrac{q_0^2}{\mathsf{h}}(\partial_x\mathsf{h})^2+\tfrac{1}{2}\tfrac{q_0^2}{\mathsf{h}}(\partial_x z)^2.$$

Then, multiplying (2.3.17) by gq_0 gives:

$$\partial_x\left(\frac{q_0}{\mathsf{h}}\Big(g\mathsf{h}^2+g\mathsf{h}z+\tfrac{1}{2}\tfrac{1}{\mathsf{h}}q_0^2+\mathsf{h}^2(\tfrac{1}{3}\ddot{\mathsf{h}}+\tfrac{1}{2}\dot{\mathsf{k}})+\tfrac{1}{6}\mathsf{h}((\dot{\mathsf{h}}+\tfrac{3}{2}(v\partial_x z)^2+\tfrac{3}{4}(v\partial_x z)^2)\Big)\right)=0. \qquad (2.3.18)$$

Upon further inspection, we see this equation is equivalent to the steady-state problem for the energy equation introduced in Proposition 2.3.6:

$$\partial_x\Big(v\big(\mathcal{E}(\boldsymbol{u}(x))+p(\boldsymbol{u}(x))\big)\Big)=0. \qquad (2.3.19)$$

To find a steady-state solution for the Serre model, we need to solve (2.3.19). This is summarized in the following assertion which is the main result from this section.

Proposition 2.3.7. *Let $q_0\in\mathbb{R}$, $a,r\in\mathbb{R}_+$ and let the bathymetry profile be defined by $z(x):=-\frac{1}{2}\frac{a}{(\cosh(rx))^2}$. Then $\mathsf{h}(x)=\mathsf{h}_0(1+\frac{a}{(\cosh(rx))^2})$ with the constant discharge q_0 is a steady state solution to* (2.3.1)–(2.3.2) *if*

$$q:=\pm\sqrt{\frac{(1+a)g\mathsf{h}_0^3}{2}},\qquad r:=\frac{1}{\mathsf{h}_0}\sqrt{\frac{3a}{1+a}}. \qquad (2.3.20)$$

Proof. Let $\boldsymbol{u}(x) := (\mathsf{h}(x), q(x))^{\mathsf{T}}$ be a smooth solution to the Serre model. Notice that the discharge $q \equiv q_0$ is constant since the solution does not depend on time. As shown above, a steady-state solution can be found by solving the steady-state problem of the energy equation in Proposition 2.3.6. Solving this equation yields a Bernoulli-like relation for the dispersive Serre model (2.3.1)–(2.3.2). More precisely, from (2.3.19), we infer that $\partial_x(\mathcal{F}(\boldsymbol{u})) = 0$, which implies $\mathcal{F}(\boldsymbol{u}(x)) = C_{\text{Ber}} q_0 g$ where C_{Ber} is the Bernoulli constant. We look for a stationary wave with the following structure $\mathsf{h}(x) = \mathsf{h}_0(1 + \frac{a}{(\cosh(rx))^2})$ and posit that the topography is of the form $z(x) = \lambda(\mathsf{h}(x) - \mathsf{h}_0)$. The problem now consists of finding relations between the parameters a, r, h_0, g, and λ so that the condition $g^{-1}q_0^{-1}\mathcal{F}(\boldsymbol{u}(x)) = C_{\text{Ber}}$ is satisfied, i.e.,

$$\mathsf{h}(1+\lambda) + \frac{q_0^2}{2g\mathsf{h}^2} - \frac{q_0^2}{6g\mathsf{h}^2}(1-3\lambda^2)(\partial_x \mathsf{h})^2 + \frac{q_0^2}{3g\mathsf{h}}(1+\tfrac{3}{2}\lambda)\partial_{xx}\mathsf{h} = C_{\text{Ber}} + \lambda \mathsf{h}_0,$$

where we used $v = \frac{q_0}{\mathsf{h}}$. By taking the limit of this identity for $|x| \to \infty$, we find that $C_{\text{Ber}} = \mathsf{h}_0 + \frac{q_0^2}{2g\mathsf{h}_0^2}$. After inserting the ansatz $\mathsf{h}(x) = \mathsf{h}_0(1 + \frac{a}{(\cosh(rx))^2})$ into the above identity, we find that the following must hold true for all $x \in \mathbb{R}$:

$$\begin{aligned}
&\big((1+\lambda)g\mathsf{h}_0^3 + (\lambda+\tfrac{2}{3})2r^2q_0^2\mathsf{h}_0^2 - q_0^2\big)\cosh(rx)^4 \\
&\qquad + \big((1+\lambda)2ag\mathsf{h}_0^3 + ((\lambda^2+\lambda+\tfrac{1}{3})a - \tfrac{3}{2}\lambda - 1)2r^2q_0^2\mathsf{h}_0^2 - \tfrac{1}{2}aq_0^2\big)\cosh(rx)^2 \\
&\qquad\qquad + \big((1+\lambda)ag\mathsf{h}_0 - 2(\lambda^2+\tfrac{3}{2}\lambda+\tfrac{2}{3})r^2q_0^2\big)a\mathsf{h}_0^2 = 0.
\end{aligned}$$

This is equivalent to asserting that following nonlinear system of equations has a solution (i.e., setting the coefficients of the polynomial for $\cosh(rx)$ equal to 0):

$$\begin{aligned}
(1+\lambda)ag\mathsf{h}_0 - 2(\lambda^2+\tfrac{3}{2}\lambda+\tfrac{2}{3})r^2q_0^2 &= 0, \\
(1+\lambda)2ag\mathsf{h}_0^3 + ((\lambda^2+\lambda+\tfrac{1}{3})a - \tfrac{3}{2}\lambda - 1)2r^2q_0^2\mathsf{h}_0^2 - \tfrac{1}{2}aq_0^2 &= 0, \\
(1+\lambda)g\mathsf{h}_0^3 + (\lambda+\tfrac{2}{3})2r^2q_0^2\mathsf{h}_0^2 - q_0^2 &= 0.
\end{aligned}$$

The only nontrivial solution to the above system of equations is

$$\lambda = -\tfrac{1}{2}, \quad q_0 = \pm\sqrt{\frac{(1+a)g\mathsf{h}_0^3}{2}}, \quad r = \frac{1}{\mathsf{h}_0}\sqrt{\frac{3a}{1+a}}.$$

□

In Figure 2.2, we illustrate this steady-state solution with the values $g = 9.81\,\mathrm{ms}^{-2}$, $\mathsf{h}_0 = 1\,\mathrm{m}$, $a = 0.2\,\mathrm{m}$ and $q_0 \approx \sqrt{5.886}\mathrm{m}^2\mathrm{s}^{-1}$. The solid blue profile represents the wave elevation $\mathsf{h}(x) + z(x)$.

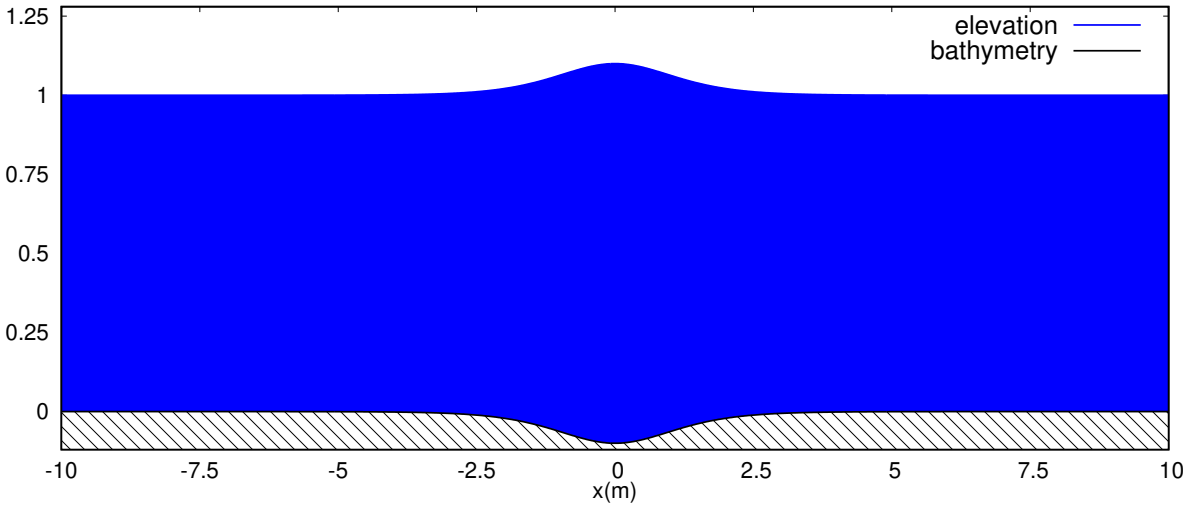


Figure 2.2: 1D Steady state solution for the Serre model

2.3.5.3 Dispersion relation

In this section, we derive the linear dispersion relation for the Serre model (2.3.1). The results shown here will be useful for analyzing the dispersive properties of a hyperbolic relaxed model of the Serre equations in Chapter 4. We note that the derivation shown here is not new and is found in the literature.

For simplicity, we restrict ourselves to one spatial dimension and assume the topography is flat. To derive the linear dispersion relation, we linearize the Serre equations about a rest state $\boldsymbol{u}_0 := (\mathsf{H}_0, 0)^\mathsf{T}$. We consider solutions of the Serre equations of the form $\boldsymbol{u} = \boldsymbol{u}_0 + \bar{\boldsymbol{u}}(x,t)$ where $\bar{\boldsymbol{u}}(x,t)$ is a small perturbation of the rest state. Substituting this solution into the equations and dropping the product of the perturbation terms (since they are small) yields the following linear system of partial differential equations:

$$\partial_t \bar{\mathsf{h}} + \mathsf{H}_0 \partial_x \bar{u} = 0, \tag{2.3.21a}$$

$$\partial_t \bar{u} + g\partial_x \bar{\mathsf{h}} + \frac{1}{3}\mathsf{H}_0 \partial_{xtt} \bar{\mathsf{h}} = 0. \tag{2.3.21b}$$

Assume now that $\bar{\boldsymbol{u}}(\boldsymbol{x},t) = \boldsymbol{u}_A \exp\big(i(kx - \sigma t)\big)$ is a plane-wave solution to (2.3.21) where $\boldsymbol{u}_A$ is the the wave amplitude, k is the wave number and σ is the wave frequency. Here $i := \sqrt{-1}$ is the imaginary unit number. A direct substitution of the plane wave solution into (2.3.21) yields the following linear system for $(\mathsf{h}_A, u_A)^\mathsf{T}$:

$$\begin{pmatrix} -\sigma & \mathsf{H}_0 k \\ gk - \frac{1}{3}\mathsf{H}_0 k\sigma^2 & -\sigma \end{pmatrix} \begin{pmatrix} \mathsf{h}_A \\ u_A \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Assuming that the solution to this linear system is non-trivial, we get the following equation for σ:

$$\sigma^2 \left(1 + \frac{1}{3}\mathsf{H}_0^2 k^2\right) - g\mathsf{H}_0 k^2 = 0. \tag{2.3.22}$$

Solving the above equation for the (squared) phase velocity squared $(c_p^S)^2 := (\frac{\sigma}{k})^2$ yields:

$$(c_p^S)^2 = \frac{g\mathsf{H}_0}{1 + \frac{1}{3}\mathsf{H}_0^2 k^2}. \tag{2.3.23}$$

In Figure 2.3, we plot the squared phase velocity $(c_p^S)^2$ as a function of the wave number k.

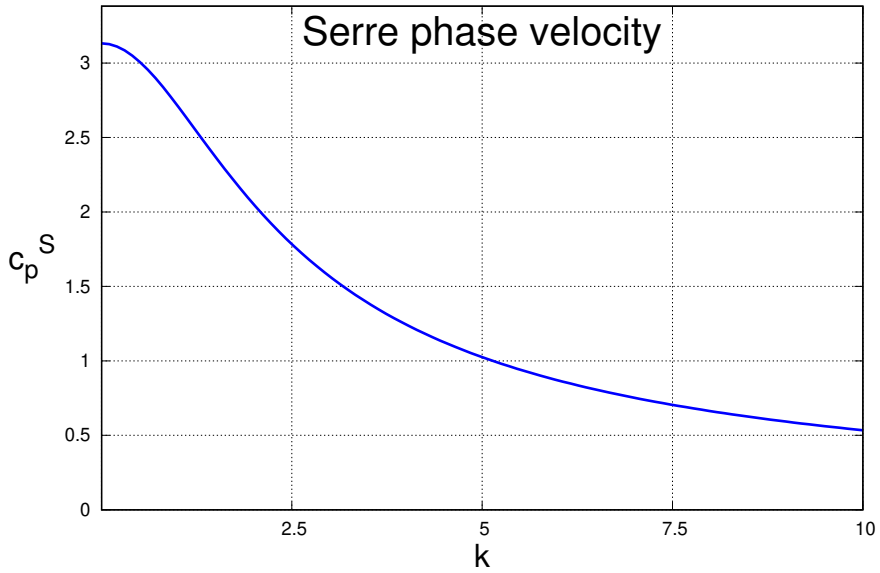


Figure 2.3: A plot of the Serre phase velocity (2.3.23) as a function of the wave number k with $\mathsf{H}_0 = 1\,\mathrm{m}$.

2.4 External physical sources

In this section we describe the mathematical formulation of the external physical sources used for applications in coastal hydrodynamics and inland flooding.

2.4.1 Preliminaries

For notation purposes, consider the condensed form of either the Saint-Venant model or the Serre–Green–Naghdi system with sources:

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

Here, the vector $\boldsymbol{R}(\boldsymbol{u}, \nabla z)$ is the natural topography source that arises in the formulation of the system of interest and 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. The mathematical formulation of the source terms will be presented in model-independent fashion.

2.4.2 Gauckler-Manning friction

Let $\mathbb{e}_{\boldsymbol{q}}$ denote the characteristic vector for the momentum equations. 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:

$$\boldsymbol{S}_{\mathrm{F}}(\boldsymbol{u}) := -gn^2\mathsf{h}^{-\gamma}\boldsymbol{q}\|\boldsymbol{v}\|_{\ell^2}\mathbb{e}_{\boldsymbol{q}}. \tag{2.4.1}$$

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 Chapter 6.

To illustrate the effects of the Gauckler-Manning friction term, we consider the propagation of a solitary wave over a flat-bottom with different values of the Manning's roughness coefficient. We set the computational domain to $D = (0, 1000\,\mathrm{m})$ and initialize the solitary wave with the profiles given by (6.3.1) at $x_0 = 250\,\mathrm{m}$. The final time is set to $T = 50\,\mathrm{s}$. In Figure 2.4, we give the final profiles for the water depth with roughness coefficients of $n = \{0, 0.1, 0.2, 0.3, 0.4\}\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$. We see that as the roughness coefficient increases, the solitary wave slows down and loses amplitude as well. For an overview on selecting the Manning roughness coefficient n for natural channels and flood plains, we refer the reader to Arcement and Schneider [2] (and references therein) where this is discussed.

2.4.3 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: Zhang et al. [68], Madsen et al. [48] 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} the water depth component of $\boldsymbol{u}_{\text{wave}}$, we assume that $\mathsf{h}_{\text{wave}}(\boldsymbol{x}, t) \geq 0$ for all $\boldsymbol{x} \in D$

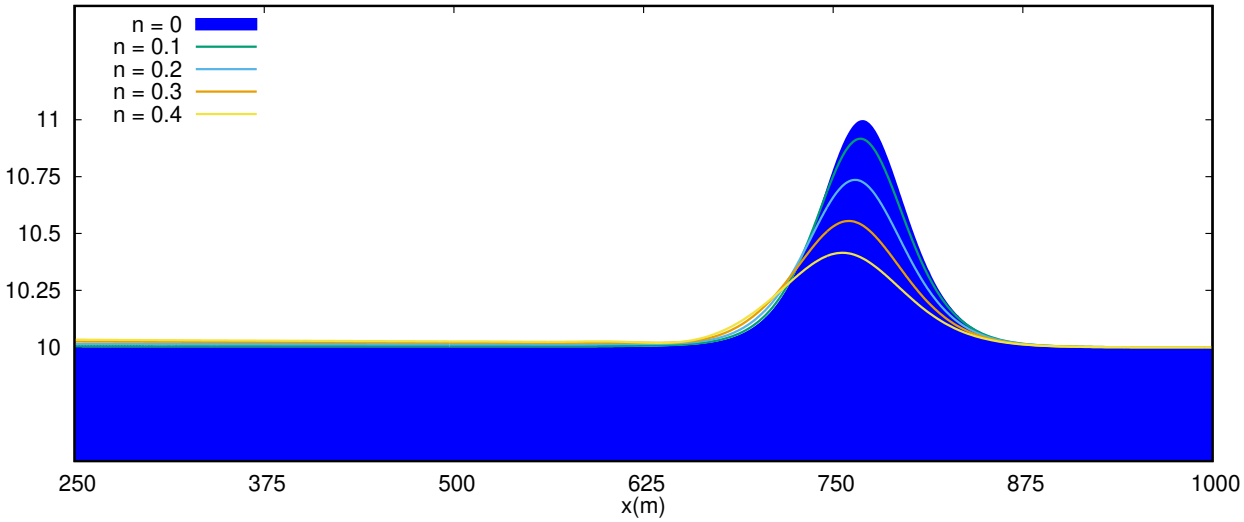


Figure 2.4: Propagation of solitary wave profile over flat-bottom with roughness coefficients of $n = \{0, 0.1, 0.2, 0.3, 0.4\}\mathrm{m}^{-\frac{1}{3}}\mathrm{s}$.

and all $t > 0$. To generate periodic waves through a relaxation zone, we introduce the following source:

$$\boldsymbol{S}_{\mathrm{G}}(\boldsymbol{u}) := -\frac{\sqrt{g\mathsf{H}_0}}{\epsilon}(\boldsymbol{u} - \boldsymbol{u}_{\mathrm{wave}}(\boldsymbol{x}, t))G(\tfrac{x - x_{\min}}{L_{\mathrm{gen}}}), \tag{2.4.2}$$

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:

$$\boldsymbol{S}_{\mathrm{A}}(\boldsymbol{u}) := -\frac{\sqrt{g\mathsf{H}_0}}{\epsilon}G(\tfrac{x_{\max} - x}{L_{\mathrm{abs}}})\boldsymbol{u}_{\mathrm{abs}}, \tag{2.4.3}$$

where $\boldsymbol{u}_{\mathrm{abs}}$ are the conserved variables to be "absorbed". That it to say, we enforce the zero value on

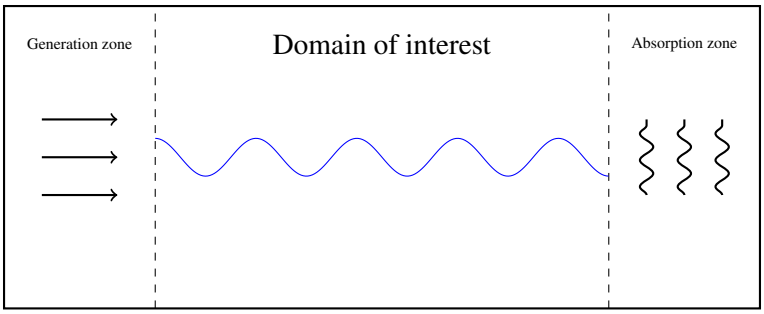


Figure 2.5: Generation and absorption zone schematic.

$\boldsymbol{u}_{\text{abs}}$ to dissipate the waves. In Figure 2.5, we show a simple schematic of the wave generation and absorption zone layout. We now illustrate numerically the wave generation and wave absorption source terms with an example. Let the computational domain be $D = (0, 50\,\text{m})$. We set the still water depth to $\mathsf{h}_0 = 0.4\,\text{m}$. The wave amplitude is set to $a = 0.01\,\text{m}$ and wave period is set to $T_p = 2\,\text{s}^{-1}$. The generation zone was set to $5\,\text{m}$ and the absorption zone to $10\,\text{m}$. In Figure 2.6, we show a space-time plot of the periodic wave solution in the range $t \in [0, 40\,\text{s}]$. We see that the waves are smoothly introduced into the domain of interest as the time increases in the generation zone ($x < 5\,\text{m}$). We then see that waves smoothly dissipate in the absorption zone ($x > 40\,\text{m}$).

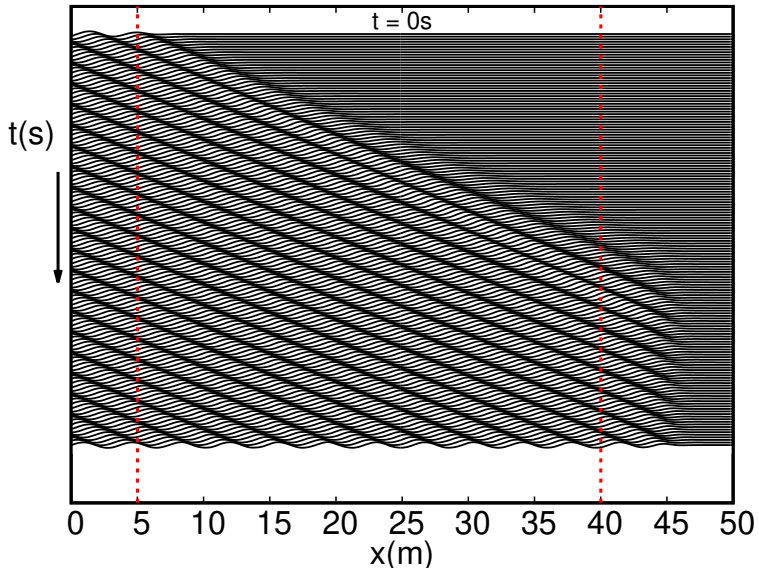


Figure 2.6: Numerical illustration of wave generation/absorption in the space-time domain.

3.1 Introduction

In this chapter, we discuss the mathematical reformulation of the dispersive Serre Equations. The reformulation of the Serre model is of great interest to the coastal modeling community since it allows one to exploit different mathematical structures of the model which open the door for efficient implicit and explicit numerical methods. The strategy of reformulating higher-order partial differential equations is ubiquitous in the literature, so various reformulations of the Serre equations exist.

The chapter is organized as follows. In Section 3.2, we propose a novel reformulation of the Serre model that yields a first-order system under two algebraic constraints. This reformulation is the foundation for a hyperbolic relaxation technique introduced in Chapter 4. In Section 3.3, we give an overview of existing reformulations of the Serre model and give direct comparisons of the reformulation proposed in this work to others seen in the literature.

3.2 Reformulation under two algebraic constraints

We now propose a novel reformulation of the Serre model. The motivation behind the reformulation is to reduce the higher-order dispersive terms in the pressure and topography source terms of the Serre model. This is achieved by introducing three first-order auxiliary equations coupled with two algebraic constraints that take the place of the dispersive terms.

We first recall the Serre model. Let $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ be the (smooth) solution variable where h is the positive water depth and $\boldsymbol{q} \in \mathbb{R}^d$ is the momentum vector where d is the spatial dimension. Let $\boldsymbol{v}$ be the velocity vector defined such that $\boldsymbol{q} := \mathsf{h}\boldsymbol{v}$. Then, the Serre model is formulated as follows:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0,$$
$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla\Big(\frac{1}{2}g\mathsf{h}^2 + \underbrace{\mathsf{h}^2(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}})}_{\widetilde{p}(\boldsymbol{u})}\Big) = -\Big(g\mathsf{h} + \underbrace{\mathsf{h}(\frac{1}{2}\ddot{\mathsf{h}} + \dot{\mathsf{k}})}_{\tilde{r}(\boldsymbol{u})}\Big)\nabla z.$$

The goal moving forward is to replace the non-hydrostatic pressure $\widetilde{p}(\boldsymbol{u})$ and non-hydrostatic part of the source $\tilde{r}(\boldsymbol{u})$ by algebraic expressions. We begin the reformulation of the Serre model with the non-hydrostatic pressure. The reformulation of the topography source term will follow naturally. Rearranging the non-hydrostatic pressure yields:

$$\begin{aligned}\widetilde{p}(\boldsymbol{u}) &= \mathsf{h}^2\left(\frac{1}{3}\ddot{\mathsf{h}}+\frac{1}{2}\dot{\mathsf{k}}\right)\\ &= \frac{1}{3}\mathsf{h}^2\left(\ddot{\mathsf{h}}+\frac{3}{2}\dot{\mathsf{k}}\right)\\ &= \frac{1}{3}\mathsf{h}^2\mathsf{D}_t\left(\dot{\mathsf{h}}+\frac{3}{2}\mathsf{k}\right),\end{aligned}$$

where D_t denotes the total derivative operator (i.e., equivalent to the $\dot{()}$ notation). We use the following property to continue the analysis.

Proposition 3.2.1. *Let* $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ *be a smooth solution of* (2.3.1)*. Assuming that* $\phi(\boldsymbol{x}, t)$ *is an arbitrary smooth function, then*

$$\mathsf{h}\dot{\phi} = \partial_t(\mathsf{h}\phi) + \nabla\cdot(\mathsf{h}\boldsymbol{v}\phi).$$

Proof. A direction computation of the left hand side yields:

$$\begin{aligned}\mathsf{h}\dot{\phi} &= \mathsf{h}\left(\partial_t\phi + \boldsymbol{v}\cdot\nabla\phi\right),\\ &= \partial_t(\mathsf{h}\phi) - \phi\partial_t\mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}\phi) - \phi\nabla\cdot(\mathsf{h}\boldsymbol{v}),\\ &= \partial_t(\mathsf{h}\phi) + \phi\nabla\cdot(\mathsf{h}\boldsymbol{v}) - \cancel{\phi(\partial_t + \nabla\cdot(\mathsf{h}\boldsymbol{v}))},\\ &= \partial_t(\mathsf{h}\phi) + \nabla\cdot(\mathsf{h}\boldsymbol{v}\phi).\end{aligned}$$

□

Using Proposition 3.2.1 and continuing the expansion for the non-hydrostatic pressure, we see:

$$\begin{aligned}\widetilde{p}(\boldsymbol{u}) &= \frac{1}{3}\mathsf{h}^2\mathsf{D}_t\left(\dot{\mathsf{h}} + \frac{3}{2}\mathsf{k}\right) \\ &= \frac{1}{3}\mathsf{h}\Big(\mathsf{h}\mathsf{D}_t(\dot{\mathsf{h}}) + \mathsf{h}\mathsf{D}_t(\frac{3}{2}\mathsf{k})\Big) \\ &= \frac{1}{3}\mathsf{h}\left(\partial_t(\mathsf{h}\dot{\mathsf{h}}) + \nabla\cdot(\boldsymbol{v}\mathsf{h}\dot{\mathsf{h}}) + \frac{3}{2}(\partial_t(\boldsymbol{q}\cdot\nabla z) + \nabla\cdot(\boldsymbol{v}(\boldsymbol{q}\cdot\nabla z)))\right).\end{aligned}$$

Let us introduce the following constraint $q_3 := \boldsymbol{q}\cdot\nabla z$ and variable $\widetilde{s}$ such that

$$\partial_t(q_3) + \nabla\cdot(\boldsymbol{v}q_3) = \widetilde{s}.$$

Notice that this is a first-order conservation law with source $\widetilde{s}$. Also note that the above equation is equivalent to $\mathsf{h}\dot{\mathsf{k}} = \widetilde{s}$. We now introduce another variable s so that

$$-s = \partial_t(\mathsf{h}\dot{\mathsf{h}}) + \nabla\cdot(\boldsymbol{v}\mathsf{h}\dot{\mathsf{h}}) + \frac{3}{2}\widetilde{s} \tag{3.2.1}$$

At this point, we can define the algebraic non-hydrostatic pressure of $\widetilde{p}(\boldsymbol{u}) = -\frac{1}{3}\mathsf{h}s$ and non-hydrostatic source $\tilde{r}(\boldsymbol{u}) = -\frac{1}{2}s + \frac{1}{4}\widetilde{s}$. However, the equation (3.2.1) still contains higher-order derivatives, so we continue the reduction. We introduce the auxiliary relation $q_2 = \mathsf{h}\dot{\mathsf{h}} + \frac{3}{2}q_3$ so that substituting in the (3.2.1) gives:

$$\partial_t(q_2) + \nabla\cdot(\boldsymbol{v}q_2) = -s.$$

This is another first-order conservation law with a source term s. Since q_2 is dependent on $\mathsf{h}\dot{\mathsf{h}}$, we can once again reduce the higher-order terms. By Proposition 3.2.1, we see that:

$$\mathsf{h}\dot{\mathsf{h}} = \partial_t(\mathsf{h}^2) + \nabla\cdot(\boldsymbol{v}\mathsf{h}^2).$$

Introducing the algebraic constraint $q_1 := \mathsf{h}^2$ and using the definitions for q_2 and q_3, the previous

equation simplifies to

$$\partial_t(q_1) + \nabla\cdot(\boldsymbol{v}q_1) = q_2 - \frac{3}{2}q_3.$$

Again, we have arrived at a first-order conservation law with source $q_2 - \frac{3}{2}q_3$. Thus, if $q_1 := \mathsf{h}^2$ and $q_3 := \boldsymbol{q}\cdot\nabla z$, the full Serre pressure and topography source terms in (2.3.2) are equivalently written as:

$$\begin{aligned}
p(\boldsymbol{u}) &= \frac{1}{2}g\mathsf{h}^2 - \frac{1}{3}\mathsf{h}s,\\
r(\boldsymbol{u}) &= g\mathsf{h} - \tfrac{1}{2} + \tfrac{1}{4}\tilde{s},\\
\partial_t q_1 + \nabla\cdot(\boldsymbol{v}q_1) &= q_2 - \frac{3}{2}q_3,\\
\partial_t q_2 + \nabla\cdot(\boldsymbol{v}q_2) &= -s,\\
\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) &= \tilde{s}
\end{aligned}$$

Note that the variables s and $\widetilde{s}$ can be interpreted as Lagrange multipliers associated with the constraints $q_1 = \mathsf{h}^2$ and $q_3 = \boldsymbol{q}\cdot\nabla z$, respectively. The above ideas are summarized in the following lemma.

Lemma 3.2.2. *Let $\boldsymbol{u} : D\times(0,T) \to \mathbb{R}_+\times\mathbb{R}^d$ be a smooth function. Then $\boldsymbol{u}$ solves the dispersive Serre model* (2.3.1)–(2.3.2) *iff $(\boldsymbol{u}, q_1, q_2, q_3)$ solves*

$$\partial_t\mathsf{h} + \nabla\cdot\boldsymbol{q} = 0, \tag{3.2.2a}$$

$$\partial_t\boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla(\tfrac{1}{2}g\mathsf{h}^2 - \tfrac{1}{3}\mathsf{h}s) = -(g\mathsf{h} - \tfrac{1}{2}s + \tfrac{1}{4}\widetilde{s})\nabla z, \tag{3.2.2b}$$

$$\partial_t q_1 + \nabla\cdot(\boldsymbol{v}q_1) = q_2 - \tfrac{3}{2}q_3 \tag{3.2.2c}$$

$$\partial_t q_2 + \nabla\cdot(\boldsymbol{v}q_2) = -s \tag{3.2.2d}$$

$$\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) = \widetilde{s} \tag{3.2.2e}$$

$$q_1 = \mathsf{h}^2, \quad q_3 = \boldsymbol{q}\cdot\nabla z \tag{3.2.2f}$$

Proof. The forward direction of the proof is summarized in the steps above, so we now prove the

converse. Let us assume that $(\boldsymbol{u}, q_1, q_2, q_3)$ solves (3.2.2a)–(3.2.2f).

We set $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})$ and $\boldsymbol{q} := \mathsf{h}\boldsymbol{v}$ where $\boldsymbol{v}$ is the velocity vector. Let us set $\dot{\mathsf{k}} := \partial_t(\boldsymbol{v}\cdot\nabla z) + \boldsymbol{v}\cdot\nabla(\boldsymbol{v}\cdot\nabla z)$. Then using (3.2.2e) and (3.2.2f) we obtain

$$\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) = \partial_t(\mathsf{h}\boldsymbol{v}\cdot\nabla z) + \nabla\cdot(\boldsymbol{v}\mathsf{h}(\boldsymbol{v}\cdot\nabla z)) = \mathsf{h}\mathsf{D}_t(\boldsymbol{v}\cdot\nabla z) = \widetilde{s},$$

which gives $\mathsf{h}\dot{\mathsf{k}} = \widetilde{s}$. Similarly, Proposition 3.2.1 implies that $\mathsf{h}\dot{\mathsf{h}} = \partial_t(\mathsf{h}^2) + \nabla\cdot(\boldsymbol{v}\mathsf{h}^2)$. This identity, together with $q_1 := \mathsf{h}^2$ and (3.2.2c), gives $\mathsf{h}\dot{\mathsf{h}} = q_2 - \frac{3}{2}q_3$. Then

$$\mathsf{h}\ddot{\mathsf{h}} = \partial_t(\mathsf{h}\dot{\mathsf{h}}) + \nabla\cdot(\boldsymbol{v}\mathsf{h}\dot{\mathsf{h}}) = \partial_t(q_2 - \tfrac{3}{2}q_3) + \nabla\cdot(\boldsymbol{v}(q_2 - \tfrac{3}{2}q_3)) = -s - \tfrac{3}{2}\widetilde{s}.$$

This implies that $s = -\mathsf{h}(\ddot{\mathsf{h}} + \frac{3}{2}\dot{\mathsf{k}})$. Let us set $p := \frac{1}{2}g\mathsf{h}^2 - \frac{1}{3}\mathsf{h}s$. Then $p = \frac{1}{2}g\mathsf{h}^2 + \mathsf{h}^2(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}})$. This is the expression of the pressure in (2.3.2a).

Finally let us set $r := g\mathsf{h} - \frac{1}{2}s + \frac{1}{4}\widetilde{s}$. Then the above computations imply that $r = g\mathsf{h} + \frac{1}{2}\mathsf{h}(\ddot{\mathsf{h}} + \frac{3}{2}\dot{\mathsf{k}}) + \frac{1}{4}\mathsf{h}\dot{\mathsf{k}}$, i.e., $r = g\mathsf{h} + \frac{1}{2}\mathsf{h}\ddot{\mathsf{h}} + \dot{\mathsf{k}}$. This is the expression of the source term r in (2.3.2b). Hence we have established that $\boldsymbol{u}$ solves (2.3.1)–(2.3.2). This completes the proof. □

The following is another way to reformulate Lemma 3.2.2.

Proposition 3.2.3 (Co-dimension 2)**.** *Let $\boldsymbol{u} : D\times(0,T) \to \mathbb{R}_+\times\mathbb{R}^d$ be a smooth function. Then $\boldsymbol{u}$ solves the dispersive Serre model* (2.3.1)–(2.3.2) *if and only if $(\boldsymbol{u}, q_1, q_2, q_3)$ solves the quasi-linear first-order system* (3.2.2a)–(3.2.2e) *on the co-dimension 2 manifold* $\{(\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3) \in \mathbb{R}_+\times\mathbb{R}^d\times\mathbb{R}_+\times\mathbb{R}^2 \mid q_1 = \mathsf{h}^2,\ q_3 = \boldsymbol{q}\cdot\nabla z\}$.

Remark 3.2.4 (Initial conditions). Let $(\mathsf{h}_0, \boldsymbol{q}_0)$ be the initial state for (2.3.1)–(2.3.2). Then the corresponding initial state for (3.2.2a)–(3.2.2f) is

$$\mathsf{h}(\boldsymbol{x}, 0) = \mathsf{h}_0(\boldsymbol{x}), \qquad \boldsymbol{q}(\boldsymbol{x}, 0) = \boldsymbol{q}_0(\boldsymbol{x}), \qquad q_1(\boldsymbol{x}, 0) = \mathsf{h}_0(\boldsymbol{x})^2, \tag{3.2.3a}$$

$$q_3(\boldsymbol{x}, 0) = \boldsymbol{q}_0(\boldsymbol{x})\cdot\nabla z, \quad q_2(\boldsymbol{x}, 0) = -\mathsf{h}_0(\boldsymbol{x})\nabla\cdot\boldsymbol{v}_0(\boldsymbol{x}) + \tfrac{3}{2}q_3(\boldsymbol{x}, 0). \tag{3.2.3b}$$

Remark 3.2.5 (Reformulation with a flat-bottom). If no topography effects are considered (i.e., $z(\boldsymbol{x}) \equiv 0$), repeating the above reformulation yields the following (smaller) system of equations for the solution variable $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2)^{\mathsf{T}}$:

$$\partial_t \mathsf{h} + \nabla\cdot\boldsymbol{q} = 0 \tag{3.2.4a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v} \otimes \boldsymbol{q}) + \nabla(\tfrac{1}{2} g \mathsf{h}^2 - \tfrac{1}{3}\mathsf{h} s) = 0, \tag{3.2.4b}$$

$$\partial_t q_1 + \nabla\cdot(\boldsymbol{v} q_1) = q_2 \tag{3.2.4c}$$

$$\partial_t q_2 + \nabla\cdot(\boldsymbol{v} q_2) = -s \tag{3.2.4d}$$

$$q_1 = \mathsf{h}^2 \tag{3.2.4e}$$

3.3 Literature review

There are many ways to reformulate the dispersive Serre model. Two popular approaches in literature are: (i) reformulate the model as a system of first-order conservation equations with sources for the construction of explicit methods; (ii) reformulation of the model with linear differential operators to construct efficient implicit methods. For instance, in Gavrilyuk and Shugrin [22, Eqs. (5.12)–(5.15)] the authors reformulated the model with a flat bottom as a first-order system with a constraint on the divergence of the velocity. In Lannes and Bonneton [46, Eq. (26)], the authors reformulated the momentum evolution equation of the Serre Equations (called Green-Naghdi Equations therein) using two linear differential operators to induce regularizing effects for easing numerical computations. In Bristeau et al. [12, Eq. (50)], the authors recover the same first-order reformulation (with a flat-bottom) originally proposed in [22]. In Escalante et al. [19, Eq. (1)], the authors present a first-order reformulation with the assumption of mild bottom variation (i.e., dropping the terms containing $\nabla\cdot(\nabla z)$ and $\|\nabla z\|^2$). In Fernandez-Nieto et al. [21, Eq. (3.16)], the authors present a 1D first-order reformulation the Serre model with full topography effects with constraints on the divergence of the velocity and the quantity $\boldsymbol{v}\cdot\nabla z$.

To put the present work in perspective with respect to these techniques, we recall the reformulations proposed in [22, 12, 19], but contrary to [19] we keep all the effects induced by the

topography. Keeping the effects induced by the topography gives a reformulation equivalent to that proposed in [21] up to a change of constant. The starting point is again the Serre model. Assume that $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^{\mathsf{T}}$ is a smooth solution to (2.3.1)–(2.3.2). Let

$$w := -\tfrac{1}{2}\mathsf{h}\nabla\cdot\boldsymbol{v} + \tfrac{3}{4}\boldsymbol{v}\cdot\nabla z. \tag{3.3.1}$$

Recall that by the mass conservation equation, we have $\dot{\mathsf{h}} = -\mathsf{h}\nabla\cdot\boldsymbol{v}$. Also recall the notation: $\dot{\mathsf{k}} := \partial_t(\boldsymbol{v}\cdot\nabla z)+\boldsymbol{v}\cdot\nabla(\boldsymbol{v}\cdot\nabla z)$. Then, applying the total derivative operator D_t to both sides of (3.3.1) yields:

$$\begin{aligned}
\mathsf{D}_t w &= \partial_t w + \boldsymbol{v}\cdot\nabla w \\
&= \frac{1}{2}\ddot{\mathsf{h}} + \frac{3}{4}\dot{\mathsf{k}} \\
&= \frac{3}{2}\Big(\frac{1}{3}\ddot{\mathsf{h}} + \frac{1}{2}\dot{\mathsf{k}}\Big) \\
&= \frac{3}{2}\frac{\overline{p}(\boldsymbol{u})}{\mathsf{h}},
\end{aligned}$$

where $\overline{p}(\boldsymbol{u}) := \frac{1}{3}\mathsf{h}\ddot{\mathsf{h}} + \frac{1}{2}\mathsf{h}\dot{\mathsf{k}}$. We multiply the above equation by h and use Proposition 3.2.1 to obtain a first-order conservation law:

$$\partial_t(\mathsf{h}w) + \nabla\cdot(\boldsymbol{u}\mathsf{h}w) = \tfrac{3}{2}\overline{p}(\boldsymbol{u}).$$

Then, another first-order reformulation of the dispersive Serre model with topography is given as follows:

$$\partial_t\mathsf{h} + \nabla\cdot\boldsymbol{q} = 0, \tag{3.3.2a}$$

$$\partial_t\boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla(\tfrac{1}{2}g\mathsf{h}^2 + \mathsf{h}\overline{p}(\boldsymbol{u})) = -(g\mathsf{h} + \tfrac{3}{2}\overline{p}(\boldsymbol{u}) + \tfrac{1}{4}\mathsf{h}\dot{\mathsf{k}})\nabla z, \tag{3.3.2b}$$

$$\partial_t(\mathsf{h}w) + \nabla\cdot(\boldsymbol{u}\mathsf{h}w) = \tfrac{3}{2}\overline{p}(\boldsymbol{u}), \tag{3.3.2c}$$

$$\nabla\cdot\boldsymbol{v} + \frac{w - \frac{3}{4}\boldsymbol{v}\cdot\nabla z}{\frac{1}{2}\mathsf{h}} = 0. \tag{3.3.2d}$$

We recover Eq. (1) in [19], up to the term $\frac{1}{4}\mathsf{h}\dot{\mathsf{k}}$, which is neglected therein, and up to the coefficient $\frac{3}{4}$ in (3.3.2d). The above system bears some resemblance to (3.2.2). In particular we observe that $\mathsf{h}\overline{p}(\boldsymbol{u}) = -\frac{1}{3}s(\boldsymbol{u})$ and $\mathsf{h}w = \frac{1}{2}q_2$. Notice however that in our system (3.2.2) the two constraints (3.2.2f) are purely algebraic (i.e., these constraints are enforced in the phase space), whereas the constraint (3.3.2d) is differential. As a result, the technique proposed in [19] to relax the differential constraint (3.3.2d) is fundamentally different from (4.4.1h).

Remark 3.3.1 (Reformulation in [21]). In [21] where the full topography effects of the Serre model are considered, two constraints are introduced for the reformulation:

$$w_s := -\mathsf{h}\nabla\cdot\boldsymbol{v} + \boldsymbol{v}\cdot\nabla z, \quad \tilde{w} := w_s + \frac{1}{2}\mathsf{h}\nabla\cdot\boldsymbol{v}$$

Combining these two constraints into one yields:

$$\tilde{w} := -\frac{1}{2}\mathsf{h}\nabla\cdot\boldsymbol{v} + \boldsymbol{v}\cdot\nabla z.$$

This constraint is of the form (3.3.1) up to the constant on the $\boldsymbol{v}\cdot\nabla z$ term. Thus, we can repeat the above process and derive the first-order formulation introduced in [21]. Again, in [21], the constraint is differential and thus different from the proposed reformulation in §3.2.

4. A HYPERBOLIC RELAXATION TECHNIQUE FOR SOLVING THE SERRE EQUATIONS

4.1 Introduction

In this chapter, we introduce a hyperbolic relaxation technique for solving the Serre model (2.3.1) for dispersive water waves with topography effects. The work shown in this chapter addresses one the main challenges for solving the Serre equations. In particular, the hyperbolic relaxation technique circumvents the dispersive time step restriction discussed in §2.3.4 since the technique reformulates the Serre model as a first-order hyperbolic system. This allows for explicit time-stepping when discretizing the equations in time under the usual hyperbolic CFL condition. The hyperbolic reformulation technique sets up the framework for introducing an explicit, continuous finite element approximation that is invariant-domain preserving and well-balanced in Chapter 5.

This chapter is organized as follows. In Section 4.2, we recall the reformulation of the Serre model introduced in Chapter 3 which is the starting point of the relaxation technique. We also introduce some notation preliminaries in this section. Then, in Section 4.3 we describe how to relax the constraints $\{q_1 := \mathsf{h}^2; q_3 := \boldsymbol{q}\cdot\nabla z\}$ discussed in §3.2 and introduce them in the model. This relaxation can be thought of as a penalty technique. Then, in Section 4.3.1, we discuss how to define the relaxed non-hydrostatic pressure so that the new model is an energy-consistent approximation of the original Serre Equations. In Section 4.4, we present a generic form of the relaxed model. Then, in Section 4.5, we derive important properties of the new relaxed model. In particular, we show the relaxed model is indeed hyperbolic. We also describe how to define the relaxation functional Γ in Section 4.5.2 so that the resulting hyperbolic system remains compatible with dry states. We then derive an energy inequality for the model and derive the linear dispersion properties. In Section 4.6, we review the literature on similar relaxation techniques for solving the Serre equations to put the work shown here in context. Finally, we summarize the chapter.

4.2 Preliminaries

For completeness, we recall here the first-order reformulation of the Serre Equations under the constraints $\{q_1 := \mathsf{h}^2; q_3 := \boldsymbol{q}\cdot\nabla z\}$. Let $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T}$ be the solution variable for the following first-order system:

$$\partial_t \mathsf{h} + \nabla\cdot\boldsymbol{q} = 0, \tag{4.2.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla(\tfrac{1}{2}g\mathsf{h}^2 - \tfrac{1}{3}\mathsf{h}s) = -(g\mathsf{h} - \tfrac{1}{2}s + \tfrac{1}{4}\widetilde{s})\nabla z, \tag{4.2.1b}$$

$$\partial_t q_1 + \nabla\cdot(\boldsymbol{v}q_1) = q_2 - \tfrac{3}{2}q_3, \tag{4.2.1c}$$

$$\partial_t q_2 + \nabla\cdot(\boldsymbol{v}q_2) = -s, \tag{4.2.1d}$$

$$\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) = \widetilde{s}, \tag{4.2.1e}$$

$$q_1 = \mathsf{h}^2, \quad q_3 = \boldsymbol{q}\cdot\nabla z. \tag{4.2.1f}$$

To simplify the notation in what follows, we adopt a similar notation as in Favrie and Gavrilyuk [20] and introduce the following primitive variables for the auxiliary quantities: $q_1 := \mathsf{h}\eta$, $q_2 := \mathsf{h}\omega$, $q_3 := \mathsf{h}\beta$. We also introduce the quantity $\overline{\lambda}$ which is a non-dimensional number of order one. Finally, we introduce a relaxation parameter ϵ, a small length scale, which we will later replace by the local mesh-size when the model is approximated in space.

4.3 Relaxing the $\{q_1 = \mathsf{h}^2; q_3 = \boldsymbol{q}\cdot\nabla z\}$ constraints

In this section, we propose a technique to relax the constraints $\{q_1 = \mathsf{h}^2; q_3 = \boldsymbol{q}\cdot\nabla z\}$ so that the resulting system is hyperbolic and remains compatible with dry states. This technique is a loose adaptation of the Lagrangian penalty technique introduced in [20].

Let $\Gamma \in C^2(\mathbb{R}; [0,\infty))$ be some smooth non-negative function with constraints $\Gamma(1) = 0$ and $\Gamma'(0) = 0$. We replace the quantity s in (4.2.1) by

$$s \to -\frac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2\Gamma'\left(\frac{\eta}{\mathsf{h}}\right). \tag{4.3.1}$$

Notice that the quantity $\frac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2$ has dimensions of $\mathrm{m}\frac{\mathrm{m}}{\mathrm{s}^2}$ which is what we expect since $-\frac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2\Gamma'\left(\frac{\eta}{\mathsf{h}}\right)$ should be an ansatz for $-\mathsf{h}(\ddot{\mathsf{h}}+\frac{3}{2}\dot{\mathsf{k}})$. The purpose of this term is to enforce the constraint $q_1=\mathsf{h}^2$ through the Γ function; that is to say, we want to enforce the ratio $\frac{\eta}{\mathsf{h}}$ to be close to 1 as $\epsilon\to 0$ since $\Gamma'(1)=0$. We call (4.3.1) the penalty term for the q_2 equation. We note that the definition of the Γ functional will be discussed in §4.5.2

Let $\Phi\in C^0(\mathbb{R};\mathbb{R})$ be a function such that $\xi\Phi(\xi)\geq 0$ for all $\xi\in\mathbb{R}$. Let h_0 be a reference water depth. We then replace $\widetilde{s}$ in (4.2.1) by:

$$\widetilde{s}\to\frac{\overline{\lambda}g}{\epsilon}\mathsf{h}_0\mathsf{h}\Phi\left(\frac{\boldsymbol{v}\cdot\nabla z-\beta}{\sqrt{g\mathsf{h}_0}}\right). \tag{4.3.2}$$

Notice that this quantity also has dimensions of $\mathrm{m}\frac{\mathrm{m}}{\mathrm{s}^2}$ since it is an ansatz for $\mathsf{h}\dot{\mathsf{k}}$. The purpose of this term is to enforce the constraint $q_3=\boldsymbol{q}\cdot\nabla z$ through the Φ function; that is to say, we want to enforce the ratio $\frac{\boldsymbol{v}\cdot\nabla z-\beta}{\sqrt{g\mathsf{h}_0}}$ to be close to 0 as $\epsilon\to 0$. We call (4.3.2) the penalty term for the q_3 equation. Note that the Φ function is defined in Remark 4.4.1.

Remark 4.3.1 (ϵ vs. $\frac{\overline{\lambda}}{\epsilon}$). Since $\overline{\lambda}$ is a non-dimensional number, it also possible to let the quantity $\frac{\overline{\lambda}}{\epsilon}$ be the relaxation parameter instead of just ϵ. Then, $\epsilon\to 0$ is equivalent to $\overline{\lambda}\to\infty$. However, this condition on $\overline{\lambda}$ is ad-hoc while the condition $\epsilon\to 0$ will be directly correlated to decreasing the local-mesh size when we approximate the model in space.

4.3.1 The correlation between the energy functional and relaxed pressure and source terms

It is only natural to assume that the new relaxed pressure is given by $p(\boldsymbol{u})=\frac{1}{2}g\mathsf{h}^2-\frac{1}{3}\mathsf{h}s\to\frac{1}{2}g\mathsf{h}^2+\frac{1}{3}\frac{\lambda g}{\epsilon}\mathsf{h}^3\Gamma'(\frac{\eta}{\mathsf{h}})$ (actually, this is similar to the augmented Lagrangian approach: see §4.5.2). However, this expression leads to a relaxed model that is not an energy-consistent relaxation of the original Serre Equations. That is to say, the relaxed energy functional with this pressure does not converge to the Serre energy functional as $\epsilon\to 0$. To find the correct expression for the relaxed pressure, we work through the derivation of an energy inequality (see: Proposition 4.5.3). The

derivation shown in §4.5.3 yields a relaxed pressure of the form:

$$p_\epsilon(\boldsymbol{u}) = \frac{1}{2} g \mathsf{h}^2 + \widetilde{p}_\epsilon(\boldsymbol{u}),$$

where

$$\begin{aligned}
\widetilde{p}_\epsilon(\boldsymbol{u}) &= -\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon}\eta^3 \partial_x \left(x^{-2}\Gamma(x)\right)_{|x=\eta \mathsf{h}^{-1}} \\
&= -\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon}\mathsf{h}^2 \left(\eta \Gamma'(\frac{\eta}{\mathsf{h}}) - 2\mathsf{h}\Gamma(\frac{\eta}{\mathsf{h}})\right) \\
&= -\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon}\mathsf{h}^2 \eta \Gamma'(\frac{\eta}{\mathsf{h}}) + \frac{1}{3}\frac{\overline{\lambda} g}{\epsilon} 2\mathsf{h}^3 \Gamma(\frac{\eta}{\mathsf{h}}).
\end{aligned}$$

By working through the energy argument, we will see that we must also replace $-\frac{3}{2}q_3$ on the right-hand side of (4.2.1c) by $-\frac{3}{2}\boldsymbol{q}\cdot\nabla z$. This is consistent since q_3 should be equal to $\boldsymbol{q}\cdot\nabla z$.

Remark 4.3.2 (Relaxed non-hydrostatic pressure). After replacing s by $\frac{\overline{\lambda} g}{\epsilon}\mathsf{h}^2\Gamma'(\frac{\eta}{\mathsf{h}})$ in (3.2.2b), we observe that the definition of $\widetilde{p}_\epsilon(\boldsymbol{u})$ is compatible with the definition $\widetilde{p}(\boldsymbol{u}) = -\frac{1}{3}\mathsf{h}s$ up to the remainder $\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon} 2\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})$. However, this remainder is small when the ratio $\frac{\eta}{\mathsf{h}}$ is close to 1. More precisely, using Taylor expansions at 1, we have

$$\Gamma(1) = 0 = \Gamma(\frac{\eta}{\mathsf{h}}) + \mathsf{h}^{-1}(\mathsf{h}-\eta)\Gamma'(\frac{\eta}{\mathsf{h}}) + \mathsf{h}^{-2}\mathcal{O}(\mathsf{h}-\eta)^2,$$

which shows that $2\mathsf{h}\Gamma(\frac{\eta}{\mathsf{h}})/\eta|\Gamma'(\frac{\eta}{\mathsf{h}})| = \mathcal{O}(\frac{|\eta-\mathsf{h}|}{\eta})$. Hence the ratio $2\mathsf{h}\Gamma(\frac{\eta}{\mathsf{h}})/\eta|\Gamma'(\frac{\eta}{\mathsf{h}})|$ is small as $\eta \to \mathsf{h}$, which proves that $\widetilde{p}_\epsilon(\boldsymbol{u})$ is indeed a consistent approximation of $\widetilde{p}(\boldsymbol{u}) = -\frac{1}{3}\mathsf{h}s$ as $\eta \to \mathsf{h}$.

4.4 The generic relaxed model

Finally, the relaxed system for the conserved variables $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^{\mathsf{T}}$is formulated as follows:

$$\partial_t \mathsf{h} + \nabla\cdot\boldsymbol{q} = 0, \tag{4.4.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla p_\epsilon(\boldsymbol{u}) = -r_\epsilon(\boldsymbol{u})\nabla z, \tag{4.4.1b}$$

$$\partial_t q_1 + \nabla\cdot(\boldsymbol{v}q_1) = q_2 - \tfrac{3}{2}\boldsymbol{q}\cdot\nabla z, \tag{4.4.1c}$$

$$\partial_t q_2 + \nabla\cdot(\boldsymbol{v}q_2) = -s_\epsilon(\boldsymbol{u}), \tag{4.4.1d}$$

$$\partial_t q_3 + \nabla\cdot(\boldsymbol{v}q_3) = \widetilde{s}_\epsilon(\boldsymbol{u}), \tag{4.4.1e}$$

$$p_\epsilon(\boldsymbol{u}) := \tfrac{1}{2}g\mathsf{h}^2 + \widetilde{p}_\epsilon(\boldsymbol{u}), \qquad \widetilde{p}_\epsilon(\boldsymbol{u}) := -\tfrac{1}{3}\tfrac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2\big(\eta\Gamma'(\tfrac{\eta}{\mathsf{h}}) - 2\mathsf{h}\Gamma(\tfrac{\eta}{\mathsf{h}})\big), \tag{4.4.1f}$$

$$r_\epsilon(\boldsymbol{u}) := g\mathsf{h} - \tfrac{1}{2}s_\epsilon(\boldsymbol{u}) + \tfrac{1}{4}\widetilde{s}_\epsilon(\boldsymbol{u}), \tag{4.4.1g}$$

$$s_\epsilon(\boldsymbol{u}) := \tfrac{\overline{\lambda}g}{\epsilon}\mathsf{h}^2\Gamma'(\tfrac{\eta}{\mathsf{h}}), \qquad \widetilde{s}_\epsilon(\boldsymbol{u}) := \tfrac{\overline{\lambda}g}{\epsilon}\mathsf{h}_0\mathsf{h}\Phi(\tfrac{\boldsymbol{v}\nabla z-\beta}{\sqrt{g\mathsf{h}_0}}). \tag{4.4.1h}$$

The relaxed model (4.4.1) is incomplete since we have yet to explicitly define the Γ and Φ functions. It is shown in Proposition 4.5.2 that the hyperbolicity of the relaxed model is dependent on the form of Γ so we defer defining it here.

Remark 4.4.1 (Defining the Φ function). The condition $\xi\Phi(\xi) \geq 0$ for all $\xi \in \mathbb{R}$ is important for deriving the energy inequality in Prop. 4.5.3. In particular, it is used to show that the energy equation is negative. This is the only information needed to construct Φ. Thus, the simplest choice is to set $\Phi(\xi) := \xi$ so that $\xi\Phi(\xi) = \xi^2 \geq 0$. This choice of Φ gives:

$$\widetilde{s}_\epsilon(\boldsymbol{u}) = \frac{\overline{\lambda}}{\epsilon}\sqrt{g\mathsf{h}_0}(\boldsymbol{q}\cdot\nabla z - q_3).$$

Remark 4.4.2 (Generality). Unless otherwise stated, we express the analysis in this work with generic functions Γ and Φ to emphasize the generality of the relaxation procedure.

4.5 Properties

In this section, we derive important results for the relaxed model (4.4.1).

4.5.1 Hyperbolicity for the relaxed model

We now verify that the relaxed system (4.4.1) is hyperbolic as defined by Definition (1.2.1). Let $\boldsymbol{n}$ be a unit vector in $\mathbb{R}^d$. Let $\mathbb{f}(\boldsymbol{u})$ denote the conservative flux of (4.4.1). We introduce the notation $\boldsymbol{q}{\cdot}\boldsymbol{n} := \mathsf{h}v$ and $\boldsymbol{q}^\perp := \boldsymbol{q} - (\boldsymbol{q}{\cdot}\boldsymbol{n})\boldsymbol{q}$, $\boldsymbol{q}^\perp := \mathsf{h}\boldsymbol{v}^\perp$ and make a change of coordinates so that the conserved variable can be re-written as $\boldsymbol{u} = (\mathsf{h}, (\boldsymbol{q}{\cdot}\boldsymbol{n}), \boldsymbol{q}^\perp, q_1, q_2, q_3)^\mathsf{T}$. Recalling the notation $q_1 := \mathsf{h}\eta$, we introduce the function $\overline{p}(\mathsf{h}, q_1)$ such that $\overline{p}(\mathsf{h}, q_1) := \widetilde{p}_\epsilon(\mathsf{h}, \eta)$. The quasi-linear form of the equation in the new change of coordinates is given by

$$\partial_t \boldsymbol{u} + D(\mathbb{f}(\boldsymbol{u})\boldsymbol{n})\partial_x \boldsymbol{u} = \boldsymbol{S},$$

where $D(\mathbb{f}(\boldsymbol{u})\boldsymbol{n})$ is the Jacobian matrix of the flux $\mathbb{f}(\boldsymbol{u})\boldsymbol{n}$ defined as follows:

$$D(\mathbb{f}(\boldsymbol{u})\boldsymbol{n}) = \begin{pmatrix} 0 & 1 & \mathbf{0}^\mathsf{T} & 0 & 0 & 0 \\ gh - v^2 + \partial_\mathsf{h}\overline{p} & 2v & \mathbf{0}^\mathsf{T} & \partial_{q_1}\overline{p} & 0 & 0 \\ -v\boldsymbol{v}^\perp & \boldsymbol{v}^\perp & v & 0 & 0 & 0 \\ -v\eta & \eta & \mathbf{0}^\mathsf{T} & v & 0 & 0 \\ -vw & w & 0 & 0 & v & 0 \\ -v\beta & \beta & 0 & 0 & 0 & v \end{pmatrix}.$$

Denoting by μ a generic eigenvalue, the characteristic polynomial of the Jacobian $D(\mathbb{f}(\boldsymbol{u})\boldsymbol{n})$ is

$$(\mu - v)^{2+d}((\mu - v)^2 - (g\mathsf{h} + \eta\partial_{q_1}\overline{p} + \partial_\mathsf{h}\overline{p})).$$

This characteristic polynomial yields eigenvalues: v, with multiplicity $d{+}2$; $v{-}\sqrt{g\mathsf{h} + \eta\partial_{q_1}\overline{p} + \partial_\mathsf{h}\overline{p}}$ and $v + \sqrt{g\mathsf{h} + \eta\partial_{q_1}\overline{p} + \partial_\mathsf{h}\overline{p}}$, each with multiplicity 1.

To verify hyperbolicity, we need to investigate the sign of $g\mathsf{h} + \eta\partial_{q_1}\overline{p} + \partial_\mathsf{h}$. In particular, we

would like this quantity to be positive. Using that $\widetilde{p}(\mathsf{h},\eta) := \overline{p}(\mathsf{h},q_1)$, we observe by the chain rule that $\partial_{\mathsf{h}}\widetilde{p} = \eta\partial_{q_1}\overline{p} + \partial_{\mathsf{h}}\overline{p}$. Let us introduce the following change of variable: $\mathsf{h}(\tau) = \tau^{-1}$, i.e., $\tau(\mathsf{h}) = \mathsf{h}^{-1}$. Then, $\widehat{p}(\tau,\eta) := \widetilde{p}(\tau^{-1},\eta)$, and recalling (4.4.1f), this gives $\widehat{p}(\tau,\eta) = -\frac{1}{3}\frac{\overline{\lambda}g}{\epsilon}\partial_\tau(\tau^{-2}\Gamma(\eta\tau))$. Using $\partial_{\mathsf{h}}\widetilde{p}(\mathsf{h},\eta) = -\tau^2\partial_\tau\widehat{p}(\tau,\eta)$, this in turn implies that the following inequality

$$\begin{aligned} g\mathsf{h} + \partial_{\mathsf{h}}\widetilde{p}(\mathsf{h},\eta) &= g\mathsf{h} + \frac{1}{3}\frac{\overline{\lambda}g}{\epsilon}\tau^2\partial_{\tau\tau}(\tau^{-2}\Gamma(\eta\tau)) \\ &= g\mathsf{h}(1 + \frac{\overline{\lambda}}{3\epsilon}\eta\times(x^3\partial_{xx}(x^{-2}\Gamma(x)))_{|x=\eta\mathsf{h}^{-1}}), \end{aligned}$$

Thus

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

is a necessary and sufficient condition for hyperbolicity. The results are summarized in the following proposition.

Proposition 4.5.1 (Hyperbolicity)**.** *Let $\mathbb{f}(\boldsymbol{u})$ be the conservative flux of the system* (4.4.1)*. For any unit vector $\boldsymbol{n} \in \mathbb{R}^d$, the $d+4$ eigenvalues of the Jacobian matrix of the flux $\mathbb{f}(\boldsymbol{u})\boldsymbol{n}$ are $\mu_k = \boldsymbol{v}\cdot\boldsymbol{n}$, $k \in \{2{:}d+3\}$ and*

$$\mu_1 = \boldsymbol{v}\cdot\boldsymbol{n} - \sqrt{g\mathsf{h} + \partial_{\mathsf{h}}\widetilde{p}(\mathsf{h},\eta)}, \quad \mu_{d+4} = \boldsymbol{v}\cdot\boldsymbol{n} + \sqrt{g\mathsf{h} + \partial_{\mathsf{h}}\widetilde{p}(\mathsf{h},\eta)}. \tag{4.5.1}$$

The system (4.4.1) *is hyperbolic iff the following holds for all $\eta \in \mathbb{R}$ and all $\mathsf{h} \geq 0$:*

$$g\mathsf{h}\left(1 + \frac{1}{3}\frac{\overline{\lambda}}{\epsilon}\eta\left(x^3\partial_{xx}(x^{-2}\Gamma(x))\right)_{|x=\eta\mathsf{h}^{-1}}\right) \geq 0. \tag{4.5.2}$$

4.5.2 Defining the Γ function

In this section, we construct the smooth functional Γ that: (i) satisfies the constraints $\Gamma \in C^2(\mathbb{R};[0,\infty))$, $\Gamma \geq 0$, $\Gamma(1) = \Gamma'(1) = 0$; (ii) satisfies the hyperbolicity condition (4.5.2); (iii) yields a relaxed model that is compatible with dry states.

The hyperbolicity condition (4.5.2) suggests that the explicit form for Γ should satisfy:

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

Note that a priori we don't know that η is non-negative (it is assumed, but not proven). Thus, it is important to construct $\Gamma(\eta\mathsf{h}^{-1})$ for all $\eta \in \mathbb{R}$. Since $\mathsf{h} > 0$, we set $\eta = \mathsf{h}\frac{\eta}{\mathsf{h}} = \mathsf{h}x$ where $x = \frac{\eta}{\mathsf{h}}$ and rewrite the above inequality:

$$\frac{1}{3}\frac{\overline{\lambda}}{\epsilon}\mathsf{h}\left(x^4\partial_{xx}(x^{-2}\Gamma(x))\right)_{|x=\eta\mathsf{h}^{-1}} \geq 0.$$

The quantity $\frac{1}{3}\frac{\overline{\lambda}}{\epsilon}\mathsf{h}$ is positive, so finding Γ is reduced to solving the following ordinary differential equation:

$$\begin{cases} 6\Gamma - 4x\Gamma' + x^2\Gamma'' = \alpha, \\ \Gamma(1) = 0, \quad \Gamma'(1) = 0, \end{cases}$$

where α is a positive constant. Solving the ordinary differential equation yields $\Gamma(x) = \frac{\alpha}{6}(1 + 2x)(x-1)^2$. Note that the condition $\alpha > 0$ is important since $\alpha = 0$ would yield $\Gamma(x) = 0$. For simplicity, we take $\alpha = 6$. This yields a relaxed pressure:

$$\begin{aligned} \widetilde{p}_\epsilon(\mathsf{h},\eta) &= \tfrac{\overline{\lambda}g}{\epsilon}\tfrac{2}{3}(\mathsf{h}^3 - \eta^3), \\ &\Longrightarrow \\ \overline{p}(\mathsf{h},q_1) &= \tfrac{\overline{\lambda}g}{\epsilon}\tfrac{2}{3}\left(\mathsf{h}^3 - \tfrac{q_1^3}{\mathsf{h}^3}\right), \end{aligned}$$

and relaxed source:

$$\begin{aligned} s_\epsilon(\mathsf{h},\eta) &= \tfrac{\overline{\lambda}g}{\epsilon}2\eta(\eta - \mathsf{h}), \\ &\Longrightarrow \\ \overline{s}(\mathsf{h},q_1) &= \tfrac{\overline{\lambda}g}{\epsilon}2\left(\tfrac{q_1^2}{\mathsf{h}^2} - q_1\right), \end{aligned}$$

The above relaxed pressure yields a relaxed model that is uniformly hyperbolic. However, it has been observed numerically that this model does not behave properly with dry states ($\mathsf{h} \to 0$) when $\eta \leq \mathsf{h}$. This can be seen as follows. Since $\lim_{\mathsf{h}->0} \frac{1}{2} g\mathsf{h}^2 = 0$, we consider the ratio of the non-hydrostatic relaxed pressure to the hydrostatic pressure: $\widetilde{p}_\epsilon(\mathsf{h}, \eta)/(\frac{1}{2} g\mathsf{h}^2)$:

$$\frac{\widetilde{p}_\epsilon(\mathsf{h}, \eta)}{(\frac{1}{2} g\mathsf{h}^2)} = \frac{\overline{\lambda}}{\epsilon} \frac{4}{3} \frac{\mathsf{h}^3 - \eta^3}{\mathsf{h}^2}$$

We see that as $\mathsf{h} \to 0$, this quantity is unbounded. Thus, this particular Γ functional gives a relaxed model that is uniformly hyperbolic, but not compatible with dry states.

To find a suitable workaround, we adapt the augmented Lagrangian approach of Favrie and Gavrilyuk [20]. That is to say, we set the relaxed pressure to $\widetilde{p}(\mathsf{h}, \eta) = -\frac{1}{3}\mathsf{h}s = -\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon}\mathsf{h}\mathsf{h}^2\Gamma'(\frac{\eta}{\mathsf{h}})$ where $s(\mathsf{h}, \eta) = \frac{\overline{\lambda} g}{\epsilon}\mathsf{h}^2\Gamma'(\frac{\eta}{\mathsf{h}})$. However, recall that by the energy argument in Proposition 4.5.3, this new relaxed non-hydrostatic pressure also has to satisfy: $\widetilde{p}(\mathsf{h}, \eta) = -\frac{1}{3}\frac{\overline{\lambda} g}{\epsilon}\eta^3 \partial_x (x^{-2}\Gamma(x))_{|x=\eta\mathsf{h}^{-1}}$. This is only possible if $\mathsf{h}^3\Gamma'(\frac{\eta}{\mathsf{h}}) = \eta^3 \partial_x (x^{-2}\Gamma(x))_{|x=\eta\mathsf{h}^{-1}}$. This condition yields the following ordinary differential equation for Γ:

$$\begin{cases} \Gamma'(x) = -2\Gamma(x) + x\Gamma'(x), \\ \Gamma(1) = 0. \end{cases}$$

The solution to this ODE is given by $\Gamma(x) = c(x-1)^2$ where c is an arbitrary, positive constant. For simplicity, we take $c = 3$. Substituting this definition of Γ into the hyperbolicity condition (4.5.2) yields:

$$g\mathsf{h} + 2\frac{\overline{\lambda} g}{\epsilon}(3\mathsf{h} - 2\eta) \geq 0.$$

We see that hyperbolicity may be lost when $2\eta > 3\mathsf{h}$ i.e., $x > \frac{3}{2}$.

The solution we propose moving forward is to combine both functionals in a piecewise manner.

That is, we define Γ as follows:

$$\Gamma(x) = \begin{cases} 3(1-x)^2 & \text{if } x \leq 1, \\ (1+2x)(1-x)^2 & \text{if } 1 \leq x. \end{cases} \tag{4.5.3}$$

Notice that $\Gamma(x) \geq 0$ for any $x \in \mathbb{R}$, $\Gamma(1) = 0$, $\Gamma'(1) = 0$, and $\Gamma \in C^2(\mathbb{R}; [0, \infty))$. A plot of this function can be seen in Figure 4.1. The final source term $\mathsf{h}^2\Gamma'(\eta\mathsf{h}^{-1})$ in the balance equation for

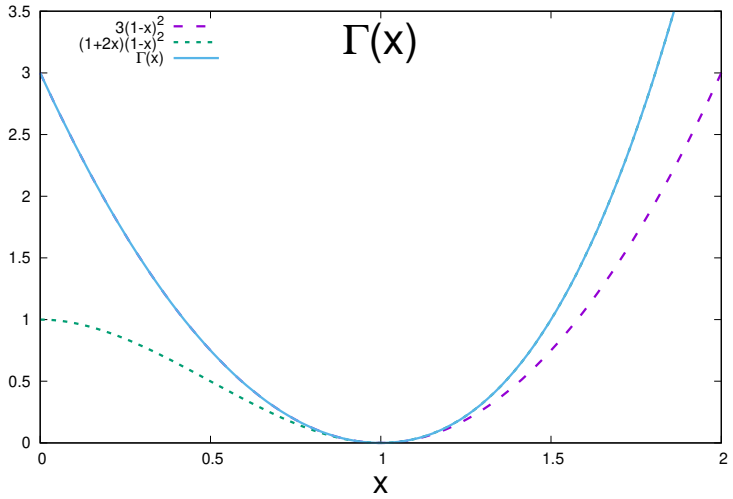


Figure 4.1: A plot of $\Gamma(x)$.

$\mathsf{h}\omega$ and pressure $\widetilde{p}(\mathsf{h}, \eta)$ are given by

$$\mathsf{h}^2\Gamma'(\eta\mathsf{h}^{-1}) = \begin{cases} 6(\eta\mathsf{h} - \mathsf{h}^2), & \text{if } \eta \leq \mathsf{h} \\ 6(\eta^2 - \eta\mathsf{h}), & \text{if } \mathsf{h} \leq \eta, \end{cases} \tag{4.5.4}$$

$$\widetilde{p}(\mathsf{h}, \eta) = -\frac{\overline{\lambda} g}{3\epsilon} \times \begin{cases} 6\mathsf{h}(\eta\mathsf{h} - \mathsf{h}^2), & \text{If } \eta \leq \mathsf{h} \\ 2(\eta^3 - \mathsf{h}^3), & \text{If } \mathsf{h} \leq \eta. \end{cases} \tag{4.5.5}$$

The above definition of Γ implies that

$$\partial_{\mathsf{h}}\widetilde{p}(\mathsf{h},\eta) = g\mathsf{h}\frac{\overline{\lambda}}{3\epsilon}\times\begin{cases}6\mathsf{h}+12(\mathsf{h}-\eta), & \text{If } \eta\leq\mathsf{h},\\ 6\mathsf{h}, & \text{If } \mathsf{h}\leq\eta.\end{cases} \tag{4.5.6}$$

Hence the hyperbolicity condition (4.5.2) is satisfied for any pair $(\eta,\mathsf{h})\in\mathbb{R}\times\mathbb{R}_+$ and the model is compatible with dry states.

Remark 4.5.2 (Infinitely many models). The Γ and Φ functionals chosen above are seemingly the most efficient, but they are not unique. Technically, there exists a family of infinitely many relaxed models assuming Γ and Φ satisfy their respective constraints.

4.5.3 Derivation of energy inequality

In this section, we derive an energy inequality for the hyperbolic relaxed model (4.4.1). The key results of this section are (i) Proposition 4.5.3 which a relaxed counterpart to Proposition 2.3.6; (ii) Corollary 4.5.5 which shows that the extra energy induced by the relaxation is small and positive.

Let $\boldsymbol{u} := (\mathsf{h},\boldsymbol{q},q_1,q_2,q_3)^{\mathsf{T}}$ be a smooth solution to (4.4.1a)–(4.4.1h) and assume ϵ is constant. To simplify the presentation, we use the primitive representation of the auxiliary variables $q_1 := \mathsf{h}\eta, q_2 := \mathsf{h}\omega, q_3 := \mathsf{h}\beta$. We first derive an equation for the potential energy. Recall this is done by multiplying the mass conservation equation (4.4.1a) with $g(\mathsf{h}+z)$. Since the mass equation of the relaxed model is equivalent to that of the Serre model (2.3.1), the potential energy equation is the same as (2.3.14):

$$\partial_t\big(\frac{1}{2}g\mathsf{h}^2+g\mathsf{h}z\big)+\nabla\cdot\big(\boldsymbol{v}\frac{1}{2}g\mathsf{h}^2\big)+\frac{1}{2}g\mathsf{h}^2(\nabla\cdot\boldsymbol{v})+gz\nabla\cdot(\mathsf{h}\boldsymbol{v})=0. \tag{4.5.7}$$

To derive the kinetic energy equation, we apply the operator $(\boldsymbol{v}\cdot)$ to (4.4.1b):

$$\partial_t(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2)+\nabla\cdot(\boldsymbol{v}(\frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2))+\boldsymbol{v}\cdot\nabla p_\epsilon(\boldsymbol{u})=\frac{1}{2}s_\epsilon(\boldsymbol{u})\boldsymbol{v}\cdot\nabla z-\tfrac{1}{4}\widetilde{s}_\epsilon(\boldsymbol{u})\boldsymbol{v}\cdot\nabla z.$$

We now find the contributions to the total energy induced by the auxiliary variables. We first

multiply (4.4.1c) by $\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}\Gamma'(\frac{\eta}{\mathsf{h}}) := \frac{1}{3\mathsf{h}}s_\epsilon(\boldsymbol{u})$:

$$\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}\Gamma'(\frac{\eta}{\mathsf{h}})\left(\partial_t(\mathsf{h}\eta)+\nabla\cdot(\mathsf{h}\eta\boldsymbol{v})\right)=\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^2\Gamma'(\frac{\eta}{\mathsf{h}})\left(\partial_t\eta+\boldsymbol{v}\cdot\nabla\eta\right).$$

We want to simplify the expression on the right hand side. By the chain rule, we have

$$\partial_t\Gamma(\tfrac{\eta}{\mathsf{h}})=\Gamma'(\tfrac{\eta}{\mathsf{h}})(\tfrac{1}{\mathsf{h}}\partial_t\eta-\tfrac{\eta}{\mathsf{h}^2}\partial_t\mathsf{h}),\qquad \boldsymbol{v}\cdot\nabla\Gamma(\tfrac{\eta}{\mathsf{h}})=\Gamma'(\tfrac{\eta}{\mathsf{h}})(\tfrac{1}{\mathsf{h}}\boldsymbol{v}\cdot\nabla\eta-\tfrac{\eta}{\mathsf{h}^2}\boldsymbol{v}\cdot\nabla\mathsf{h}).$$

Using the above identities, we see

$$\begin{aligned}
\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^2\Gamma'(\frac{\eta}{\mathsf{h}})(\partial_t\eta+\boldsymbol{v}\cdot\nabla z) &= \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\left(\partial_t\Gamma(\tfrac{\eta}{\mathsf{h}})+\boldsymbol{v}\cdot\nabla\Gamma(\tfrac{\eta}{\mathsf{h}})\right)+\frac{\overline{\lambda}g}{3\epsilon}\Gamma'(\frac{\eta}{\mathsf{h}})\mathsf{h}^2\frac{\eta}{\mathsf{h}}(\partial_t\mathsf{h}+\boldsymbol{v}\cdot\nabla\mathsf{h})\\
&= \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\left(\partial_t\Gamma(\tfrac{\eta}{\mathsf{h}})+\boldsymbol{v}\cdot\nabla\Gamma(\tfrac{\eta}{\mathsf{h}})\right)+\frac{\overline{\lambda}g}{3\epsilon}\Gamma'(\frac{\eta}{\mathsf{h}})\eta\mathsf{h}(-\mathsf{h}\nabla\cdot\boldsymbol{v})\\
&= \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\mathsf{D}_t\left(\Gamma(\tfrac{\eta}{\mathsf{h}})\right)-\frac{\overline{\lambda}g}{3\epsilon}\Gamma'(\frac{\eta}{\mathsf{h}})\eta\mathsf{h}^2\nabla\cdot\boldsymbol{v}\\
&= \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^2\left(\mathsf{D}_t(\mathsf{h}\Gamma(\tfrac{\eta}{\mathsf{h}}))+\mathsf{h}\Gamma(\frac{\eta}{\mathsf{h}})\nabla\cdot\boldsymbol{v})\right)-\frac{\overline{\lambda}g}{3\epsilon}\Gamma'(\frac{\eta}{\mathsf{h}})\eta\mathsf{h}^2\nabla\cdot\boldsymbol{v}\\
&= \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}\left(\mathsf{D}_t(\mathsf{h}^2\Gamma(\tfrac{\eta}{\mathsf{h}}))+2\mathsf{h}^2\Gamma(\frac{\eta}{\mathsf{h}})\nabla\cdot\boldsymbol{v})\right)-\frac{\overline{\lambda}g}{3\epsilon}\Gamma'(\frac{\eta}{\mathsf{h}})\eta\mathsf{h}^2\nabla\cdot\boldsymbol{v}\\
&= \frac{\overline{\lambda}g}{3\epsilon}\left(\partial_t(\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})+\nabla\cdot(\boldsymbol{v}\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})\right)+\underbrace{\frac{\overline{\lambda}g}{3\epsilon}\left(2\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})-\mathsf{h}^2\eta\Gamma(\frac{\eta}{\mathsf{h}})\right)}_{\widetilde{p}_\epsilon(\boldsymbol{u})}\nabla\cdot\boldsymbol{v}\\
&= \frac{\overline{\lambda}g}{3\epsilon}\left(\partial_t(\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})+\nabla\cdot(\boldsymbol{v}\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})\right)+\widetilde{p}_\epsilon(\boldsymbol{u})\nabla\cdot\boldsymbol{v}.
\end{aligned}$$

Notice that as mentioned in §4.3.1 the definition of the relaxed non-hydrostatic pressure $\widetilde{p}_\epsilon$ is important for the above simplification.

Multiplying the source terms of (4.4.1c) yields:

$$\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}\Gamma'(\frac{\eta}{\mathsf{h}})(\mathsf{h}\omega-\frac{3}{2}\boldsymbol{q}\cdot\nabla z) = \frac{1}{3}s_\epsilon(\boldsymbol{u})\omega-\frac{1}{2}s_\epsilon(\boldsymbol{u})\boldsymbol{v}\cdot\nabla z.$$

Notice again, as mentioned in §4.3.1, replacing $-\frac{3}{2}q_3$ by $-\frac{3}{2}\mathsf{h}\boldsymbol{v}\cdot\nabla z$ in (4.4.1c) is important here. Without this substitution we would have $-\frac{1}{2}s_\epsilon(\boldsymbol{u})\beta$ on the right-hand side of the above identity

instead of $-\frac{1}{2}s_\epsilon(\boldsymbol{u})\boldsymbol{v}\cdot\nabla z$. The above terms combine to the following equation:

$$\partial_t(\tfrac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}}))+\nabla\cdot(\tfrac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}})\boldsymbol{v})+\widetilde{p}_\epsilon(\mathsf{h},\eta)\nabla\cdot\boldsymbol{v}=\tfrac{1}{3}s_\epsilon(\boldsymbol{u})\omega-\tfrac{1}{2}s_\epsilon(\boldsymbol{u})\boldsymbol{v}\cdot\nabla z.$$

We continue by multiplying (4.4.1d) by $\frac{1}{3}\omega$ and we obtain

$$\partial_t(\tfrac{1}{6}\mathsf{h}\omega^2)+\nabla\cdot(\tfrac{1}{6}\mathsf{h}\omega^2\boldsymbol{v})=-\tfrac{1}{3}s_\epsilon(\boldsymbol{u})\omega.$$

Finally we multiply (4.4.1e) by $\frac{1}{4}\beta$ and obtain

$$\partial_t(\tfrac{1}{8}\mathsf{h}\beta^2)+\nabla\cdot(\tfrac{1}{8}\mathsf{h}\beta^2\boldsymbol{v})=\tfrac{1}{4}\widetilde{s}_\epsilon(\boldsymbol{u})\beta.$$

Notice that the definition of $\widetilde{s}_\epsilon(\boldsymbol{u})$ and $\Phi(\xi)$ gives

$$\tfrac{1}{4}\widetilde{s}_\epsilon(\boldsymbol{u})(\beta-\boldsymbol{v}\cdot\nabla z)=-\tfrac{g\mathsf{h}_0}{4}\tfrac{\mathsf{h}}{\epsilon}\Phi(\tfrac{\boldsymbol{v}\nabla z-\beta}{\sqrt{g\mathsf{h}_0}})(\boldsymbol{v}\cdot\nabla z-\beta)\leq 0.$$

Combining the above details, we have the following energy inequality.

Proposition 4.5.3 (Energy inequality for relaxed system)**.** *Let* $\boldsymbol{u}$ *be a smooth solution to* (4.4.1a)–(4.4.1h). *Then the following holds true:* $\partial_t\mathcal{E}_\epsilon(\boldsymbol{u})+\nabla\cdot(\boldsymbol{\mathcal{F}}_\epsilon(\boldsymbol{u}))=\frac{1}{4}\widetilde{s}_\epsilon(\boldsymbol{u})(\beta-\boldsymbol{v}\cdot\nabla z)\leq 0$*, with*

$$\mathcal{E}_\epsilon(\boldsymbol{u}):=\tfrac{1}{2}g\mathsf{h}^2+gz\mathsf{h}+\tfrac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2+\tfrac{1}{6}\mathsf{h}\omega^2+\tfrac{1}{8}\mathsf{h}\beta^2+\frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}}),\tag{4.5.8a}$$

$$\boldsymbol{\mathcal{F}}_\epsilon(\boldsymbol{u}):=\boldsymbol{v}(\mathcal{E}_\epsilon(\boldsymbol{u})+p_\epsilon(\boldsymbol{u})).\tag{4.5.8b}$$

Remark 4.5.4 ((4.5.8a) vs. (2.3.15a)). By comparing the expression (4.5.8a) to (2.3.15a), and recalling that ω is meant to be an approximation for $\dot{\mathsf{h}}+\frac{3}{2}\boldsymbol{v}\cdot\nabla z$ and β an approximation for $\boldsymbol{v}\cdot\nabla z$, we see that $\frac{1}{6}\mathsf{h}\omega^2+\frac{1}{8}\mathsf{h}\beta^2=\frac{1}{6}\mathsf{h}(\omega^2+\frac{3}{4}\beta^2)$ in (4.5.8a) is the approximation of $\frac{1}{6}\mathsf{h}\Big(\big(\dot{\mathsf{h}}+\frac{3}{2}(\boldsymbol{v}\cdot\nabla z)\big)^2+\frac{3}{4}(\boldsymbol{v}\cdot\nabla z)^2\Big)$ in (2.3.15a). Notice that we observe an extra term in the energy $\mathsf{h}^3\Gamma(\frac{\eta}{\mathsf{h}})$, but this was

shown in Remark 4.3.2 to be a small, positive quantity.

Corollary 4.5.5 (Bound on relaxation induced energy term)**.** *Let* $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T}$ *be a smooth solution to* (4.4.1a)–(4.4.1h) *and assume* ϵ *is constant (i.e., does not depend on* $\boldsymbol{x}$ *and* t*). Let* $T > 0$ *be some final time. Assume that the boundary conditions for* $\boldsymbol{u}$ *are such that* $\boldsymbol{\mathcal{F}}(\boldsymbol{u})\cdot\boldsymbol{n}_{\partial D} = 0$ *for all* $t \in (0, T)$*. Also assume that the frame of reference for the model problem is such that the topography is positive (i.e.,* $z(\boldsymbol{x}) > 0$*). Then, there is a* $c(\boldsymbol{u}_0)$ *such that:*

$$\int_D \left(\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}})\right)_{|t=T} \leq c(\boldsymbol{u}_0)\epsilon. \tag{4.5.9}$$

Proof. Integrating $\partial_t\mathcal{E}_\epsilon(\boldsymbol{u}) + \nabla\cdot\boldsymbol{\mathcal{F}}_\epsilon(\boldsymbol{u}) \leq 0$ over the spatial domain and applying the boundary condition $\boldsymbol{\mathcal{F}}(\boldsymbol{u})\cdot\boldsymbol{n}_{\partial D} = 0$ gives:

$$\int_D \partial_t\mathcal{E}_\epsilon(\boldsymbol{u}) \leq 0.$$

Then, integrating again over the time interval $(0, T)$ yields:

$$\begin{aligned} \int_D (\mathcal{E}_\epsilon(\boldsymbol{u}(T)) - \mathcal{E}_\epsilon(\boldsymbol{u}(0))) \quad &\leq \quad 0, \\ &\Longrightarrow \\ \int_D \mathcal{E}_\epsilon(\boldsymbol{u}(T)) \quad &\leq \quad c(\boldsymbol{u}_0). \end{aligned}$$

Then, by dropping the positive quantities $(\frac{1}{2}g\mathsf{h}^2 + gz\mathsf{h} + \frac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2 + \frac{1}{6}\mathsf{h}\omega^2 + \frac{1}{8}\mathsf{h}\beta^2)_{t=T}$, we have

$$\int_D \frac{\overline{\lambda}g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}}) \leq c(\boldsymbol{u}_0).$$

The result follow immediately. □

Remark 4.5.6 (Energy consistent approximation). The previous result shows that the extra energy term induced by the relaxation vanishes as $\epsilon \to 0$. That is to say, the energy functional (4.5.8a) converges to the Serre energy functional (2.3.15a). Thus, our hyperbolic relaxed model is a consistent approximation of the original Serre equations. In Figure 4.2 we show an updated version of

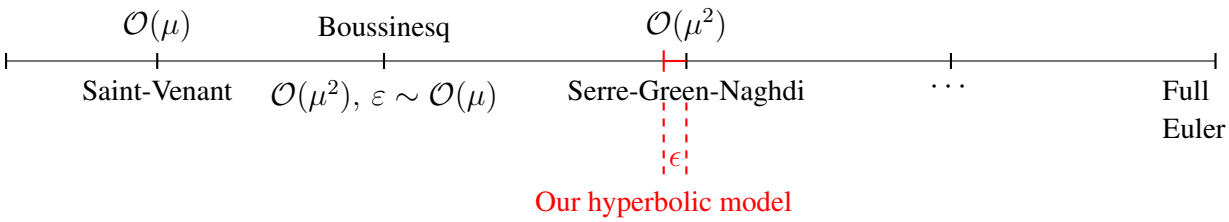


Figure 4.2: A representation of the physical accuracy of the models derived in Section 1.4 compared to the full free-surface Euler problem now including the hyperbolic relaxed model.

Figure 1.2 that now includes the "location" of the hyperbolic relaxed model.

Corollary 4.5.7 (Energy equality for relaxed system with flat topography)**.** *If the topography is flat ($z(\boldsymbol{x}) \equiv 0$), the following holds true: $\partial_t \mathcal{E}_{\text{flat}} + \nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}) = 0$, with*

$$\mathcal{E}_{\text{flat}}(\boldsymbol{u}) := \tfrac{1}{2} g \mathsf{h}^2 + \tfrac{1}{2}\mathsf{h}\|\boldsymbol{v}\|^2 + \tfrac{1}{6}\mathsf{h}\omega^2 + \tfrac{\overline{\lambda} g}{3\epsilon}\mathsf{h}^3\Gamma(\tfrac{\eta}{\mathsf{h}}), \tag{4.5.10a}$$

$$\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u}) := \boldsymbol{v}(\mathcal{E}_{\text{flat}}(\boldsymbol{u}) + p(\boldsymbol{u})). \tag{4.5.10b}$$

The following result is essential for constructing higher-order accurate (in space) graph viscosity coefficients for the numerical approximation in Chapter 5.

Proposition 4.5.8 (Entropy chain rule)**.** *The energy functional and energy flux defined in* (4.5.10) *for the relaxed Serre model with flat topography satisfy the following property:*

$$\nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u})) = (\nabla_{\boldsymbol{u}}\mathcal{E}_{\text{flat}}(\boldsymbol{u}))^{\mathsf{T}}\left(\nabla\cdot\mathbb{f}_{\text{flat}}(\boldsymbol{u}) - \boldsymbol{R}(\boldsymbol{u})\right), \tag{4.5.11}$$

where $\boldsymbol{R}(\boldsymbol{u}) = (0, \mathbf{0}, q_2, -\frac{\overline{\lambda} g}{\epsilon}\mathsf{h}^2\Gamma(\frac{\eta}{\mathsf{h}}))^{\mathsf{T}}$.

Proof. Let $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q}, q_1, q_2)^{\mathsf{T}}$ be a smooth solution of the hyperbolic relaxed model with no

topography effects (3.2.4). Then,

$$\begin{aligned}
\partial_t \mathcal{E}_{\text{flat}}(\boldsymbol{u}) + \nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u})) &= (\nabla_{\boldsymbol{u}}\mathcal{E}_{\text{flat}})^{\mathsf{T}}\partial_t \boldsymbol{u} + \nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u})) \\
&= (\nabla_{\boldsymbol{u}}\mathcal{E}_{\text{flat}})^{\mathsf{T}}(-\nabla\cdot\mathbb{f}_{\text{flat}}(\boldsymbol{u}) + \boldsymbol{R}(\boldsymbol{u})) + \nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u}) \\
&= 0, \\
&\Longrightarrow \\
\nabla\cdot(\boldsymbol{\mathcal{F}}_{\text{flat}}(\boldsymbol{u})) &= (\nabla\mathcal{E}_{\text{flat}}(\boldsymbol{u}))^{\mathsf{T}}\left(\nabla\cdot\mathbb{f}_{\text{flat}}(\boldsymbol{u}) - \boldsymbol{R}(\boldsymbol{u})\right).
\end{aligned}$$

□

4.5.4 Dispersion relation

In this section, we derive the dispersion relation for the hyperbolic relaxed model (4.4.1). We then compare the relaxed phase velocity to that of the original Serre model (2.3.1). For simplicity, we restrict ourselves to one space-dimension and assume the topography is flat. Notice that since the topography is flat, we can drop the dependency of the q_3 variable (see: Remark 3.2.5).

We proceed similarly as in §2.3.5.3 by considering the linearized problem of (4.4.1) about a rest state $\boldsymbol{u}_0 := (\mathsf{H}_0, 0, \mathsf{H}_0^2, 0)^{\mathsf{T}}$:

$$\partial_t \mathsf{h} + \mathsf{H}_0 \partial_x u = 0, \tag{4.5.12a}$$

$$\partial_t u + g\partial_x \mathsf{h} + 2\frac{\overline{\lambda} g}{\epsilon}\mathsf{H}_0 \partial_x\left(\mathsf{h} - \eta\right) = 0, \tag{4.5.12b}$$

$$\partial_t \eta = \omega, \tag{4.5.12c}$$

$$\partial_t \omega = 6\frac{\overline{\lambda} g}{\epsilon}(\mathsf{h} - \eta). \tag{4.5.12d}$$

We look for solutions in the form of $\boldsymbol{u}(x,t) = \boldsymbol{u}_A \exp(i(kx - \sigma t))$ where $\boldsymbol{u}_A$ are the wave amplitudes, k is the wave number and σ is the wave frequency. Here $i := \sqrt{-1}$ is the imaginary unit number. A direct substitution of the plane wave solution into (4.5.12) yields the following linear

system:

$$\begin{pmatrix} -\sigma & \mathsf{H}_0 k & 0 & 0 \\ gk + 2\frac{\overline{\lambda} g}{\epsilon}\mathsf{H}_0 k & -\sigma & -2\frac{\overline{\lambda} g}{\epsilon} k\mathsf{H}_0 & 0 \\ 0 & 0 & i\sigma & 1 \\ -6\frac{\overline{\lambda} g}{\epsilon} & 0 & 6\frac{\overline{\lambda} g}{\epsilon} & -i\sigma \end{pmatrix} \begin{pmatrix} \mathsf{h}_A \\ u_A \\ \eta_A \\ \omega_A \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Assuming that the solution to the linear system is non-trivial, we get the following equation for σ:

$$\sigma^4 - \left(g\mathsf{H}_0 + 6\frac{\overline{\lambda} g}{\epsilon}\left(\frac{1 + \frac{1}{3}\mathsf{H}_0^2 k^2}{k^2}\right)\right) k^2\sigma^2 + 6\frac{\overline{\lambda} g}{\epsilon} g\mathsf{H}_0 k^2 = 0. \tag{4.5.13}$$

We want to solve this equation for σ to find the linear dispersion relation for the relaxed model (4.4.1). However, we first show that this equation can be simplified to find the Serre dispersion relation. Dividing the equation by $6\frac{\overline{\lambda} g}{\epsilon} g\mathsf{H}_0$ yields:

$$\frac{\epsilon}{6\overline{\lambda} g}\frac{1}{g\mathsf{H}_0}\sigma^4 - \left(\frac{\epsilon}{6\overline{\lambda} g} + \left(\frac{1 + \frac{1}{3}\mathsf{H}_0^2 k^2}{g\mathsf{H}_0 k^2}\right)\right) k^2\sigma^2 + k^2 = 0.$$

Taking the limit $\epsilon \to 0$ the above equation yields:

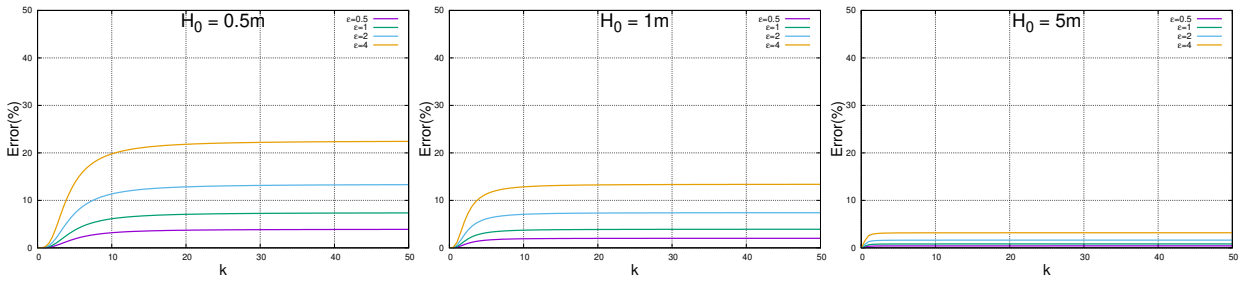
$$-\left(\frac{1 + \frac{1}{3}\mathsf{H}_0^2 k^2}{g\mathsf{H}_0 k^2}\right)\sigma^2 k^2 + k^2 = 0.$$

Solving for σ in the above equation yields the Serre dispersion relation:

$$\sigma^2 = \frac{g\mathsf{H}_0 k^2}{1 + \frac{1}{3}\mathsf{H}_0^2 k^2},$$

which was previously derived in §2.3.5.3. Thus, the dispersion relation for the relaxed system converges to that of the original Serre model as $\epsilon \to 0$.

We continue by giving the expression for the square of the phase velocity $c_p^2 := (\frac{\sigma}{k})^2$. This is



(a) $\mathsf{H}_0 = 0.5\,\mathrm{m}$ (b) $\mathsf{H}_0 = 1\,\mathrm{m}$ (c) $\mathsf{H}_0 = 5\,\mathrm{m}$

Figure 4.3: The error $\frac{c_p^S - c_p^-}{c_p^S}$ as a function of the wave number k for different water depths.

found by solving the polynomial equation (4.5.13) and dividing by k^2:

$$(c_p^{\pm})^2 = \frac{1}{2} g\mathsf{H}_0 + 3\frac{\overline{\lambda} g}{\epsilon}\left(\frac{1+\frac{1}{3}\mathsf{H}_0^2 k^2}{g\mathsf{H}_0 k^2}\right) \pm \frac{1}{2k}\sqrt{k^2\left(g\mathsf{H}_0 + 6\frac{\overline{\lambda} g}{\epsilon}\left(\frac{1+\frac{1}{3}\mathsf{H}_0^2 k^2}{g\mathsf{H}_0 k^2}\right)\right)^2 - 24\frac{\overline{\lambda} g}{\epsilon} g\mathsf{H}_0}. \tag{4.5.14}$$

The phase velocity above bears resemblance to the phase velocity derived in Favrie and Gavrilyuk [20, Eq .19] for a different hyperbolic relaxed model. Just as in [20], we define a slow phase velocity (corresponding to $(c_p^-)^2$) and a rapid phase velocity (corresponding to $(c_p^+)^2$). In Figure 4.4, we plot the slow phase velocity c_p^- as a function of the wave number k for the values $\overline{\lambda} = 1, \mathsf{H}_0 = 1\,\mathrm{m}$ and $\epsilon = \{1, 2, 4, 8\}$. We see that as ϵ decreases the relaxed slow phase velocity approaches that of the Serre model. In Figure 4.3, we show the error (in percent) between the relaxed slow phase velocity and the Serre phase velocity for several reference water depths: $\mathsf{H}_0 = \{0.5\,\mathrm{m}, 1\,\mathrm{m}, 5\,\mathrm{m}\}$. It can be seen that as the water depth decreases, we must decrease the relaxation parameter to capture the dispersive effects. In Figure 4.5, we plot the rapid phase velocity c_p^+ for the same values. We see that c_p^+ tends away from the Serre phase velocity as ϵ decreases. As described in [20], this can be interpreted as the evolution of parasitic high-frequency waves induced by the relaxation. The above results are summarized in the following proposition.

Proposition 4.5.9 (Linear dispersion relation for the relaxed model (4.4.1))**.** *The hyperbolic re-*

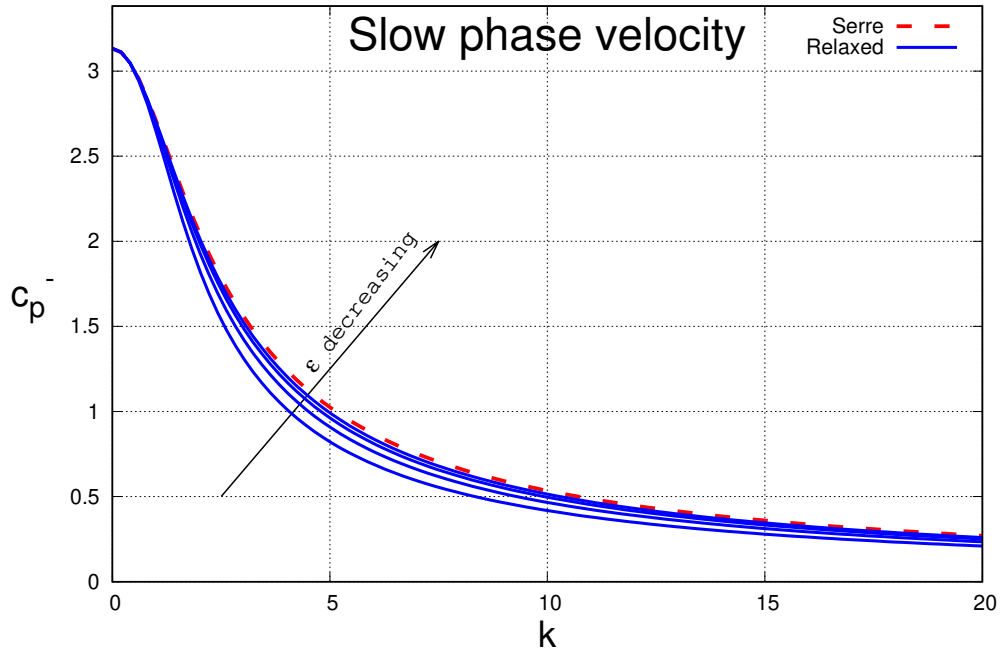


Figure 4.4: Comparison of slow phase velocity of hyperbolic relaxed model (4.4.1) with $\epsilon = \{1, 2, 4, 8\}$ and Serre phase velocity (2.3.23).

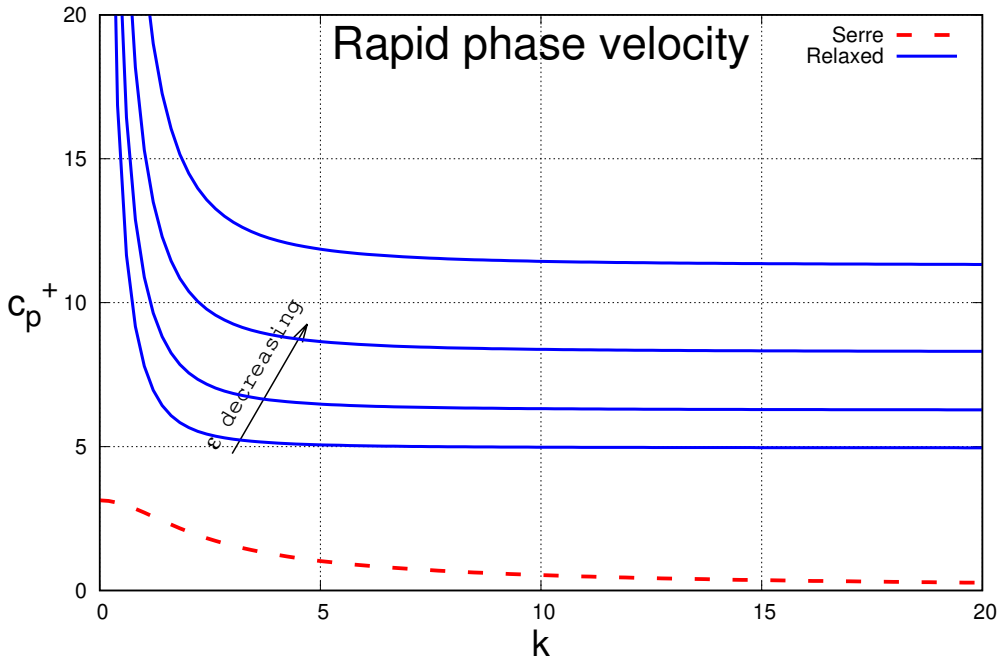


Figure 4.5: Comparison of rapid phase velocity of hyperbolic relaxed model (4.4.1) with $\epsilon = \{1, 2, 4, 8\}$ and Serre phase velocity (2.3.23).

laxed system (4.4.1) *admits the linear (squared) phase velocity* $(c_p^{\pm})^2 := (\frac{\sigma}{k})^2$*:*

$$(c_p^{\pm})^2 = \frac{1}{2}g\mathsf{H}_0 + 3\frac{\overline{\lambda}g}{\epsilon}\left(\frac{1+\frac{1}{3}\mathsf{H}_0^2k^2}{g\mathsf{H}_0k^2}\right) \pm \frac{1}{2k}\sqrt{k^2\left(g\mathsf{H}_0 + 6\frac{\overline{\lambda}g}{\epsilon}\left(\frac{1+\frac{1}{3}\mathsf{H}_0^2k^2}{g\mathsf{H}_0k^2}\right)\right)^2 - 24\frac{\overline{\lambda}g}{\epsilon}g\mathsf{H}_0}.$$

Additionally, the phase velocities $(c_p^{\pm})$ *satisfy the following condition:*

$$0 < c_p^-(\epsilon) < c_p^S < c_p^+(\epsilon),$$

where c_p^S *is the Serre phase velocity* (2.3.23)*.*

4.6 Literature review

In this section, we discuss similar hyperbolic relaxation techniques for solving the Serre Equations seen in the literature. In particular, we give special attention to the work of Favrie and Gavrilyuk [20] which was the motivation for the hyperbolic relaxation technique described above (and many other relaxation techniques). We note that the technique described in [20] was done in one spatial dimension $\mathbb{R}^1$ and extended in Tkachenko [60] for $\mathbb{R}^2$. The other relaxation techniques discussed are those that are extensions to the reformulations discussed in §3.3.

4.6.1 Favrie and Gavrilyuk model

The relaxation approach introduced in [20] is based on a Hamiltonian mechanics point of view. The authors relax the constraint $q_1 := \mathsf{h}^2$ (or equivalently $\eta := \mathsf{h}$) by modifying the Lagrangian of the Serre model:

$$\mathcal{L}(\mathsf{h}, \boldsymbol{q}) = \int_D \left((\frac{1}{2}\mathsf{h}^{-1}\|\boldsymbol{q}\|^2 + \frac{1}{6}\mathsf{h}\dot{\mathsf{h}}^2) - \frac{1}{2}g\mathsf{h}^2\right) \mathrm{d}x,$$

with an augmented Lagrangian:

$$\widehat{\mathcal{L}}(\mathsf{h}, \boldsymbol{q}) = \int_D \left(\frac{1}{2}\mathsf{h}^{-1}\|\boldsymbol{q}\|^2 - \frac{1}{2}g\mathsf{h}^2 + \frac{1}{6}\mathsf{h}\dot{\mathsf{h}}^2 - \frac{1}{6}\overline{\lambda}(\frac{\eta}{\mathsf{h}} - 1)^2\right) \mathrm{d}x,$$

and deriving a new system from this augmented Lagrangian. Recall that the Lagrangian of a mechanical system can be thought of as a functional measuring the difference of the system's

kinetic energy and potential energy.

The relaxed system (with a flat-bottom) derived from the augmented Lagrangian (see [20, 60] for details) is of the form:

$$\partial_t \mathsf{h} + \nabla\cdot(\mathsf{h}\boldsymbol{v}) = 0 \tag{4.6.1a}$$

$$\partial_t \boldsymbol{q} + \nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q}) + \nabla\Big(\frac{1}{2}g\mathsf{h}^2 - \frac{1}{3}\eta\overline{\lambda}(\frac{\eta}{\mathsf{h}} - 1)\Big) = 0 \tag{4.6.1b}$$

$$\partial_t q_1 + \nabla\cdot(q_1\boldsymbol{v}) = q_2 \tag{4.6.1c}$$

$$\partial_t q_2 + \nabla\cdot(q_2\boldsymbol{v}) = -\overline{\lambda}(\frac{\eta}{\mathsf{h}} - 1) \tag{4.6.1d}$$

By directly comparing this model to the one defined by by (4.4.1), we observe that the Hamiltonian strategy proposed in [20] can be re-interpreted of as:

(i) Define the pressure as $p := \frac{1}{2}g\mathsf{h}^2 - \frac{1}{3}\frac{q_1}{\mathsf{h}}s = \frac{1}{2}g\mathsf{h}^2 - \frac{1}{3}\eta s$;

(ii) Replacing s by $-\overline{\lambda}(\frac{\eta}{\mathsf{h}} - 1)$ to relax the constraint $q_1 := \mathsf{h}^2$ where $\overline{\lambda} > 0$ is the relaxation parameter which scales like the square of a velocity.

This relaxed model is shown in [20] to be hyperbolic, but not compatible with dry states.

5. APPROXIMATION

5.1 Introduction

In this chapter, we introduce a space/time approximation of the Saint-Venant model (described in Chapter 2) and the hyperbolic Serre model (described in Chapter 4). The approximation in space is done using continuous finite elements with explicit time stepping. The goal of this chapter is to introduce a high-order (accurate in space) invariant-domain-preserving method that is well-balanced (with respect to rest states) for solving the two models with external physical sources. The approximation is carried out as follows. We first introduce a low-order (i.e., first-order accurate in space) method that is robust with respect to dry states and well-balanced with respect to rest states. We then introduce a provisional high-order method (i.e., second-order accurate in space) using the entropy viscosity methodology that may violate the invariant-domain-preserving property. We then describe a convex limiting technique used for correcting (i.e., limiting) the provisional higher-order method. This is done by enforcing local physical bounds that are naturally satisfied by the low-order method. We give an emphasis on how to apply the convex limiting methodology to hyperbolic systems with sources since it is not well-documented in the literature. The key results of this chapter are: Proposition 5.4.2 and Proposition 5.4.3 which prove the low-order method is invariant-domain-preserving and well-balanced for each model; Theorem 5.7.8 which proves the limited high-order method is invariant-domain-preserving for the Hyperbolic Serre model; and Proposition 5.7.9 which proves that the higher-order method is well-balanced. We note that the algorithms introduced in this chapter are presented in a generic fashion (i.e., independent of the model) unless otherwise specified.

The chapter is organized as follows. In Section 5.2, we introduce the continuous finite element setting for approximating the models in space. In Section 5.2.1, we introduce the finite element representation of the approximate solution. Then, in Section 5.3, we describe the low-order approximation that is invariant-domain-preserving. The details on how to construct the graph-viscosity

for both models is given in this section along with a discussion on the discretization of the external physical sources. In Section 6.4, we give a discussion on the invariant domain preserving and well-balancing properties of the low-order method. Then, in Section 5.6, we introduce a provisional higher-order method that is more accurate, but may violate the invariant domain preserving property. Finally, in Section 5.7, we describe how to apply the convex limiting technique to the provisional higher-order method.

5.2 Finite element setting

In this section, we introduce the continuous finite element setting used for the approximation of the Saint-Venant model (2.2.1) and the hyperbolic Serre model (4.4.1). We note that the techniques shown here can be also be adapted using discontinuous finite elements and finite volumes as discussed in Guermond et al. [32].

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. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be a reference finite element. For every cell $K \in \mathcal{T}_h$ we denote by $T_K : \widehat{K} \to K$ the geometric bijective transformation that maps the reference element $\widehat{K}$ to the current element K. Let $\mathcal{V}$ denote the set of integers that represent the degrees of freedom. Given some mesh $\mathcal{T}_h$, we consider a scalar-valued finite element space $P(\mathcal{T}_h)$ with positive global shape functions $\{\varphi_i\}_{i\in\mathcal{V}}$ associated with the Lagrange nodes $\{\boldsymbol{a}_i\}_{i\in\mathcal{V}}$:

$$P(\mathcal{T}_h) := \{v \in C^0(D;\mathbb{R}) \mid v_{|K} \circ T_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}. \tag{5.2.1}$$

Note that $\dim(P(\mathcal{T}_h)) := \text{card}(\mathcal{V})$. The approximation in space of the conserved variable $\boldsymbol{u}$ is done in the space of $\mathbb{R}^m$-valued finite elements $\boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^m$ where m is the PDE system size. Letting d represent the spatial dimension, we have that $m := d + 1$ for the Saint-Venant model or $m := d + 4$ for the hyperbolic Serre model. For both models, the bottom topography $z(\boldsymbol{x})$ 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

$$\mathcal{I}(i) := \{j \in \mathcal{V} \mid \text{supp}(\varphi_i) \cap \text{supp}(\varphi_j) \neq \emptyset\}. \tag{5.2.2}$$

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:

$$m_{ij} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x, \qquad m_i := \int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}x, \tag{5.2.3a}$$

$$\boldsymbol{c}_{ij} := \int_D \varphi_i \nabla\varphi_j\,\mathrm{d}x, \qquad \boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}, \tag{5.2.3b}$$

where $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. Here m_{ij} represents 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 of the numerical scheme:

(i) $\sum_{i\in\mathcal{V}} \boldsymbol{c}_{ij} = \mathbf{0}$ (partition of unity property);

(ii) $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ if either φ_i or φ_j is zero on ∂D (integration by parts);

(iii) $\sum_{i\in\mathcal{V}} \boldsymbol{c}_{ij} = \mathbf{0}$ if φ_j is zero on ∂D (partition of unity property).

The quantity $\boldsymbol{c}_{ij}$ is used to approximate the gradient (or divergence) operator. Let us expand on this. Let $f(\boldsymbol{x})$ be a sufficiently smooth scalar function and let $\sum_{i\in\mathcal{V}} \mathsf{F}_i\varphi_i(\boldsymbol{x})$ be its finite element approximation. Then, we define its discrete gradient as follows: $(\nabla f)_i := \sum_{j\in\mathcal{I}(i)} \mathsf{F}_j\boldsymbol{c}_{ij}$. Now, let $\boldsymbol{f}(\boldsymbol{x})$ be a sufficiently smooth vector-valued function and let $\sum_{i\in\mathcal{V}} \mathbf{F}_i\varphi_i(\boldsymbol{x})$ be its finite element approximation. Then we define the approximate divergence of $\boldsymbol{f}$ as: $(\nabla\cdot\boldsymbol{f})_i := \sum_{j\in\mathcal{I}(i)} \mathbf{F}_j\cdot\boldsymbol{c}_{ij}$

Remark 5.2.1 (Finite element spaces). In the numerical simulations reported in Chapter 6, we use the following finite element spaces:

(i) The linear, continuous $\mathbb{P}_1$ Lagrange finite elements on unstructured, Delaunay meshes.

(ii) The bi-linear, continuous $\mathbb{Q}_1$ finite elements on quadrangular meshes.

5.2.1 Finite element representations

Since we would like to present the space/time approximation in a model-independent fashion, we first represent quantities and definitions in a generic way. Then, when appropriate, we give explicit definitions for each respective model. The finite element approximation of the conserved variable $\boldsymbol{u}$ at time t is denoted by $\boldsymbol{u}_h(t)$ where

$$\boldsymbol{u}_h(t) = \sum_{i \in \mathcal{V}} \mathbf{U}_i(t) \varphi_i \in \boldsymbol{P}(\mathcal{T}_h).$$

We denote by $z_h := \sum_{i \in \mathcal{V}} \mathsf{Z}_i \varphi_i \in P(\mathcal{T}_h)$ the approximation of the topography map. Let $(\nabla Z)_i := \sum_{j \in \mathcal{I}(i)} Z_j \boldsymbol{c}_{ij}$ denote the approximate gradient of the topography.

Let $\mathsf{H}_{0,\max}$ be some reference scale for the water depth. For instance we can take $\mathsf{H}_{0,\max} := \operatorname{ess\,sup}_{\boldsymbol{x} \in D} \mathsf{h}_0(\boldsymbol{x})$, where $\mathsf{h}_0(\boldsymbol{x})$ is the initial water depth profile. We introduce the regularized water depth quantity H^δ:

$$\mathsf{H}_i^\delta := \left(\frac{2\mathsf{H}_i}{\mathsf{H}_i^2 + \max(\mathsf{H}_i, \delta \mathsf{H}_{0,\max})^2} \right)^{-1} \tag{5.2.4}$$

where δ is a small dimensionless parameter. The quantity H^δ is used for defining the approximate primitive variables and is used to make sense of $1/\mathsf{H}^\delta$. For example, the approximate velocity is defined by $\boldsymbol{v}_h := \sum_{i \in \mathcal{V}} \mathbf{V}_i \varphi_i$ with

$$\mathbf{V}_i := \frac{\mathbf{Q}_i}{\mathsf{H}_i^\delta} \tag{5.2.5}$$

Notice that this regularization is active only when dry state occurs. For example, $\mathbf{V}_i := \frac{1}{\mathsf{H}_i} \mathbf{Q}_i$ if $\mathsf{H}_i \geq \delta \mathsf{H}_{0,\max}$. When $0 \leftarrow \mathsf{H}_i < \delta \mathsf{H}_{0,\max}$, we see that $1/\mathsf{H}^\delta$ approaches 0; thus avoiding the division by 0. Though this may seem un-physical, when H_i is small the momentum $\mathbf{Q}_i$ will be close to 0. The reader is referred to Kurganov and Petrova [43, Eq. (2.17)], Chertock et al. [13, Eq. (3.10)], and [5, 28, §5.1], where this technique is also adopted.

Remark 5.2.2 (Setting the δ quantity). We set $\delta = 10^{-5}$ in the numerical illustrations described in Chapter 6 for both models. We note that it is possible to take δ to be smaller for each model. However, for the hyperbolic Serre model, a smaller δ requires a smaller CFL number.

Remark 5.2.3 (Saint-Venant representation). The finite element representation for the Saint-Venant model is given by $\boldsymbol{u}_h(t) := (\mathsf{h}_h, \boldsymbol{q}_h)^\mathsf{T}$ with

$$\mathbf{U}_i(t) := \Big(\mathsf{H}_i(t), \mathbf{Q}_i(t)\Big)^\mathsf{T}. \tag{5.2.6}$$

Remark 5.2.4 (Hyperbolic Serre representation). The finite element representation for the hyperbolic Serre model is given by: $\boldsymbol{u}_h(t) := (\mathsf{h}_h, \boldsymbol{q}_h, q_{1,h}, q_{2,h}, q_{3,h})^\mathsf{T}$ with

$$\mathbf{U}_i(t) := (\mathsf{H}_i(t), \mathbf{Q}_i(t), \mathsf{Q}_{1,i}(t), \mathsf{Q}_{2,i}(t), \mathsf{Q}_{3,i}(t))^\mathsf{T}. \tag{5.2.7}$$

and approximate auxiliary quantities $\eta_h, \omega_h, \beta_h$ defined with regularization for all $i \in \mathcal{V}$:

$$\mathsf{N}_i := \frac{\mathsf{Q}_{1,i}}{\mathsf{H}_i^\delta}, \quad \mathsf{W}_i := \frac{\mathsf{Q}_{2,i}}{\mathsf{H}_i^\delta}, \quad \mathsf{B}_i := \frac{\mathsf{Q}_{3,i}}{\mathsf{H}_i^\delta}, \tag{5.2.8}$$

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).

5.3 The low-order method

In this section, we describe the low-order approximation of the shallow water flow models (2.2.1) and (2.3.1) 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 by using any explicit strong stability preserving Runge-Kutta method (more on this is discussed in §5.3.8).

5.3.1 Numerical flux and hydrostatic pressure/source

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 $\boldsymbol{u}_0$ at time t^0. Let $n \in \mathbb{N}$ and t^n be the current time. 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 $\boldsymbol{u}$. Recall that the admissible set for both models is: $\mathcal{A} = \{\boldsymbol{u} \mid \mathsf{h} > 0\}$.

We define the numerical gas dynamics flux $\nabla\cdot(\boldsymbol{v}\otimes\boldsymbol{q})$ for all $i\in\mathcal{V}$ and $j\in\mathcal{I}(i)$:

$$\mathbf{G}_{ij}^n := \mathbf{U}_j^n(\mathbf{V}_j^n\cdot\boldsymbol{c}_{ij}) \tag{5.3.1}$$

We now want to discretize the hydrostatic part of the pressure $\nabla\frac{1}{2}g\mathsf{h}^2$ and topography source term $g\mathsf{h}\nabla z$. To ensure well-balancing, we combine the two terms as follows:

$$\begin{aligned}\nabla\tfrac{1}{2}g\mathsf{h}^2 + g\mathsf{h}\nabla z &= g\mathsf{h}\nabla\mathsf{h} + g\mathsf{h}\nabla z\\ &= g\mathsf{h}\nabla\mathsf{h} + g\mathsf{h}\nabla z\\ &= g\mathsf{h}\nabla(\mathsf{h}+z).\end{aligned}$$

Notice that the quantity $g\mathsf{h}\nabla(\mathsf{h}+z)$ shows up directly in the lake-at-rest problem for both models (see: Prop. 2.2.6). We discretize $g\mathsf{h}\nabla(\mathsf{h}+z)$ as follows:

$$\mathsf{P}_{ij}^{\mathrm{SV}} = g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\boldsymbol{c}_{ij} \tag{5.3.2}$$

Remark 5.3.1 (Saint-Venant numerical flux). For all $i\in\mathcal{V}$ and all $j\in\mathcal{I}(i)$, we define the Saint-Venant numerical flux as follows:

$$\mathbf{F}_{ij}^n := \mathbf{G}_{ij}^n + (0,\mathsf{P}_{ij}^{\mathrm{SW}})^{\mathsf{T}}. \tag{5.3.3}$$

Remark 5.3.2 (Hyperbolic Serre numerical flux). For all $i\in\mathcal{V}$ and all $j\in\mathcal{I}(i)$, we define the Hyperbolic Serre numerical flux as follows:

$$\mathbf{F}_{ij}^n := \mathbf{G}_{ij}^n + \left(0,\mathsf{P}_{ij}^{\mathrm{SW}} + \widetilde{\mathsf{P}}(\mathbf{U}_j^n)\boldsymbol{c}_{ij},0,0,0\right)^{\mathsf{T}}, \tag{5.3.4a}$$

$$\widetilde{\mathsf{P}}(\mathbf{U}) := -\frac{\overline{\lambda}g}{3\mathcal{E}}\times\begin{cases}6\mathsf{H}(\mathsf{Q}_1-\mathsf{H}^2), & \text{if } \mathsf{Q}_1\le\mathsf{H}^2\\ 2\frac{(\mathsf{Q}_1-\mathsf{H}^2)}{\mathsf{H}^\delta}(\mathsf{N}^2+\mathsf{Q}_1+\mathsf{H}^2), & \text{if } \mathsf{H}^2<\mathsf{Q}_1.\end{cases} \tag{5.3.4b}$$

where $\widetilde{\mathsf{P}}(\mathbf{U})$ represents the numerical relaxed non-hydrostatic pressure.

5.3.2 Well-balancing star states

We now define the hydrostatic reconstructed star states $\mathbf{U}_i^{*,j,n}$ which ensure that the numerical method is well-balanced with respected to rest states. The concept of well-balancing (originally introduced for the Saint-Venant model) has roots in the work of Bermúdez and Vázquez [9, Def. 1] and Greenberg and Le Roux [27]. In this work, we follow the ideas presented in Audusse et al. [4] (and used in [5, 28]) and define the hydrostatic reconstructed water depth for all $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$ as follows:

$$\mathsf{H}_i^{*,j,n} := \max(0, \mathsf{H}_i^n + \mathsf{Z}_i - \max(\mathsf{Z}_i, \mathsf{Z}_j)). \tag{5.3.5}$$

The hydrostatic reconstructed water depth $\mathsf{H}_i^{*,j,n}$ is used to define the full reconstructed state $\mathbf{U}_i^{*,j,n}$ for each model below.

Remark 5.3.3 (Saint-Venant star states). We define the hydrostatic reconstructed star states for the Saint-Venant model for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$:

$$\mathbf{U}_i^{*,j,n} := \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}} \left(\mathsf{H}_i^n, \mathbf{Q}_i^n\right)^{\mathsf{T}}, \tag{5.3.6a}$$

$$\mathbf{U}_j^{*,i,n} := \frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}} \left(\mathsf{H}_j^n, \mathbf{Q}_j^n\right)^{\mathsf{T}}, \tag{5.3.6b}$$

Remark 5.3.4 (Hyperbolic Serre star states). We define the Hyperbolic Serre hydrostatic reconstructed star states for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$:

$$\mathbf{U}_i^{*,j,n} := \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}} \left(\mathsf{H}_i^n, \mathbf{Q}_i^n, \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}} \mathsf{Q}_{1,i}^n, \mathsf{Q}_{2,i}^n, \mathsf{Q}_{3,i}^n\right)^{\mathsf{T}}, \tag{5.3.7a}$$

$$\mathbf{U}_j^{*,i,n} := \frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}} \left(\mathsf{H}_j^n, \mathbf{Q}_j^n, \frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}} \mathsf{Q}_{1,j}^n, \mathsf{Q}_{2,j}^n, \mathsf{Q}_{3,j}^n\right)^{\mathsf{T}}, \tag{5.3.7b}$$

Note that the star state for the quantity Q_1 in (5.3.7) is defined with the square of the ratio $\left(\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\right)^2$; that is $\mathsf{Q}_{1,i}^{*,j,n} := \left(\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\right)^2$. This is crucial for well-balanced and is seen in the proof of Proposi-

tion 5.4.3. We expand on this in Remark 5.4.1.

5.3.3 PDE source

We now discretize the PDE source for both models.

Remark 5.3.5 (Saint-Venant PDE source). Recall $\boldsymbol{R}(\boldsymbol{u}, \nabla z) = g\mathsf{h}\nabla z$ for the Saint-Venant model. However, since we discretized the hydrostatic pressure and source together in (5.3.2), we set

$$\boldsymbol{R}_i^n = \mathbf{0}. \tag{5.3.8}$$

Remark 5.3.6 (Hyperbolic Serre PDE source). To approximate the PDE source term $\boldsymbol{R}(\boldsymbol{u}, \nabla z)$ in (4.4.1), we introduce the following quantities:

$$\mathsf{R}_1(\mathbf{U}, \nabla Z) := \mathsf{Q}_2 - \tfrac{3}{2}\mathbf{Q} \cdot \nabla Z, \tag{5.3.9a}$$

$$\mathsf{R}_2(\mathbf{U}) := \frac{\overline{\lambda} g}{\mathcal{E}} \times \begin{cases} 6(\mathsf{Q}_1 - \mathsf{H}^2), & \text{if } \mathsf{Q}_1 \leq \mathsf{H}^2, \\ 6\mathsf{N}\frac{(\mathsf{Q}_1 - \mathsf{H}^2)}{\mathsf{H}^\delta}, & \text{if } \mathsf{H}^2 < \mathsf{Q}_1, \end{cases} \tag{5.3.9b}$$

$$\mathsf{R}_3(\mathbf{U}, \nabla Z) := \frac{\overline{\lambda}}{\mathcal{E}} \sqrt{g\mathsf{H}_{0,\max}} (\mathbf{Q} \cdot \nabla Z - \mathsf{Q}_3). \tag{5.3.9c}$$

Then, the Hyperbolic Serre discrete PDE source is defined by

$$\begin{aligned} \boldsymbol{R}_i^n := \Bigg(0, \Big(\tfrac{1}{2}\mathsf{R}_2(\mathbf{U}_i^n) - \tfrac{1}{4}\mathsf{R}_3(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i) \Big) \tfrac{1}{m_i}(\nabla Z)_i, \\ \mathsf{R}_1(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i), -\mathsf{R}_2(\mathbf{U}_i^n), \mathsf{R}_3(\mathbf{U}_i^n, \tfrac{1}{m_i}(\nabla Z)_i) \Bigg)^\mathsf{T}. \end{aligned} \tag{5.3.10}$$

5.3.4 External physical sources

We discretize the Gauckler-Manning friction source term (2.4.1) by setting

$$\boldsymbol{S}_\mathrm{F}(\mathbf{U}_i^n) := \frac{-2gn^2 \mathbf{Q}_i^n \|\mathbf{V}_i\|_{\ell^2}}{(\mathsf{H}_i^n)^\gamma + \max((\mathsf{H}_i^n)^\gamma, 2gn^2\tau_n \|\mathbf{V}_i\|_{\ell^2})} \mathbb{e}_{\boldsymbol{q}}. \tag{5.3.11}$$

Note that a regularization for the quantity $\mathsf{h}^{-\gamma}$ was introduced to avoid division by 0 when $\mathsf{h} \to 0$. This expression was introduced in Guermond et al. [28] and is shown 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 (2.4.2) as follows:

$$\boldsymbol{S}_{\mathrm{G}}(\mathbf{U}_i^n) := -\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}(\mathbf{U}_i^n - \boldsymbol{u}_{\mathrm{wave}}(\boldsymbol{a}_i, t^n))G(\tfrac{\boldsymbol{a}_i - x_{\min}}{L_{\mathrm{gen}}}), \tag{5.3.12}$$

The absorption zone source term (2.4.3) is approximated as follows:

$$\boldsymbol{S}_{\mathrm{A}}(\mathbf{U}_i^n) := -\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\mathbf{U}_{\mathrm{abs},i}^n G(\tfrac{x_{\max} - \boldsymbol{a}_i}{L_{\mathrm{abs}}}), \tag{5.3.13}$$

Remark 5.3.7 (Saint-Venant absorption). The quantities enforced in the absorption zone for the Saint-Venant model are:

$$\mathbf{U}_{\mathrm{abs},i}^n := (0, \mathbf{Q}_i^n)^{\mathsf{T}}.$$

Remark 5.3.8 (Hyperbolic Serre absorption). The quantities enforced in the absorption zone for the hyperbolic Serre model are:

$$\mathbf{U}_{\mathrm{abs},i}^n := (0, \mathbf{Q}_i^n, 0, \mathsf{Q}_{2,i}, 0)^{\mathsf{T}}.$$

5.3.5 Low-order graph-viscosity coefficients

We now define the low-order graph-viscosity coefficients $d_{ij}^{\mathrm{L},n}$ and $\mu_{ij}^{\mathrm{L},n}$ that make the numerical method invariant domain preserving and well-balanced. Here, the superindex $^{\mathrm{L}}$ denotes that the quantities are associated with a (L)ow-order solution (i.e., first-order accurate in space).

For all $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$, we set

$$\mu_{ij}^{\mathrm{L},n} := \max(|\boldsymbol{V}_i^n \cdot \boldsymbol{n}_{ij}| \|\boldsymbol{c}_{ij}\|_{\ell^2}, |\boldsymbol{V}_j^n \cdot \boldsymbol{n}_{ji}| \|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{5.3.14a}$$

$$d_{ij}^{\mathrm{L},n} := \max(\mu_{ij}^{\mathrm{L},n}, \max(\lambda_{ij}^n \|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{ji}^n \|\boldsymbol{c}_{ji}\|_{\ell^2})), \tag{5.3.14b}$$

such that

$$d_{ij}^{\mathrm{L},n} = d_{ji}^{\mathrm{L},n},\ \mu_{ij}^{\mathrm{L},n} = \mu_{ji}^{\mathrm{L},n},\ d_{ij}^{\mathrm{L},n} \geq \mu_{ij}^{\mathrm{L},n} \geq 0, \quad i \neq j.$$

The quantity λ_{ij}^n is some wave speed yet to be defined for each model.

Remark 5.3.9 (Finding λ_{ij}^n for the Saint-Venant model). The wave speed λ_{ij}^n for the Saint-Venant model can be found by deriving a guaranteed maximum wave speed from the associated Riemann problem (see Azerad et al. [5, Sec. 3.3] and Guermond et al. [28, Sec. 4]). For the sake of brevity, we only give the exact formula here and not the derivation.

Let $\partial_t \boldsymbol{w} + \partial_x(\mathbb{f}_{1D}(\boldsymbol{w})) = 0$ be the associated directional 1D Riemann problem with left states

$$(\mathsf{h}_L, \boldsymbol{q}_L \cdot \boldsymbol{n})^\mathsf{T} := (\mathsf{H}_i^n, \mathbf{Q}_i^n \cdot \boldsymbol{n}_{ij})^\mathsf{T},$$

and right states

$$(\mathsf{h}_R, \boldsymbol{q}_R \cdot \boldsymbol{n})^\mathsf{T} := (\mathsf{H}_j^n, \mathbf{Q}_j^n \cdot \boldsymbol{n}_{ji})^\mathsf{T},$$

where $\boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|}$. Let $\mathsf{h}_{\min} = \min(\mathsf{h}_L, \mathsf{h}_R)$ and $\mathsf{h}_{\max} = \max(\mathsf{h}_L, \mathsf{h}_R)$. We set the left and right sound speeds as $a_L = \sqrt{g\mathsf{h}_L}$ and $a_R = \sqrt{g\mathsf{h}_R}$, respectively. Then, λ_{ij}^n for the Saint-Venant model is defined as follows:

$$\lambda_{ij}^n = \max(\lambda_1^-(\overline{\mathsf{h}}_*), \lambda_1^+(\overline{\mathsf{h}}_*)) \tag{5.3.15}$$

with

$$\lambda_1^-(\mathsf{h}_*) = u_L - a_L \sqrt{\Big(1 + \Big(\frac{\mathsf{h}_* - \mathsf{h}_L}{2\mathsf{h}_L}\Big)_+\Big)\Big(1 + \Big(\frac{\mathsf{h}_* - \mathsf{h}_L}{\mathsf{h}_L}\Big)_+\Big)}, \tag{5.3.16a}$$

$$\lambda_2^-(\mathsf{h}_*) = u_R + a_R \sqrt{\Big(1 + \Big(\frac{\mathsf{h}_* - \mathsf{h}_R}{2\mathsf{h}_R}\Big)_+\Big)\Big(1 + \Big(\frac{\mathsf{h}_* - \mathsf{h}_R}{\mathsf{h}_R}\Big)_+\Big)}, \tag{5.3.16b}$$

$$\overline{\mathsf{h}}_* = \begin{cases} \frac{1}{16g}\big(\max(0, u_L - u_R + 2a_L + 2a_R)\big)^2, & \text{if } 0 \le f(x_0\mathsf{h}_{\min}), \\ \sqrt{\mathsf{h}_{\min}\mathsf{h}_{\max}}\big(1 + \frac{\sqrt{2}(u_L - u_R)}{\sqrt{g\mathsf{h}_{\min}} + \sqrt{g\mathsf{h}_{\max}}}\big), & \text{if } f(x_0\mathsf{h}_{\max}) < 0, \\ \Big(-\sqrt{2\mathsf{h}_{\min}} + \sqrt{3\mathsf{h}_{\min}} + 2\sqrt{2\mathsf{h}_{\min}\mathsf{h}_{\max}} + \sqrt{\frac{2}{g}\mathsf{h}_{\min}}(u_L - u_R)\Big)^2, & \text{otherwise.} \end{cases} \tag{5.3.17}$$

Here, $x_0 = (2\sqrt{2} - 1)^2$ and $f(\mathsf{h}) := f_L(\mathsf{h}) + f_R(\mathsf{h}_r) - u_R - u_L$ is the depth function where for $Z = \{L, R\}$

$$f_Z(\mathsf{h}) = \begin{cases} 2(\sqrt{g\mathsf{h}} - \sqrt{g\mathsf{h}_Z}) & \text{if } \mathsf{h} \le \mathsf{h}_Z, \\ (\mathsf{h} - \mathsf{h}_Z)\sqrt{\frac{g(\mathsf{h}+\mathsf{h}_Z)}{2\mathsf{h}\mathsf{h}_Z}} & \text{if } \mathsf{h} > \mathsf{h}_Z. \end{cases} \tag{5.3.18}$$

Remark 5.3.10 (Finding λ_{ij}^n for the Hyperbolic Serre model). The Riemann problem associated with the system (4.4.1) is quite complicated, so in this work, we avoid solving it and define λ_{ij}^n using the eigenvalues established in Proposition 4.5.1. That is, for all $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$:

$$\lambda_{ij}^n = \max(|\boldsymbol{V}_i^n \cdot \boldsymbol{n}_{ij} - (g\mathsf{H}_i^n + \theta_i^n)^{\frac{1}{2}}|, |\boldsymbol{V}_j^n \cdot \boldsymbol{n}_{ij} + (g\mathsf{H}_j^n + \theta_j^n)^{\frac{1}{2}}|) \tag{5.3.19}$$

with $\theta_i^n := \partial_{\mathsf{h}}\widetilde{p}(\mathsf{H}_i^n, \mathsf{N}_i^n)\big(\frac{\mathcal{E}_i}{\max(\mathcal{E}_i, \mathsf{H}_i^n)}\big)^2$.

Notice that when $\mathsf{H}_i^n \le \mathcal{E}_i$, the expressions $\boldsymbol{V}_i^n \cdot \boldsymbol{n}_{ij} - (g\mathsf{H}_i^n + \theta_i^n)^{\frac{1}{2}}$ and $\boldsymbol{V}_i^n \cdot \boldsymbol{n}_{ij} + (g\mathsf{H}_i^n + \theta_i^n)^{\frac{1}{2}}$ are the two extreme eigenvalues of the relaxed system (4.4.1) as established in Proposition 4.5.1. This means that the viscosity accounts for the large wave speeds induced by dispersion only when $\mathsf{H}_i^n \lesssim \mathcal{E}_i$.

5.3.6 Defining the time-step

We define the time-step in the following way:

$$\tau_n := \text{CFL} \times \max_{i \in \mathcal{V}} \left(\frac{m_i}{\sum_{j \in \mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n}} \right), \tag{5.3.20}$$

where CFL is a user-defined positive constant.

5.3.7 Generic low-order update for both models

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 generic low-order update $\mathbf{U}_i^{n+1}$ for all $i \in \mathcal{V}$ is computed as follows:

$$\frac{m_i}{\tau_n}(\mathbf{U}_i^{L,n+1} - \mathbf{U}_i^n) = m_i(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n) + \sum_{j\in\mathcal{I}(i)} -\mathbf{F}_{ij}^n + \sum_{j\in\mathcal{I}^*(i)} \Big((d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) + \mu_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) \Big). \quad (5.3.21)$$

Definition 5.3.1 (Saint-Venant low-order scheme)**.** Let $\boldsymbol{T}_{\mathrm{SV}} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}_{\mathrm{SV}}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ be the low-order Saint-Venant scheme defined by (5.2.6)–(5.3.21).

Definition 5.3.2 (Hyperbolic Serre low-order scheme)**.** Let $\boldsymbol{T}_{\mathrm{HS}} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}_{\mathrm{HS}}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ be the low-order Hyperbolic Serre scheme defined by (5.2.7)–(5.3.21).

5.3.8 Higher-order time stepping

To preserve the invariant domain preserving property while attaining higher-order accuracy in time, we use the the strong stability preserving (SSP) discretization method. In particular, we use SSPRK(3,3): a third-order accurate, three-stage explicit Runge-Kutta method that is SSP. For an overview of higher-order time stepping schemes, we refer the reader to Ern and Guermond [18] and the references therein.

The SSPRK(3,3) method for solving the problem $\partial_t \boldsymbol{u} = L(t, \boldsymbol{u})$ is carried out as follows:

$$\begin{aligned}
\boldsymbol{w}^{(1)} &:= \boldsymbol{u}^n + \tau L(t^n, \boldsymbol{u}^n),\\
\boldsymbol{w}^{(2)} &:= \frac{3}{4}\boldsymbol{u}^n + \frac{1}{4}\Big(\boldsymbol{w}^{(1)} + \tau L(t^n + \tau, \boldsymbol{w}^{(1)})\Big),\\
\boldsymbol{w}^{(3)} &:= \frac{1}{3}\boldsymbol{u}^n + \frac{2}{3}\Big(\boldsymbol{w}^{(2)} + \tau L(t^n + \frac{1}{2}\tau, \boldsymbol{w}^{(2)})\Big)
\end{aligned}$$

Then, the final update at t^{n+1} is given by $\boldsymbol{u}^{n+1} := \boldsymbol{w}^{(3)}$.

5.4 Well-balancing and invariant domain preserving properties

In this section, we show that the low-order algorithm (5.3.21) is well-balanced and invariant domain preserving (i.e., positivity-preserving) for each model. We begin by recalling the precise definitions of the respective properties for each model.

Definition 5.4.1 (Exact rest for Saint-Venant)**.** A numerical state $(\mathsf{h}_h, \boldsymbol{q}_h)^\mathsf{T}$ is said to be exactly at rest if $\boldsymbol{q}_h = \mathbf{0}$ 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 5.4.2 (Exact rest for Hyperbolic Serre)**.** 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 5.4.3 (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.

Remark 5.4.1 (Reconstructed star states at rest). When the numerical state is at rest, we have that by definition:

$$\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n} = \mathbf{0}.$$

We show this holds for each component for the Hyperbolic Serre model using Definition 5.3.7 and Definition 5.4.2.

Assume the numerical state is at rest as defined by Definition 5.4.2. To avoid the trivial case of $\mathsf{H}_j^n = \mathsf{H}_i^n = 0$, we further assume that $\mathsf{H}_i^n \geq \delta \mathsf{H}_{0,\max}$ for all $i \in \mathcal{V}$ (which implies $\mathsf{H}_i^\delta = \mathsf{H}_i^n$). Fix $i \in \mathcal{V}$. Without loss of generality, assume that $Z_i > Z_j$ for a fixed $j \in \mathcal{I}(i)$. Then, $\mathsf{H}_i^{*,j,n} = \mathsf{H}_i^n$ and $\mathsf{H}_j^{*,i,n} = \max(0, \mathsf{H}_j^n + Z_j - Z_i)$. Note that $\mathsf{H}_i^n + Z_i = \mathsf{H}_j^n + Z_j \implies \mathsf{H}_i^n = \mathsf{H}_j^n + Z_j - Z_i$ and since $\mathsf{H}_i^n \geq \delta \mathsf{H}_{0,\max}$, we have that $\mathsf{H}_j^n + Z_j - Z_i > 0$. Thus $\mathsf{H}_j^{*,i,n} = \mathsf{H}_j^n + Z_j - Z_i$. Then, for the water

depth component h we have that

$$\begin{aligned}
\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}}\mathsf{H}_j^n - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\mathsf{H}_i^n &= \frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{n}}\mathsf{H}_j^n - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{n}}\mathsf{H}_i^n \\
&= \mathsf{H}_j^{*,i,n} - \mathsf{H}_i^{*,j,n} \\
&= \mathsf{H}_j^n + Z_j - Z_i - \mathsf{H}_i^n \\
&= \mathsf{H}_j^n + Z_j - (\mathsf{H}_i^n + Z_i) \\
(\text{Def. } 5.3.7) &= 0.
\end{aligned}$$

Since the flow is at rest, $\mathbf{Q}_i^n = \mathbf{0}$ for all $i \in \mathcal{V}$ and consequently $\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}}\mathbf{Q}_j^n - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\mathbf{Q}_i^n = \mathbf{0}$. Similarly, $\mathsf{Q}_{2,i}^n = 0$ and $\mathsf{Q}_{3,i}^n = 0$ for all $i \in \mathcal{V}$ so $\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}}\mathsf{Q}_{2,j}^n - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\mathsf{Q}_{2,i}^n = 0$ and $\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}}\mathsf{Q}_{3,j}^n - \frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}}\mathsf{Q}_{3,i}^n = 0$. We have left to show that the q_1 component of the star state difference vanishes.

Since the flow is at rest, we have that $\mathsf{Q}_{1,i}^n = (\mathsf{H}_i^n)^2$ for all $i \in \mathcal{V}$. Then, we see that:

$$\begin{aligned}
(\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{\delta}})^2\mathsf{Q}_{1,j}^n - (\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{\delta}})^2\mathsf{Q}_{1,i}^n &= (\frac{\mathsf{H}_j^{*,i,n}}{\mathsf{H}_j^{n}})^2\mathsf{Q}_{1,j}^n - (\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^{n}})^2\mathsf{Q}_{1,i}^n \\
&= \frac{(\mathsf{H}_j^n + Z_j - Z_i)^2}{(\mathsf{H}_j^n)^2}\mathsf{Q}_{1,j}^n - \frac{(\mathsf{H}_i^n)^2}{(\mathsf{H}_i^n)^2}\mathsf{Q}_{1,i}^n \\
(\text{Def. } 5.3.7) &= \frac{(\mathsf{H}_i^n + Z_i - Z_i)^2}{(\mathsf{H}_j^n)^2}(\mathsf{H}_j^n)^2 - (\mathsf{H}_i^n)^2 \\
&= \frac{(\mathsf{H}_i^n)^2}{(\mathsf{H}_j^n)^2}(\mathsf{H}_j^n)^2 - (\mathsf{H}_i^n)^2 \\
&= (\mathsf{H}_i^n)^2 - (\mathsf{H}_i^n)^2 \\
&= 0.
\end{aligned}$$

Definition 5.4.4 (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 5.4.2 (Saint-Venant IDP and WB)**.** *Let $\boldsymbol{T}_{SV} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}_{SV} := \boldsymbol{u}_h^{n+1}$ be the Saint-Venant scheme.* (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* $\tau_n(\chi\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}+\frac{1}{m_i}\sum_{j\in\mathcal{I}^*(i)}d_{ij}^n)\leq 1$.

Proof. (i) Since $\boldsymbol{u}_h^n$ is exactly at rest, we have that $\boldsymbol{V}_i^n=\boldsymbol{V}_j^n=0$ or all $i\in\mathcal{V}$ and $j\in\mathcal{I}(i)$, thus $\mu_{ij}^{\mathrm{L},n}=0$. We also have that $\mathbf{U}_i^{*,j,n}=\mathbf{U}_j^{*,i,n}$ as a consequence of the definition (5.3.6). Since $\boldsymbol{q}_h=0$, we have that $\boldsymbol{S}_i^n=\mathbf{0}$. Thus, the viscosity and source terms vanish and we have the leftover terms:

$$\begin{aligned}
&\frac{m_i}{\tau_n}(\mathsf{H}_i^{\mathrm{L},n+1}-\mathsf{H}_i^n)=0,\\
&\frac{m_i}{\tau_n}(\mathbf{Q}_i^{\mathrm{L},n+1}-0)=\sum_{j\in\mathcal{I}^*(i)}\left(g\mathsf{H}_i^n(\mathsf{H}_j^n+Z_j)\boldsymbol{c}_{ij}\right).
\end{aligned}$$

Assume $\mathsf{H}_i=0$ for all $i\in\mathcal{V}$. Then $\mathsf{H}_i^{\mathrm{L},n+1}=0$ and $\mathbf{Q}_i^{\mathrm{L},n+1}=\mathbf{0}$. If $\mathsf{H}_i\neq 0$, then by assumption $\mathsf{H}_j^{n+1}+Z_j=\mathsf{H}_i^n+Z_i$ for all $i\in\mathcal{V}$ and $j\in\mathcal{I}(i)$. This gives:

$$\begin{aligned}
\frac{m_i}{\tau_n}(\mathbf{Q}_i^{\mathrm{L},n+1}-0)&=\sum_{j\in\mathcal{I}^*(i)}\left(g\mathsf{H}_i^n(\mathsf{H}_j^n+Z_j)\boldsymbol{c}_{ij}\right)\\
&=\sum_{j\in\mathcal{I}^*(i)}(g\mathsf{H}_i^n(\mathsf{H}_i^n+Z_i)\boldsymbol{c}_{ij})\\
&=(g\mathsf{H}_i^n(\mathsf{H}_i^n+Z_i))\underbrace{\sum_{j\in\mathcal{I}^*(i)}\boldsymbol{c}_{ij}}_{=\mathbf{0}}\\
&=\mathbf{0}.
\end{aligned}$$

Thus, since u_h is at exact rest and $\mathsf{H}_i^{\mathrm{L},n+1}=\mathsf{H}_i^n$, $\mathbf{Q}_i^{\mathrm{L},n+1}=\mathbf{0}$, the scheme is well-balanced with respect to rest states.

(ii) Referring to (5.3.12), we recall that $S_{\mathsf{h}}^n=-\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}(\mathsf{H}_i^n-\mathsf{h}_{\text{wave}}(\boldsymbol{a}_i,t^n))G(\frac{\boldsymbol{a}_i-x_{\min}}{L_{\text{gen}}})$ is the source in the mass balance equation, where $\mathsf{h}_{\text{wave}}(\boldsymbol{x},t)\geq 0$ for all t and all $\boldsymbol{x}\in D$, and $G\in[0,1]$. Fixing $i\in\mathcal{V}$ and assuming $\mathsf{H}_j^n\geq 0$ for all $j\in\mathcal{I}(i)$, the water height update in (5.3.21) can be arranged

as follows:

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

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

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

The conclusion follows from the condition on τ_n and the definition (5.3.14a). □

Proposition 5.4.3 (Hyperbolic Serre IDP and WB)**.** *Let $\boldsymbol{T}_{HS} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}_{HS}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ be the low-order Hyperbolic Serre scheme.* (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 $\tau_n(\chi\frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i} + \frac{1}{m_i}\sum_{j\in\mathcal{I}^*(i)} d_{ij}^n) \leq 1$.*

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 (5.3.7). Note that it is important that $\mathsf{Q}_{1,i}^{*,j} := \left(\frac{\mathsf{H}_i^{*,j,n}}{\mathsf{H}_i^\delta}\right)^2 \mathsf{Q}_{1,i}$ 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 Proposition 5.4.2.

(ii) The proof is exactly the same as Proposition 5.4.2. □

5.5 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 (5.3.21) (for both models) can be rewritten as a convex combination of auxiliary states.

Let us recall the hydrostatic contribution P_{ij}^{SW} to the numerical flux $\mathbf{F}_{ij}^n$:

$$\sum_{j\in\mathcal{I}(i)} \mathsf{P}_{ij}^{SW} = \sum_{j\in\mathcal{I}(i)} g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\boldsymbol{c}_{ij}.$$

We want to show that this can be re-written using the numerical conservative hydrostatic flux ($\nabla\frac{1}{2}g\mathsf{h}^2$) and a modified source term. We use the algebraic property $ab = -\frac{1}{2}(a-b)^2 + \frac{1}{2}a^2 + \frac{1}{2}b^2$ and the partition of unity properties for $\boldsymbol{c}_{ij}$ to rewrite the above definition:

$$\begin{aligned}
\sum_{j\in\mathcal{I}(i)} \mathsf{P}_{ij}^{SW} &= \sum_{j\in\mathcal{I}(i)} g\mathsf{H}_i^n(\mathsf{H}_j^n + Z_j)\boldsymbol{c}_{ij} \\
&= \sum_{j\in\mathcal{I}(i)} \Big[-\tfrac{1}{2}g(\mathsf{H}_i^n - \mathsf{H}_j^n)^2 + \tfrac{1}{2}g(\mathsf{H}_i^n)^2 + \tfrac{1}{2}g(\mathsf{H}_j^n)^2 + g\mathsf{H}_i^n Z_j \Big]\boldsymbol{c}_{ij} \\
&= \sum_{j\in\mathcal{I}(i)} \Big(\tfrac{1}{2}g(\mathsf{H}_j^n)^2 - \tfrac{1}{2}g(\mathsf{H}_i^n)^2\Big)\boldsymbol{c}_{ij} - \Big[\tfrac{1}{2}g(\mathsf{H}_j^n - \mathsf{H}_i^n)^2 + g\mathsf{H}_i^n Z_j\Big]\boldsymbol{c}_{ij}
\end{aligned}$$

Let $\mathbb{e}_{\boldsymbol{q}}$ denote the characteristic vector for the momentum equations. That is, $\mathbb{e}_{\boldsymbol{q}} = (0, \mathbb{1}_d)^\mathsf{T}$ for the Saint-Venant model and $\mathbb{e}_{\boldsymbol{q}} = (0, \mathbb{1}_d, 0, 0, 0)^\mathsf{T}$ for the Hyperbolic Serre model where d is the spatial dimension. Then, we have the following lemma.

Lemma 5.5.1 (Generic 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*

$$\widetilde{\boldsymbol{R}_i^n} := \boldsymbol{R}_i^n + \boldsymbol{S}_i^n + \sum_{j\in\mathcal{I}(i)}\big(\tfrac{1}{2}g(\mathsf{H}_j^n - \mathsf{H}_i^n)^2 - g\mathsf{H}_i^n Z_j\big)\,\boldsymbol{c}_{ij}\mathbb{e}_{\boldsymbol{q}}. \tag{5.5.1}$$

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:*

$$\boldsymbol{W}_i^{\mathrm{L},n+1} = \boldsymbol{U}_i^n\Big(1 - \frac{\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} 2d_{ij}^{\mathrm{L},n}\Big) + \frac{\tau_n}{m_i}\sum_{j\in\mathcal{I}^*(i)} 2d_{ij}^{\mathrm{L},n}\left(\overline{\boldsymbol{U}_{ij}^n} + \widetilde{\boldsymbol{U}_{ij}^n}\right). \tag{5.5.2}$$

with the auxiliary states defined by

$$\overline{\boldsymbol{U}_{ij}^n} = -\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}\cdot(\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n)) + \frac{1}{2}(\boldsymbol{U}_j^n + \boldsymbol{U}_i^n), \tag{5.5.3a}$$

$$\widetilde{\boldsymbol{U}_{ij}^n} = \frac{d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n}}{2d_{ij}^{\mathrm{L},n}}(\boldsymbol{U}_j^{*,i,n} - \boldsymbol{U}_j^n - (\boldsymbol{U}_i^{*,j,n} - \boldsymbol{U}_i^n)). \tag{5.5.3b}$$

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

Remark 5.5.2 (Generality). Note that the convex combination is again written independently of the mathematical model of interest. Substituting the respective quantities for each model gives the appropriate convex scheme.

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 §5.7 to perform the convex limiting technique. 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 (5.5.2). Let us define the kinetic energy functional $\psi(\boldsymbol{u}) := \frac{1}{2}\frac{1}{\mathsf{h}(\boldsymbol{u})}\|\boldsymbol{q}(\boldsymbol{u})\|_{\ell^2}^2$. Then we have the following local in space and time bounds for each model:

Remark 5.5.3 (Saint-Venant bounds). The local bounds for the Saint-Venant model are given by:

$$\mathsf{h}_i^{n,\min} := \min_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}), \qquad \mathsf{h}_i^{n,\max} := \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}) \tag{5.5.4a}$$

$$\mathsf{K}_i^{n,\max} := \max_{j\in\mathcal{I}(i)} \psi(\overline{\mathbf{U}_{ij}^n} + \widetilde{\mathbf{U}_{ij}^n}). \tag{5.5.4b}$$

Remark 5.5.4 (Hyperbolic Serre bounds). The local bounds for the Hyperbolic Serre model are given by:

$$\mathsf{h}_i^{n,\min} := \min_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}), \qquad \mathsf{h}_i^{n,\max} := \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{H}_{ij}^n} + \widetilde{\mathsf{H}_{ij}^n}), \tag{5.5.5a}$$

$$q_{1,i}^{n,\min} := \min_{j\in\mathcal{I}(i)}(\overline{\mathsf{Q}_{1,ij}^n} + \widetilde{\mathsf{Q}_{1,ij}^n}), \qquad q_{1,i}^{n,\max} := \max_{j\in\mathcal{I}(i)}(\overline{\mathsf{Q}_{1,ij}^n} + \widetilde{\mathsf{Q}_{1,ij}^n}), \tag{5.5.5b}$$

$$\mathsf{K}_i^{n,\max} := \max_{j\in\mathcal{I}(i)} \psi(\overline{\mathbf{U}_{ij}^n} + \widetilde{\mathbf{U}_{ij}^n}). \tag{5.5.5c}$$

Let us expand on the relationship between the local bounds defined above and the convex combination update (5.5.2) with an example. We denote the components of the low-order solutions without sources, $\mathbf{W}^{\mathrm{L}}$, as follows:

$$\mathbf{W}^{\mathrm{L}} := \mathbf{U}(\mathbf{W}^{\mathrm{L}}) = \left(\mathsf{H}(\mathbf{W}^{\mathrm{L}}), \mathbf{Q}(\mathbf{W}^{\mathrm{L}}) \dots\right)^{\mathsf{T}}.$$

Using the properties of convexity, we can extract the following inequality on the water height update $\mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})$ from the convex scheme (5.5.2): or equivalently:

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

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

Thus, these local bounds arise naturally from the low-order numerical scheme. More precise statements are made in §5.7.1.

5.6 Provisional high-order method

In this section, we introduce a provisional higher-order method (second-order accurate in space) that may violate the invariant domain preserving property (i.e., $\mathsf{h} > 0$). 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 viscosity coefficients $d_{ij}^{\mathrm{H},n}, \mu_{ij}^{\mathrm{H},n}$ via the estimation of an entropy residual/commutator using the energy results of both models.

5.6.1 Wave generation

If active, the wave generation mechanism must be modified in the high-order method since this source can potentially create dry states if the amplitude of the wave generated is larger than the mean water depth. We formalize this by setting $\mathsf{h}_{\text{wave}}^{\min} := \min_{\boldsymbol{x}\in D, t>0} \mathsf{h}_{\text{wave}}(\boldsymbol{x}, t)$ and by introducing the cut-off 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 (5.3.21) by setting

$$\boldsymbol{S}_i^n := \boldsymbol{S}_{\mathrm{F}}(\mathbf{U}_i^n) + \chi\Big(\tfrac{\mathsf{h}_i^{i,\max} - \mathsf{h}_i^{i,\min}}{\mathsf{h}_{\mathrm{wave}}^{\min}}\Big)\boldsymbol{S}_{\mathrm{G}}(\mathbf{U}_i^n) + \boldsymbol{S}_{\mathrm{A}}(\mathbf{U}_i^n). \tag{5.6.1}$$

We also use this definition for the high-order update (see (5.6.8)). 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 cut-off is therefore inactive. In particular, it is never active in the simulations reported in Chapter 6. This cut-off is necessary for theoretical purposes (see Theorem 5.7.8).

5.6.2 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 Guermond et al. [28]. The key idea consists of measuring the smoothness of an entropy by measuring how well a chain rule is satisfied by the discretization described above.

Let $(E(\boldsymbol{u}), \boldsymbol{F}(\boldsymbol{u}))$ be any entropy pair that satisfies the following relation:

$$\nabla\cdot(\boldsymbol{F}(\boldsymbol{u})) = (\nabla E(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot(\mathbb{f}(\boldsymbol{u})), \tag{5.6.2}$$

coupled with a respective flux function $\mathbb{f}(\boldsymbol{u})$. We want to estimate the entropy production by inserting the approximate solution in (5.6.2). For all $i \in \mathcal{V}$ we define the entropy commutator as follows:

$$C_i^n := \sum_{j\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\Big(\boldsymbol{F}(\mathbf{U}_j^n) - (\nabla E(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n)\Big) \tag{5.6.3}$$

This quantity measures how well the finite element approximation satisfies the chain rule (5.6.2). 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 resid-

ual/commutator to be

$$R_i^n := \frac{|C_i^n|}{D_i^n}, \tag{5.6.4a}$$

$$D_i^n := \Big| \sum_{i\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot\boldsymbol{F}(\mathbf{U}_j^n)\Big| + \Big| \sum_{i\in\mathcal{I}(i)} \boldsymbol{c}_{ij}\cdot((\nabla E(\mathbf{U}_i^n))^\mathsf{T}\mathbb{f}(\mathbf{U}_j^n))\Big|, \tag{5.6.4b}$$

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

$$d_{ij}^{\mathrm{H},n} = d_{ij}^{\mathrm{L},n}\max(R_i^n, R_j^n), \qquad d_{ii}^{\mathrm{H},n} := -\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{H},n} \tag{5.6.5}$$

$$\mu_{ij}^{\mathrm{H},n} = \mu_{ij}^{\mathrm{L},n}\max(R_i^n, R_j^n), \qquad \mu_{ii}^{\mathrm{H},n} := -\sum_{j\in\mathcal{I}^*(i)} \mu_{ij}^{\mathrm{H},n}. \tag{5.6.6}$$

Notice that $R_i^n \in [0,1]$. Denoting by $\mathrm{diam}(D)$ the diameter of D, it is argued in [31] 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, (5.6.5)-(5.6.6), 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 5.6.1 (The Hyperbolic Serre entropy pair). Recall that it was shown in Proposition 4.5.8 that the Hyperbolic Serre energy functional and energy flux defined in (4.5.8a) do not satisfy the entropy chain rule (5.6.2), but instead satisfy an equation of the form:

$$\nabla\cdot(\boldsymbol{F}(\boldsymbol{u})) = (\nabla E(\boldsymbol{u}))^\mathsf{T}(\nabla\cdot\mathbb{f}(\boldsymbol{u}) - \boldsymbol{S}).$$

Remark 5.6.2 (The Saint-Venant entropy pair). Let $\boldsymbol{u}_{\mathrm{SV}} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ be the Saint-Venant conserved

variable. Recall the following entropy pair for the Saint-Venant model defined in (2.2.3):

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

This pair 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 the Saint-Venant hyperbolic flux.

Remark 5.6.3 (Comparison). In Chapter 6, we use the Saint-Venant entropy pair for the Saint-Venant high-order method and in some of the numerical illustrations for the Hyperbolic Serre model as well. We also show that the convergence behavior of the numerical method with either entropy pair, (4.5.10) or (2.2.3), is similar.

5.6.3 Consistent mass matrix

Numerical dispersion errors induced by the lumped mass matrix can be significantly reduced by using the consistent mass matrix for the discretization of the time derivative term $\boldsymbol{u}_t$ (at least for piecewise linear approximation). For more details on this issue, we refer the reader to Guermond and Pasquetti [29] and the references therein.

We now replace the lumped mass matrix in (5.3.21) with the consistent mass matrix defined in (5.2.3) and the low-order graph viscosity coefficients in (5.3.21) with the entropy-viscosity coefficients (5.6.5)-(5.6.6). Then, the provisional higher-order update (for either the Saint-Venant or Hyperbolic Serre model) is given as follows:

$$\sum_{j\in\mathcal{I}(i)} m_{ij}\frac{\widetilde{\mathbf{U}}_j^{\mathrm{H},n+1}-\mathbf{U}_j^n}{\tau_n} = m_i(\boldsymbol{R}_i^n+\boldsymbol{S}_i^n) + \sum_{j\in\mathcal{I}(i)} -\mathbf{F}_{ij}^n + \sum_{j\in\mathcal{I}^*(i)}\Big((d_{ij}^{\mathrm{H},n}-\mu_{ij}^{\mathrm{H},n})\big(\mathbf{U}_j^{*,i,n}-\mathbf{U}_i^{*,j,n}\big)+\mu_{ij}^{\mathrm{H},n}\big(\mathbf{U}_j^n-\mathbf{U}_i^n\big)\Big). \quad (5.6.8)$$

We see that finding $\widetilde{\mathbf{U}}^{\mathrm{H},n+1}$ in (5.6.8) requires the inversion of the consistent mass matrix at every time step. Since this may be computationally costly, we follow the ideas of [29] and Maier

and Kronbichler [49, 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 (5.6.8) and rewrite (5.6.8) as follows:

$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{m_j}\frac{m_j}{\tau_n}(\widetilde{\mathbf{U}}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n) = \widetilde{\boldsymbol{S}}_i^n. \tag{5.6.9}$$

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 $\sum_{j\in\mathcal{I}(i)} b_{ji} = 0$ (by the partition of unity), 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

$$\begin{aligned}\frac{m_i}{\tau_n}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n) = m_i(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n) + \sum_{j\in\mathcal{I}(i)} -\mathbf{F}_{ij}^n \\ + \sum_{j\in\mathcal{I}^*(i)} \Big(b_{ij}\widetilde{\boldsymbol{S}}_j^n - b_{ji}\widetilde{\boldsymbol{S}}_i^n + (d_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{H},n})\big(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}\big) + \mu_{ij}^{\mathrm{H},n}\big(\mathbf{U}_j^n - \mathbf{U}_i^n\big)\Big).\end{aligned} \tag{5.6.10}$$

Remark 5.6.4 (Generality). The details described above for the provisional high-order method were independent of the mathematical model.

5.6.4 Loss of positivity

It is proved in Guermond et al. [30, Theorm 3.2] that the presence of the consistent mass matrix in any scheme that uses continuous finite elements based on artificial viscosity (such as (5.6.8)) and explicit time stepping violates the maximum principle for scalar conservation laws. A consequence of this result is that the scheme (5.6.8) is non-positivity-preserving regardless of the definition of the artificial viscosity coefficients. It is also observed numerically in [28] that the use of the higher-order entropy viscosity coefficients (5.6.5)-(5.6.6) in (5.3.21) can cause the scheme to be

non-positivity-preserving as well. We correct the loss of positivity in the following section.

5.7 Convex limiting with sources

In this section, we describe the convex limiting technique for both models that is used to make the higher-order methods described above invariant domain preserving (i.e., positivity-preserving). For the work shown here, we build off the ideas presented in Guermond et al. [31, 32], but 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.

5.7.1 Quasiconcave functionals and bounds

We now give some definitions and results that will illustrate the notion of *convex* limiting. Let m be again the problem system size (e.g., $m := d + 1$ for Saint-Venant or $m := d + 4$ for the hyperbolic Serre model).

Definition 5.7.1 (Quasiconcavity)**.** Given a convex set $\mathcal{B} \subset \mathbb{R}^m$, 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 $\lambda \in \mathbb{R}$.

Definition 5.7.2. A function $\Psi : \mathcal{A} \to \mathbb{R}$ is quasiconvex if $-\Psi$ is quasiconcave.

Lemma 5.7.1. *Let $\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^m \mid \mathsf{h} > 0\} \subset \mathbb{R}^{d+4}$. Let $\Psi : \mathcal{B} \to \mathbb{R}$ and assume that the product $\mathsf{h}\Psi$ is concave. Then the function Ψ is quasiconcave.*

Proof. This is a special case of the result in Guermond et al. [32, Lem. 7.4]. □

The key idea behind the convex limiting technique is to correct (i.e., limit) the provisional high-order update (5.6.10) so that it satisfies the same (yet to be defined) quasiconcave constraints as the low-order update (5.7.8). Let $\mathcal{L}$ be an index set denoting the total number of quasiconcave functionals numbered from 1 to $\hat{\ell}$.

Remark 5.7.2 (Quasiconcave functionals for Saint-Venant). Recall that the conserved variable for the system (2.2.1) is $\boldsymbol{u} := (\mathsf{h}, \boldsymbol{q})^\mathsf{T}$ where h is the water depth and $\boldsymbol{q}$ the momentum. The functional

$\Psi_1 : \mathbb{R}^{d+1} \ni (\mathsf{h}, \boldsymbol{q})^\mathsf{T} \mapsto \mathsf{h} \in \mathbb{R}$ is linear, hence concave, hence quasiconcave; this functional is also well-defined over $\mathbb{R}^{d+1}$. Let us set

$$\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+1} \mid \mathsf{h} > 0\} \tag{5.7.1}$$

Observe that $\mathcal{A}$ is convex. Then, another important example is the (negative) kinetic energy $\Psi_2 : \mathcal{A} \ni (\mathsf{h}, \boldsymbol{q})^\mathsf{T} \mapsto -\frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2$. Since the function $\mathsf{h}\Psi_2 := -\frac{1}{2}\|\boldsymbol{q}\|_{\ell^2}^2$ is concave, using Lemma 5.7.1 we conclude the (negative) kinetic energy is quasiconcave. We are going to use the functionals Ψ_1 and Ψ_2 and bounds defined in (5.5.4) to enforce positivity of the water height, positivity and a local maximum principle on the kinetic energy. Letting $\mathcal{L} := \{1{:}3\}$ and $\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+1} \mid \mathsf{h} > 0\} \subset \mathbb{R}^{d+1}$, 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:

$$\Psi_1^{i,n}(\boldsymbol{u}) = \mathsf{h} - \mathsf{h}_i^{n,\min}, \qquad \Psi_2^{i,n}(\boldsymbol{u}) = \mathsf{h}_i^{n,\max} - \mathsf{h}, \tag{5.7.2a}$$

$$\Psi_2^{i,n}(\boldsymbol{u}) = \mathsf{K}_i^{n,\max} - \frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2. \tag{5.7.2b}$$

Remark 5.7.3 (Quasiconcave functionals for Hyperbolic Serre). Recall that the conserved variable for the system (4.4.1) 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

$$\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\} \tag{5.7.3}$$

Observe that $\mathcal{A}$ is convex. Similarly as above, another example is the (negative) kinetic energy $\Psi_3 : \mathcal{A} \ni (\mathsf{h}, \boldsymbol{q}, q_1, q_2, q_3)^\mathsf{T} \mapsto -\frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2$. We use the above functionals and bounds defined in (5.5.5) to enforce positivity of the water height, positivity of the auxiliary variable q_1, and a local maximum

principle on the kinetic energy. 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:

$$\Psi_1^{i,n}(\boldsymbol{u}) = \mathsf{h} - \mathsf{h}_i^{n,\min}, \qquad \Psi_2^{i,n}(\boldsymbol{u}) = \mathsf{h}_i^{n,\max} - \mathsf{h}, \tag{5.7.4a}$$

$$\Psi_3^{i,n}(\boldsymbol{u}) = q_1 - q_{1,i}^{n,\min}, \qquad \Psi_4^{i,n}(\boldsymbol{u}) = q_{1,i}^{\max} - q_1, \tag{5.7.4b}$$

$$\Psi_5^{i,n}(\boldsymbol{u}) = \mathsf{K}_i^{n,\max} - \frac{1}{2\mathsf{h}}\|\boldsymbol{q}\|_{\ell^2}^2. \tag{5.7.4c}$$

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

Lemma 5.7.4. *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* (5.5.2) *is in $\mathcal{A}$ and satisfies the following constraints for all $l \in \mathcal{L}$:*

$$\Psi_l^{i,n}(\boldsymbol{W}_i^{\mathrm{L},n+1}) \geq 0. \tag{5.7.5}$$

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 (5.5.3a)-(5.5.3b); thus, by Lemma 5.5.1, 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 (5.5.2), the definitions of the bounds in §5.7.1 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.

5.7.2 Limiting process

We discuss in this section the proposed convex limiting methodology. The main 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 all 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 §5.7.1. 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 (see Zalesak [67], [28] and references therein), we now compute the difference $\mathbf{W}^{\mathrm{H},n+1} - \mathbf{W}^{\mathrm{L},n+1}$ by subtracting (5.3.21) from (5.6.10). This gives

$$m_i(\mathbf{W}_i^{\mathrm{H},n+1} - \mathbf{W}_i^{\mathrm{L},n+1}) = \sum_{j \in \mathcal{I}^*(i)} \mathbf{A}_{ij}^n, \tag{5.7.6}$$

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

$$\mathbf{A}_{ij}^n := \tau_n \Big[b_{ij}\widetilde{\boldsymbol{S}}_j^n - b_{ji}\widetilde{\boldsymbol{S}}_i^n + \big((d_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{H},n}) - (d_{ij}^{\mathrm{L},n} - \mu_{ij}^{\mathrm{L},n})\big)(\mathbf{U}_j^{*,i,n} - \mathbf{U}_i^{*,j,n}) + (\mu_{ij}^{\mathrm{H},n} - \mu_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n) \Big]. \tag{5.7.7}$$

Notice that $\mathbf{A}_{ij}^n = -\mathbf{A}_{ji}^n$, which implies global mass conservation

$$\sum_{i \in \mathcal{V}} m_i \mathbf{W}_i^{\mathrm{H},n+1} = \sum_{i \in \mathcal{V}} m_i \mathbf{W}_i^{\mathrm{L},n+1},$$

which is equivalent to

$$\sum_{i \in \mathcal{V}} m_i \mathbf{U}_i^{\mathrm{H},n+1} = \sum_{i \in \mathcal{V}} m_i \mathbf{U}_i^{\mathrm{L},n+1}.$$

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

$$\mathbf{W}_i^{n+1} = \sum_{j \in \mathcal{I}^*(i)} \theta_j \Big(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij} \mathbf{P}_{ij}^n \Big), \quad \text{with} \quad \mathbf{P}_{ij}^n := \frac{1}{m_i \theta_j} \mathbf{A}_{ij}^n, \tag{5.7.8}$$

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 $\theta_j := \frac{1}{\text{card}(\mathcal{I}(i))-1}$. 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 5.7.4. The following lemma proved in Guermond et al. [31, Lem. 4.4] is paramount for the convex limiting technique and sums up how to efficiently find the limiting parameters ℓ_{ij}.

Lemma 5.7.5. *Let $\mathcal{A} \subset \mathbb{R}^m$ 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*

$$\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell_j^i \boldsymbol{P}_{ij}^n) = 0, \tag{5.7.9}$$

$\Psi(\boldsymbol{W}_i^{\mathrm{L},n+1} + \ell \boldsymbol{P}_{ij}^n) \geq 0$ *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* (5.7.8)*, then $\Psi(\boldsymbol{W}_i^{n+1}) \geq 0$.*

5.7.3 Application to the system (4.4.1)

We now illustrate Lemma 5.7.5 with $\Psi := \Psi_l^{i,n}, l \in \mathcal{L}$, defined in (5.7.4) and $\mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\}$. For the sake of brevity, we restrict ourselves to discussing the application of the limiting technique only to the Hyperbolic Serre model since application to Saint-Venant model is similar. 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:

$$\ell_j^{i,\mathsf{h}} = \begin{cases} \min\left(\frac{\left|\mathsf{h}_i^{n,\min} - \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{\mathsf{h}}\right| + \delta \mathsf{h}_i^{n,\max}}, 1\right), & \text{if } \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}} < \mathsf{h}_i^{n,\min}, \\ 1, & \mathsf{h}_i^{n,\min} \le \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}} \le \mathsf{h}_i^{n,\max}, \\ \min\left(\frac{\left|\mathsf{h}_i^{n,\max} - \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{\mathsf{h}}\right| + \delta \mathsf{h}_i^{n,\max}}, 1\right), & \text{if } \mathsf{h}_i^{n,\max} < \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{\mathsf{h}}. \end{cases} \tag{5.7.10}$$

This guarantees that $\Psi_1(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \ge 0$ and $\Psi_2(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \ge 0$ 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

$$\ell_j^{i,q_1} = \begin{cases} \min\left(\frac{\left|q_{1,i}^{n,\min} - \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathsf{P}_{ij}^{q_1}\right| + \delta q_{1,i}^{n,\max}}, 1\right), & \text{if } \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1} < q_{1,i}^{n,\min}, \\ 1, & q_{1,i}^{n,\min} \le \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1} \le q_{1,i}^{n,\max}, \\ \min\left(\frac{\left|q_{1,i}^{n,\max} - \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1})\right|}{\left|\mathbf{P}_{ij}^{q_1}\right| + \delta q_{1,i}^{n,\max}}, 1\right), & \text{if } q_{1,i}^{n,\max} < \mathsf{Q}_1(\mathbf{W}_i^{\mathrm{L},n+1}) + \mathsf{P}_{ij}^{q_1}. \end{cases} \tag{5.7.11}$$

This guarantees that $\Psi_3(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \ge 0$ and $\Psi_4(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \ge 0$ 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 5.7.6 (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 [28] where this is shown for the Saint-Venant model.

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 $\Psi_5^{i,n}(\mathbf{W}_i^{\mathrm{L},n+1} + \ell \mathbf{P}_{ij}) \ge 0$ 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 $\Psi_5^{i,n}(\mathbf{U}) \ge 0$ iff $\Phi(\mathbf{U}) \ge 0$ 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

$$a = -\frac{1}{2}\|\mathbf{P}_{ij}^{\boldsymbol{q}}\|_{\ell^2}^2, \tag{5.7.12a}$$

$$b = \mathsf{K}_i^{n,\max}\mathsf{P}_{ij}^{\mathsf{h}} - \mathbf{Q}(\mathbf{W}_i^{\mathrm{L},n+1}) \cdot \mathbf{P}_{ij}^{\boldsymbol{q}}, \tag{5.7.12b}$$

$$c = \mathsf{H}(\mathbf{W}_i^{\mathrm{L},n+1})\mathsf{K}_i^{n,\max} - \frac{1}{2}\|\mathbf{Q}(\mathbf{W}_i^{\mathrm{L},n+1})\|_{\ell^2}^2. \tag{5.7.12c}$$

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

$$\ell_j^{i,\mathsf{K}} = \min(t_0, \ell_j^{i,q_1}). \tag{5.7.13}$$

It is proved in [31] that the definition (5.7.13) 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

$$\ell_{ij} = \min(\ell_j^{i,\mathsf{K}}, \ell_i^{j,\mathsf{K}}). \tag{5.7.14}$$

Then with the above definition and by Lemma 5.7.5, the update $\mathbf{W}_i^{n+1}$ computed by (5.7.8) 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 $\mathbf{U}_i^{n+1}$:

$$\mathbf{U}_i^{n+1} = \tau_n(\boldsymbol{R}_i^n + \boldsymbol{S}_i^n) + \sum_{j \in \mathcal{I}^*(i)} \theta_j \Big(\mathbf{W}_i^{\mathrm{L},n+1} + \ell_{ij} \mathbf{P}_{ij}^n \Big). \tag{5.7.15}$$

Remark 5.7.7 (Saint-Venant limiting). To apply the limiting process described above to the Saint-Venant model, one just has to skip the limiting on the q_1 functionals (i.e., Ψ_4 and Ψ_5).

Theorem 5.7.8 (Invariant domain preserving property for limited scheme)**.** *Let $i \in \mathcal{V}$ and $n \geq 0$. Assume that $\boldsymbol{U}_j^n \in \mathcal{A} := \{\boldsymbol{u} \in \mathbb{R}^{d+4} \mid \mathsf{h} > 0\}$ for all $j \in \mathcal{I}(i)$. Suppose that the time step τ_n is small enough so that $\tau_n \max\left(\frac{2}{m_i}\sum_{j \in \mathcal{I}^*(i)} d_{ij}^{\mathrm{L},n}, \frac{\sqrt{g\mathsf{H}_0}}{\mathcal{E}_i}\right) \leq 1$. Let $\boldsymbol{W}_i^{n+1}$ be defined by* (5.7.8) *with the limiter ℓ_{ij} given by* (5.7.14)*. Then $\boldsymbol{W}_i^{n+1} \in \mathcal{A}$. Consequently, the full update $\boldsymbol{U}_i^{n+1}$ defined*

by (5.7.15) *is in $\mathcal{A}$ as well.*

Proof. By construction, the definition (5.7.14) along with Lemma 5.7.5 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}:

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

where we used the fact that $\mathsf{h}_i^{n,\max} \geq \mathsf{H}_i^n$. If the cut-off 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 cut-off function χ is not active (i.e., when $\mathsf{h}_i^{n,\max} - \mathsf{h}_i^{n,\min} < \mathsf{h}_{\text{wave}}^{\min}$), then

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

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

Proposition 5.7.9 (Well-balancing property for limited scheme)**.** *Let $\boldsymbol{T}_{HS} : \boldsymbol{u}_h^h \mapsto \boldsymbol{T}_{HS}(\boldsymbol{u}_h^n) := \boldsymbol{u}_h^{n+1}$ be the high-order scheme defined by* (5.7.15) *for the Hyperbolic Serre model. This scheme is exactly well-balanced if $\boldsymbol{S}_G \equiv \mathbf{0}$.*

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 5.4.3.

□

Remark 5.7.10 (Iterative limiting). We note here that the limiting process described above can be iterated multiple times by observing from (5.7.6) 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 Chapter 6, 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

$\mathbf{U}^{n+1} = \mathbf{W}^{(k_{\max})} + \tau_n(\boldsymbol{R}^n + \boldsymbol{S}^n)$

5.7.4 Relaxation of the bounds

The methodology described above leads to second-order accuracy in the L^1-norm, but the bounds defined in (5.5.4) and (5.5.5) are too tight to make the method higher-order or even second-order in the L^∞-norm in the presence of smooth extrema. A similar phenomena was observed in Khobalatte and Perthame [42] in the context of the Euler Equations for gas dynamics and explained in Guermond et al. [31, Sec. 4.7]. We want to avoid this reduction in accuracy since

we want to model smooth solitary waves and periodic waves with the Hyperbolic Serre system (2.3.1)–(2.3.2). To recover the full accuracy in the L^∞-norm, one should *relax* the bounds (5.5.4) and (5.5.5) for smooth solutions. We briefly discuss the relaxation here and refer the reader to [32, Sec. 7.6] where this is discussed in detail. All the numerical results reported in Chapter 6 are done using this relaxation technique.

The relaxation is done as follows. Let $\Psi_i^{\min}$ and $\Psi_i^{\max}$ denote either a $\min$ or $\max$ bound defined in §5.7.1 and $\Psi_l(\boldsymbol{u})$ the associated functional; for example, $\Psi_i^{\min} = \mathsf{h}_i^{n,\min}$ and $\Psi_1(\boldsymbol{u}) = \mathsf{h}$. For each $l \in \mathcal{L}$ and all $i \in \mathcal{V}$, we set

$$(\Delta^2\Psi_l)_i = \frac{1}{\sum_{j\in\mathcal{I}^*(i)}\beta_{ij}} \sum_{j\in\mathcal{I}^*(i)} \beta_{ij}\Big(\Psi_l(\mathbf{U}_j^n) - \Psi_l(\mathbf{U}_i^n)\Big), \tag{5.7.16}$$

where the coefficients β_{ij} are meant to make the computation linearity-preserving (see: [32, Rem. 6.2]). In the computations reported below, we take $\beta_{ij} = \int_D \nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x$. We then compute an average of the above quantity

$$\overline{(\Delta^2\Psi_l)_i} = \frac{1}{2(\mathrm{card}(\mathcal{I}(i)) - 1)} \sum_{j\in\mathcal{I}^*(i)} \Big(\frac{1}{2}(\Delta^2\Psi_l)_i + \frac{1}{2}(\Delta^2\Psi_l)_j\Big), \tag{5.7.17}$$

Finally, we define the relaxation of $\Psi^{i,\min}$ and $\Psi^{i,\max}$ by redefining them as follows:

$$\Psi^{i,\min} \leftarrow \max\Big((1-s_i)\Psi^{i,\min}, \Psi^{i,\min} - \overline{(\Delta^2\Psi_l)_i}\Big), \tag{5.7.18a}$$

$$\Psi^{i,\max} \leftarrow \min\Big((1+s_i)\Psi^{i,\max}, \Psi^{i\,\max} + \overline{(\Delta^2\Psi_l)_i}\Big), \tag{5.7.18b}$$

where $s_i = (\frac{m_i}{|D|})^{\frac{3}{2d}} \in (0,1)$. Here $|D|$ is the measure of the computational domain (surface area if $d = 2$, or length if $d = 1$).

In the applications reported in Chapter 6 that involve periodic waves, we observed numerically that the above relaxation for the (negative) kinetic energy can still be too restrictive. To remedy

this, we consider the following less restrictive relaxation for limiting the (negative) kinetic energy:

$$\Psi^{i,\max} \leftarrow \min\Big((1+s_i^{\frac{2}{3}})\Psi^{i,\max}, 0\Big). \tag{5.7.19}$$

6. NUMERICAL ILLUSTRATIONS

6.1 Introduction

In this Chapter, we conclude the thesis by illustrating the performance of the hyperbolic relaxed Serre model (4.4.1) and the associated proposed convex limiting method described in Chapter 5 in both spatial dimensions ($\mathbb{R}^d$ with $d = \{1, 2\}$). We focus on numerical illustrations for the Hyperbolic Serre model (4.4.1), but consider tests with the Saint-Venant model as well. In particular, we verify the accuracy of the method using analytical solutions of the Serre model and then again with the method of manufactured solutions. We then show the method is well-balanced with and without the effects of topography. We perform several numerical tests involving the Riemann Problem for the hyperbolic relaxed system (4.4.1). We also reproduce several academic benchmarks proposed in the literature and then introduce some new benchmarks as well. We conclude the chapter by reproducing several laboratory experiments that validate the hyperbolic relaxed model.

The chapter is organized as follows. In Section 6.2, we discuss the implementation details for the numerical method. Then in Section 6.3, we verify the accuracy of the convex limiting numerical method (5.7.15). In Section 6.4, we verify the numerical method is indeed well-balanced for both models. In Section 6.5, we investigate the numerical solution of the Riemann problem for the Serre model and compare with the Saint-Venant solution. Then in Section 6.6, we reproduce different benchmarks introduced in the literature and propose new benchmarks with interesting properties. Finally, in Section 6.7, we reproduce several laboratory experiments seen in the literature.

6.2 Preliminaries

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, the relaxation is done with $\overline{\lambda} = 1$ and $\mathcal{E}_i = m_i^{\frac{1}{d}}$, 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. All 1D results are done with the `TAMU` code. The second code has been written at the US Army Engineer Research and Development Center using the `Proteus` toolkit (the reader is referred to Kees and Farthing [41]). Both codes use continuous $\mathbb{P}_1$ Lagrange elements on triangles and unstructured, non-nested, Delaunay meshes. The third code is `Ryujin` [49, 34], a high-performance finite-element solver based on the `deal.II` library Arndt et al. [3] 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).

6.3 Convergence tests

In this section we verify the accuracy of the method by performing several convergence tests with the `TAMU` code described in 6.2. Similar tests have been performed with the `Proteus` and `Ryujin`, but are omitted here for the sake of brevity. The goal of this section is to show that when the relaxation parameter ϵ of the Hyperbolic Serre model (4.4.1) is chosen to be proportional to the mesh-size, the numerical solution of this model converges to that of the original Serre model (2.3.1).

6.3.1 Solitary wave solution of Serre model (2.3.1)

The Serre model (2.3.1) admits an exact solution in the form of a solitary wave propagating over a flat bottom. The goal of this particular test is to show that solutions of the relaxed model (4.4.1) converge with a first-order rate with respect to the mesh-size to solutions of the Serre model (2.3.1). Note that the hyperbolic relaxed model (4.4.1) does *not* analytically support exact solitary waves; more on this is discussed in §6.3.2.

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

$$\tilde{\mathsf{h}}(x,t) = \mathsf{h}_0 + \frac{\alpha}{(\cosh(r(x-x_0-ct)))^2}, \qquad \tilde{u}(x,t) = c\frac{\tilde{\mathsf{h}}(x,t)-\mathsf{h}_0}{\tilde{\mathsf{h}}(x,t)}, \tag{6.3.1}$$

I	E_1		E_2		E_∞	
100	1.546E-04	Rate	3.796E-04	rate	8.383E-04	Rate
200	4.509E-05	1.78	7.910E-05	2.26	1.272E-04	2.72
400	3.141E-05	0.52	7.258E-05	0.12	3.025E-04	-1.25
800	2.317E-05	0.44	5.509E-05	0.4	2.396E-04	0.34
1600	1.369E-05	0.76	3.155E-05	0.8	1.376E-04	0.8
3200	7.533E-06	0.86	1.733E-05	0.86	7.551E-05	0.87
6400	3.815E-06	0.98	8.869E-06	0.97	3.871E-05	0.96

Table 6.1: Convergence rates for solitary wave solution $T = 50\,\mathrm{s}$, CFL $= 0.05$.

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

$$\mathsf{h}(x,0) = \max\{\tilde{\mathsf{h}}(x,0) - z(x), 0\}, \qquad q(x,0) = \mathsf{h}(x,0)\tilde{u}(x,0), \tag{6.3.2}$$

with $z(x) \equiv 0$. Note that the variables q_1, q_2, and q_3 are initialized with (6.4).

We consider a 1D uniform grid on the domain $D = (0, 1000\,\mathrm{m})$. We set $\mathsf{h}_0 = 10\,\mathrm{m}$ and $\alpha = 1\,\mathrm{m}$. The solitary wave is initiated at $x_0 = 250\,\mathrm{m}$. The final time is set to $T = 50\,\mathrm{s}$ which allows the solitary wave to travel approximately $519.4\,\mathrm{m}$. Dirichlet conditions are set at the boundaries for all variables. In Table 6.1, we show the numerical results for this problem. 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 the method converges to the solution of the Serre model (2.3.1) with first-order rate with respect to the mesh-size. The first-order rate is a consequence of the relaxation parameter in the relaxed system (4.4.1) being proportional to the local mesh-size. In Figure 6.1, we plot the final solution at $T = 50\,\mathrm{s}$. The solid blue profile represents the elevation $\mathsf{h}(x) + z(x)$.

We now use this analytical solution to verify the accuracy of the numerical method using the different mathematical entropies defined by (4.5.10) and (2.2.3) for the entropy-viscosity graph coefficients (5.6.5) and (5.6.6). The goal of this test is to show that the numerical method 5.7.15

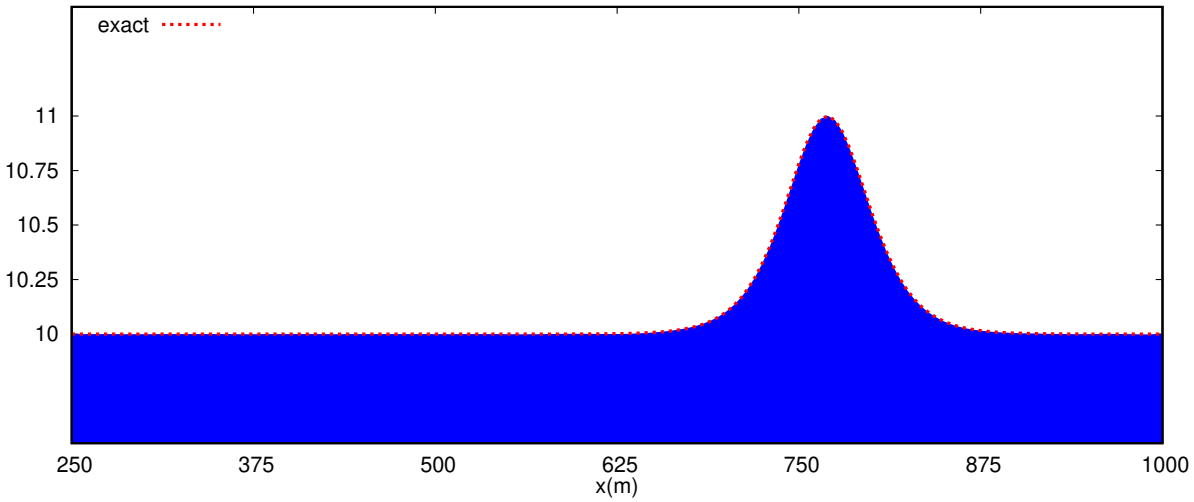


Figure 6.1: Computational solitary wave solution at $T = 50\,\mathrm{s}$ with I =200.

behaves similarly with either entropy. We use the same computational set-up as above. In Table 6.2, we compare the results using (i) the Galerkin method (i.e., no artificial viscosity); (ii) the method (5.7.15) using the Shallow Water Equations entropy pair (2.2.3); (iii) the method (5.7.15) using the hyperbolic Serre entropy pair (4.5.10).

I	Galerkin		EV with (4.5.10)		EV with (2.2.3)	
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

Table 6.2: Convergence table using $\|\mathsf{h} - \mathsf{h}_h\|_{L^1}/\|\mathsf{h}\|_{L^1}$ for solitary wave solution of Serre model (2.3.1). $T = 50\,\mathrm{s}$, CFL$= 0.05$.

6.3.2 Method of manufactured solutions for Hyperbolic Serre model (4.4.1)

We now verify the second-order accuracy of the method (5.7.15). This is typically done by comparing a computational solution to a reference analytical solution and then looking at the be-

havior of the L^1, L^2, L^∞ errors as the mesh-size decreases. However, deriving exact solutions for the hyperbolic relaxation model (4.4.1) is a highly non-trivial task and thus no known solution exists. To test the accuracy of our numerical method for solving the hyperbolic model (4.4.1), we use the method of manufactured solutions. We discuss the details below.

Assume there are no external sources and assume the topography is flat (i.e., $\boldsymbol{S}(\boldsymbol{u}) = \mathbf{0}$ and $\boldsymbol{R}(\boldsymbol{u}, \nabla z) = \mathbf{0}$). Let $\tilde{\mathsf{h}}(x,t)$ and $\tilde{u}(x,t)$ be the same profiles defined by (6.3.1). We want to derive a source term $\boldsymbol{S}_{\text{man}}(x,t)$ for the hyperbolic Serre model (4.4.1) such that

$$\partial_t \boldsymbol{u} + \nabla \cdot \mathbb{f}(\boldsymbol{u}) = \boldsymbol{S}_{\text{man}}(x,t),$$

holds exactly. To do so, we assume the following the profiles for the conserved variables $\boldsymbol{u} = (\mathsf{h}, u, q_1, q_2, q_3)^\mathsf{T}$:

$$\mathsf{h}(\boldsymbol{x},t) = \tilde{\mathsf{h}}(x,t), \qquad q(x,t) = \tilde{\mathsf{h}}(x,t)\tilde{u}(x,t), \tag{6.3.3a}$$

$$q_1(x,t) = \tilde{\mathsf{h}}^2(x,t), \qquad q_2(x,t) = -\tilde{\mathsf{h}}^2(x,t)\partial_x(\tilde{u}(x,t)), \qquad q_3(x,t) = 0. \tag{6.3.3b}$$

Substituting (6.3.3) into (4.4.1), we see that the manufactured source term should be of the form:

$$\boldsymbol{S}_{\text{man}}(x,t) = \big(0, q_{\text{man}}(x,t), 0, q_{2_{\text{man}}}(x,t), 0\big),^\mathsf{T} \tag{6.3.4a}$$

with

$$q_{\text{man}}(x,t) = 2r\alpha\Bigg(\frac{2c^2\mathsf{h}_0^2\sinh(2r(x-x_0-ct))}{(\mathsf{h}_0+2\alpha+\mathsf{h}_0\cosh(2r(x-x_0-ct)))^2} - \frac{g}{\alpha}\tilde{\mathsf{h}}(x,t)(\tilde{\mathsf{h}}(x,t)-\mathsf{h}_0)\tanh(r(x-x_0-ct))\Bigg), \tag{6.3.4b}$$

$$q_{2_{\text{man}}}(x,t) = \frac{2c^2\mathsf{h}_0^2r^2\alpha\left(-3\mathsf{h}_0-4\alpha-2\mathsf{h}_0\cosh(2r(x-x_0-ct))+\frac{\mathsf{h}_0}{\alpha}\cosh(4r(x-x_0-ct))\right)(\tilde{\mathsf{h}}(x,t)-\mathsf{h}_0)}{(\mathsf{h}_0+2\alpha+\mathsf{h}_0\cosh(2r(x-x_0-ct)))^2} \tag{6.3.4c}$$

The profiles defined above were obtained using the `Mathematica` software [65].

The computational domain is set to $D = (0, 1000\,\mathrm{m})$. We initialize the conserved variables with (6.3.3) at $t = 0\,\mathrm{s}$ with $\mathsf{h}_0 = 10\,\mathrm{m}$, $\alpha = 0.1\mathsf{h}_0$, $x_0 = 250\,\mathrm{m}$. The final time is set to $T = 50\,\mathrm{s}$ and CFL $= 0.05$. Dirichlet conditions are set at the boundaries for all variables. We show in Table 6.3 the numerical results obtained at $T = 50\,\mathrm{s}$. We observe that that all the quantities converge with second-order rate with respect to the mesh-size as expected.

I	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

Table 6.3: Convergence rates using manufactured solution. $T = 50\,\mathrm{s}$, CFL $= 0.05$

6.3.3 Steady-state solution with topography

We now verify the accuracy of the relaxation technique using the steady state solution described in Section 2.3.5.2. Recall that this is an analytical solution to the Serre model (2.3.1)–(2.3.2) with topography effects. In particular, the water height profile and bathymetry are given by:

$$\mathsf{h}(x) = \mathsf{h}_0(1 + \frac{a}{(\cosh(rx))^2}), \quad z(x) = -\frac{1}{2}(\mathsf{h}(x) - \mathsf{h}_0)$$

Here, $\mathsf{h}(x)$ describes a stationary solitary wave and the bathymetry is a depression (see Figure 2.2). We set $\mathsf{h}_0 = 1\,\mathrm{m}$, $a = 0.2$, $g = 9.81\,\mathrm{ms}^{-2}$. This gives a discharge value of $q = \sqrt{5.886}\mathrm{m}^2\mathrm{s}^{-1}$ and coefficient $r = \sqrt{0.5}\mathrm{m}^{-1}$ given by the expressions in (2.3.20). The simulations are done with $D = (-10\,\mathrm{m}, 15\,\mathrm{m})$. The discharge is enforced at the inflow boundary $x = -10\,\mathrm{m}$. The water height is enforced at $x = -10\,\mathrm{m}$ and $x = 15\,\mathrm{m}$. We show in Table 6.4 the numerical results obtained at $t = 1000\,\mathrm{s}$. The number of grid points is shown in the leftmost column. The relative

errors on the water height measured in the L^1-norm, $E_1 := \|\mathsf{h}-\mathsf{h}_h\|_{L^1}/\|\mathsf{h}\|_{L^1}$, and the relative errors measured in the L^∞-norm, $E_2 := \|\mathsf{h}-\mathsf{h}_h\|_{L^\infty}/\|\mathsf{h}\|_{L^\infty}$, are shown in the second and third columns. The relative L^1-norm of the difference between h_h^2 and q_{1h}, $E_3 := \|\mathsf{h}_h^2 - q_{1h}\|_{L^1}/\|q_{1h}\|_{L^1}$, and the relative L^1-norm of the difference between $q_h\partial_x z$ and q_{3h}, $E_4 := \|q_h\partial_x z - q_{3h}\|_{L^1}/\|q_h\partial_x z\|_{L^1}$, are shown in the fourth and fifth columns. We observe that all the quantities converge with a first-order rate with respect to the mesh size. This is consistent since we chose the relaxation parameter in the relaxed system (4.4.1) to be proportional to the local mesh size which gives a first-order approximation of the fully coupled system (3.2.2).

I	E_1		E_2		E_3		E_4	
100	1.98E-03	Rate	7.55E-03	rate	8.54E-05	Rate	1.33E-01	Rate
200	1.09E-03	0.86	3.15E-03	1.26	3.83E-05	1.16	6.77E-02	0.97
400	4.23E-04	1.36	1.05E-03	1.58	1.76E-05	1.12	3.40E-02	0.99
800	1.73E-04	1.29	4.07E-04	1.37	8.51E-06	1.05	1.70E-02	1.00
1600	7.92E-05	1.13	1.82E-04	1.16	4.20E-06	1.02	8.51E-03	1.00
3200	3.62E-05	1.13	8.51E-05	1.10	2.09E-06	1.01	4.25E-03	1.00
6400	1.85E-05	0.97	4.31E-05	0.98	1.05E-06	1.00	2.13E-03	1.00

Table 6.4: Convergence rates table for steady state solution with topography.

6.4 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 for the hyperbolic Serre model:

$$\delta_\infty(t) := \frac{\|\mathsf{h}_h(t)-\mathsf{h}_0\|_{L^\infty(D)}}{\mathsf{H}_0} + \frac{\|\boldsymbol{q}_h(t)-\boldsymbol{q}_0\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}} + \frac{\|\mathsf{Q}_{1,h}(t)-\mathsf{Q}_{1,0}\|_{L^\infty(D)}}{\mathsf{H}_0^2} + \frac{\|\mathsf{Q}_{2,h}(t)-\mathsf{Q}_{2,0}\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}} + \frac{\|\mathsf{Q}_{3,h}(t)-\mathsf{Q}_{3,0}\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}}, \quad (6.4.1)$$

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. The quantities $\mathsf{h}_h(t), \boldsymbol{q}_h(t)$,$\mathsf{Q}_{1,h}(t)$,$\mathsf{Q}_{2,h}(t)$, $\mathsf{Q}_{3,h}(t)$ are the finite element approximations at time t for the respective conserved variables. We show that this quantity stays close to round-off error for the numerical

tests.

Remark 6.4.1. The error indicator for the Saint-Venant model is defined by:

$$\delta_\infty(t) := \frac{\|\mathsf{h}_h(t) - \mathsf{h}_0\|_{L^\infty(D)}}{\mathsf{H}_0} + \frac{\|\boldsymbol{q}_h(t) - \boldsymbol{q}_0\|_{L^\infty(D)}}{\mathsf{H}_0\sqrt{g\mathsf{H}_0}} \tag{6.4.2}$$

For these tests, we consider the set up of the 1995 experiments by Briggs et al. [11] where the bathymetry is defined by a conical island (see: §6.7). The experimental domain is given by $D = (0, 25\,\mathrm{m}) \times (0, 30\,\mathrm{m})$. 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 $(12.96\,\mathrm{m}, 13.80\,\mathrm{m})$.

We first consider the case where the initial profile is a complete wet state. The reference water depth is set to $\mathsf{H}_0 = 1\,\mathrm{m}$ (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 $\boldsymbol{q}_0 = 0\,\mathrm{m/s}$. The simulations are run until $T = 50\,\mathrm{s}$. All the computations are done with CFL $= 0.5$. In Table 6.2a, we report the well-balancing quantity (6.4.1) 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

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

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

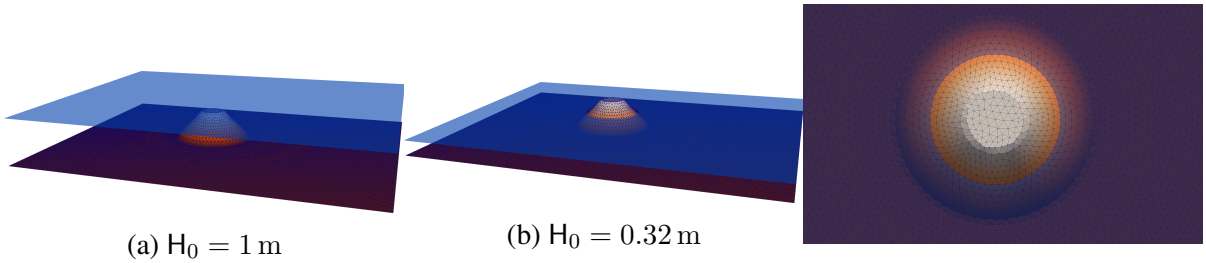
I	$\delta_\infty(t)$
3587	7.7165E-12
14023	1.0692E-11

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

Figure 6.2: Error tables for well-balancing tests using conical island topography.

to be $\mathsf{H}_0 = 0.32\,\mathrm{m}$ so that the water elevation intersects the cone at $r(\boldsymbol{x}) = 2.32\,\mathrm{m}$. 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 Figure 6.3c. In Table 6.2b, we report the well-balancing quantity (6.4.1) for two different meshes and see that it also stays close to machine precision 0. We also compute the δ_∞ error indicator



(a) $\mathsf{H}_0 = 1\,\mathrm{m}$

(b) $\mathsf{H}_0 = 0.32\,\mathrm{m}$

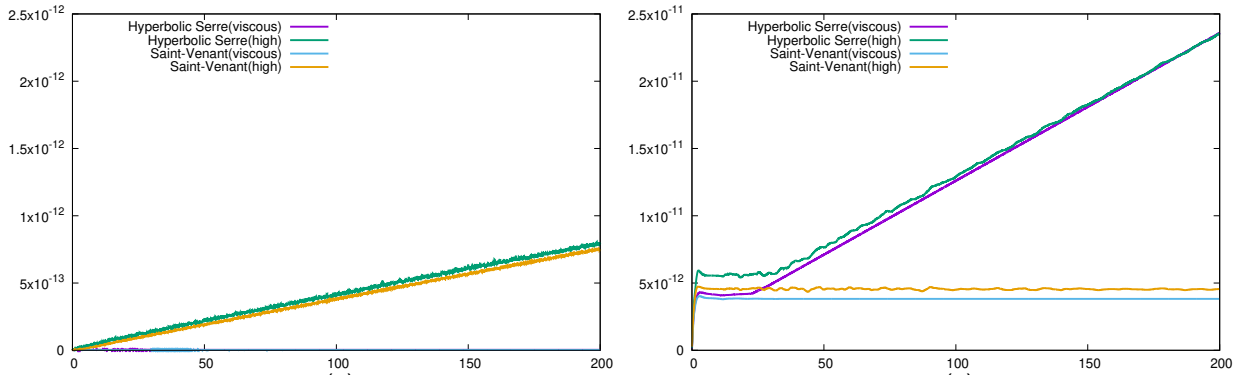
(c) Top view of $\mathsf{h} + z$ with $\mathsf{H}_0 = 0.32\,\mathrm{m}$.

Figure 6.3: Figures for well-balancing tests with conical island topography.

for the Saint-Venant model. In Figure 6.4, we plot the $\delta_\infty(t)$ indicator as a function of time over the interval $t \in [0, 200\,\mathrm{s}]$ for the Saint-Venant low/high-order schemes and the Hyperbolic Serre low/high-order schemes. We see that all quantities for each test stay close to machine precision 0 for long times.

6.5 Riemann problem for Serre Equations

In this section, we investigate numerically the solution to the Riemann Problem for the Serre model. It has been numerically observed that the solutions of the Riemann Problem can produce solutions in the form of dispersive shock waves which can have practical applications in studying undular bores. For a detailed study of the Riemann Problem for the Serre Equations, we refer the reader to El et al. [17] and Gavrilyuk et al. [23] and the references therein.



(a) $\mathsf{H}_0 = 1\,\mathrm{m}$ (b) $\mathsf{H}_0 = 0.32\,\mathrm{m}$

Figure 6.4: A plot of $\delta_\infty(t)$ as a function of time for $t \in [0, 200\,\mathrm{s}]$ for the Hyperbolic Serre model and the Saint-Venant model for the low and high-order schemes.

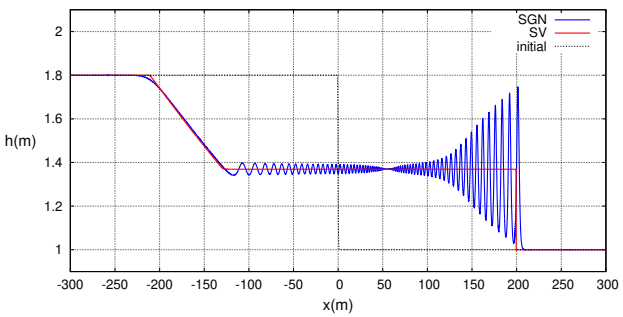
6.5.1 1D – Dam break over wet bed

We consider a one-dimensional dam-break problem introduced in El et al. [17, Sec. 7] with initial conditions:

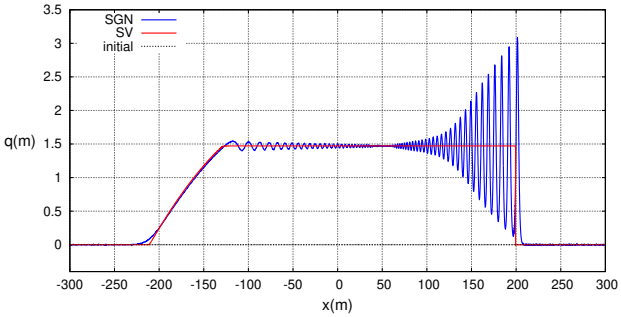
$$\mathsf{h}(x,0) = \begin{cases} 1.8\,\mathrm{m}, & x < 0, \\ 1\,\mathrm{m}, & x > 0, \end{cases}, \quad q(x,0) = 0.$$

The computational domain is defined to be $D = (-300, 300\,\mathrm{m})$. We set the mesh-size to be $h = 0.05\,\mathrm{m}$ corresponding to 12,000 $\mathbb{P}_1$ elements. The final time is set to $T = 50\,\mathrm{s}$ and the CFL number is chosen to be 0.05. In Figure 6.7, we show the computed water depth and momentum (respectively) of both the hyperbolic Serre–Green–Naghdi model (4.4.1) and the Saint-Venant model (2.2.1). The computed profiles for the hyperbolic Serre model (blue profiles) quantitatively look similar to that of El et al. [17, Fig. 8] and Tkachenko [60, Fig. 2.4]. We see that in Figure 6.5, the hyperbolic Serre model produces dispersive a shock wave traveling to the right and an expansion wave traveling to the left followed by a dispersive tail. This structure is fundamentally different than the Saint-Venant solution (red profiles) which produces a right-going shock and a left-going expansion wave.

We now modify the dam-break problem with a smaller jump in the initial water depth. That is,



(a) Case 1 – water depth



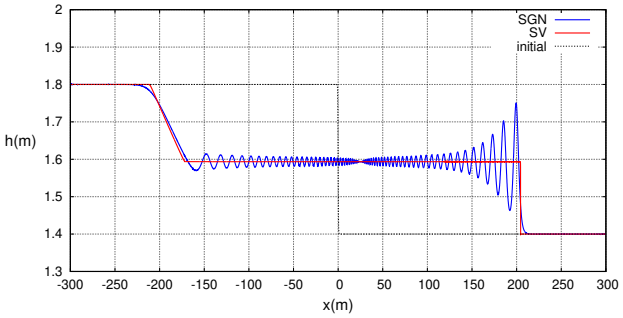
(b) Case 1 – momentum

Figure 6.5: Numerical solution to 1D dam-break problem with dry bed with $\Delta \mathsf{h} = 1.8\,\mathrm{m}$.

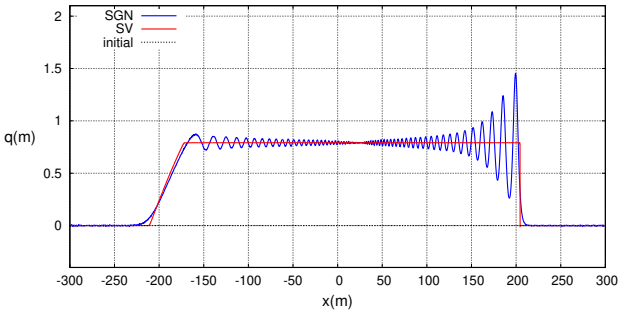
we set the initial conditions to be:

$$\mathsf{h}(x,0) = \begin{cases} 1.8\,\mathrm{m}, & x < 0, \\ \\ 1.4\,\mathrm{m}, & x > 0, \end{cases}, \quad q(x,0) = 0.$$

The computational set-up is the same as above. We show in Figure 6.6 the comparison of the water depth and flow discharge using both models. We see similar profiles as in Figure 6.5.



(a) Case 2 – water depth



(b) Case 2 – momentum

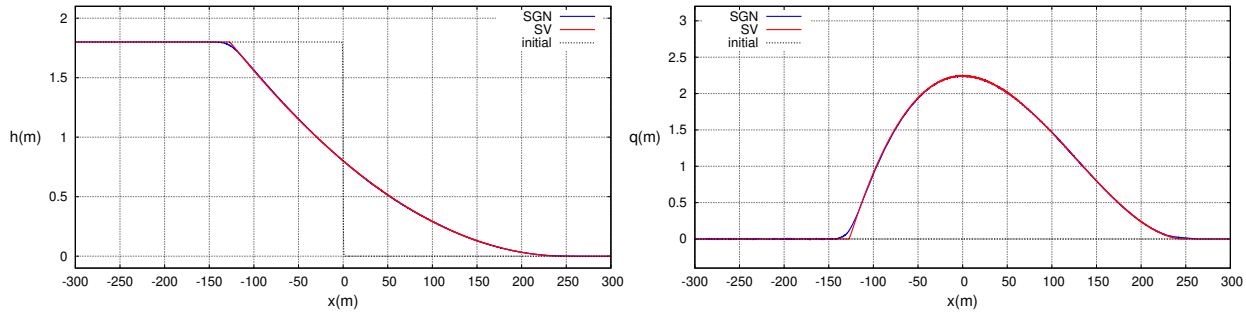
Figure 6.6: Numerical solution to 1D dam-break problem with a wet bed with $\Delta \mathsf{h} = 0.4\,\mathrm{m}$.

6.5.2 1D – Dam break over dry bed

We consider a one-dimensional dam-break over a dry bed. This test case highlights the ability of the proposed numerical methods to handle dry states. We set the initial conditions to be:

$$\mathsf{h}(x,0) = \begin{cases} 1.8\,\mathrm{m}, & x < 0, \\ 0\,\mathrm{m}, & x > 0, \end{cases}, \quad q(x,0) = 0.$$

The computational domain is defined to be $D = (-300, 300\,\mathrm{m})$. We set the mesh-size to be $h = 0.05\,\mathrm{m}$ corresponding to 12,000 $\mathbb{P}_1$ elements. The final time is set to $T = 30\,\mathrm{s}$ and the CFL number is chosen to be 0.05. In Figure 6.7, we show the computed water depth and momentum (respectively) of both the hyperbolic Serre–Green–Naghdi model (4.4.1) and the Saint-Venant model (2.2.1).



(a) Case 1 – water depth

(b) Case 1 – momentum

Figure 6.7: Numerical solution to 1D dam-break problem with a dry bed with $\Delta\mathsf{h} = 1.8\,\mathrm{m}$.

6.5.3 2D – Circular dam break

We now consider a two-dimensional circular dam-break problem. A similar test case was proposed in Tkachenko [60, Sec. 5.2]. Let the still water depth be $\mathsf{H}_0 = 1\,\mathrm{m}$. We initialize circular dam at rest with amplitude $0.8\,\mathrm{m}$ centered at $(250\,\mathrm{m}, 250\,\mathrm{m})$ with radius $R = 40\,\mathrm{m}$. That is to say, the initial condition is set to:

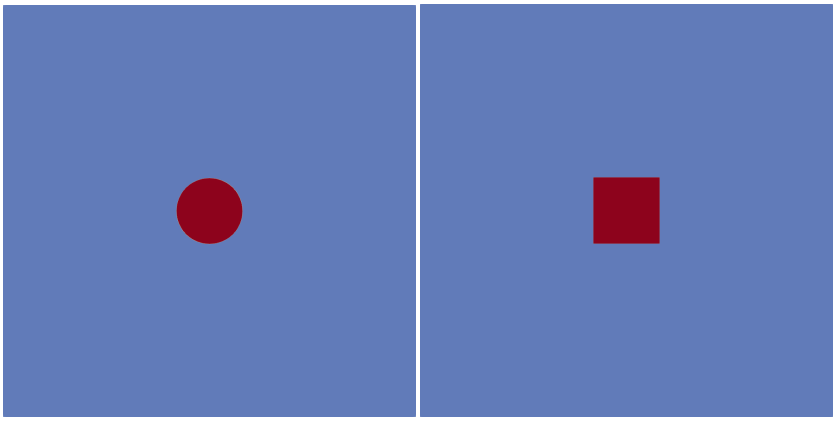


Figure 6.8: Initial profiles for circular and square Riemann Problems.

$$\mathsf{h}(\boldsymbol{x},0) = \begin{cases} 1.8\,\mathrm{m}, & \sqrt{(x-250)^2+(y-250)^2} < R, \\ 1\,\mathrm{m}, & \text{otherwise}, \end{cases}, \quad \boldsymbol{q}(\boldsymbol{x},0) = 0,$$

The initial condition is shown in Figure 6.8 (left). For this test case, we use the `Ryujin` software described in §6.2. The computational domain is set to $D = (0, 500\,\mathrm{m}) \times (0, 500\,\mathrm{m})$ with 4,004,001 $\mathbb{Q}_1$ nodes. The final time is set to $T = 50\,\mathrm{s}$ with CFL number of 0.075. In Figure 6.9, we show the water depth profiles for both the hyperbolic Serre model (top) and the Saint-Venant model (bottom) at the times $t = \{10, 30, 50\}\mathrm{s}$.

6.5.4 2D – Square dam break

We consider a two-dimensional square dam-break problem introduced in Tkachenko [60, Sec. 5.2]. Let the still water depth be $\mathsf{H}_0 = 1\,\mathrm{m}$. The square dam of side length $d_s = 80\,\mathrm{m}$ of amplitude $0.8\,\mathrm{m}$ is initialized as follows:

$$\mathsf{h}(\boldsymbol{x},0) = \begin{cases} 1.8\,\mathrm{m}, & |x-250| \le \frac{d_s}{2} \text{ and } |y-250| \le \frac{d_s}{2}, \\ 1\,\mathrm{m}, & \text{otherwise}, \end{cases}, \quad \boldsymbol{q}(\boldsymbol{x},0) = 0,$$

The initial condition is shown in Figure 6.8 (right). For this test case, we use the `Ryujin` software described in §6.2. The computational domain is set to $D = (0, 500\,\mathrm{m}) \times (0, 500\,\mathrm{m})$ with 4,004,001

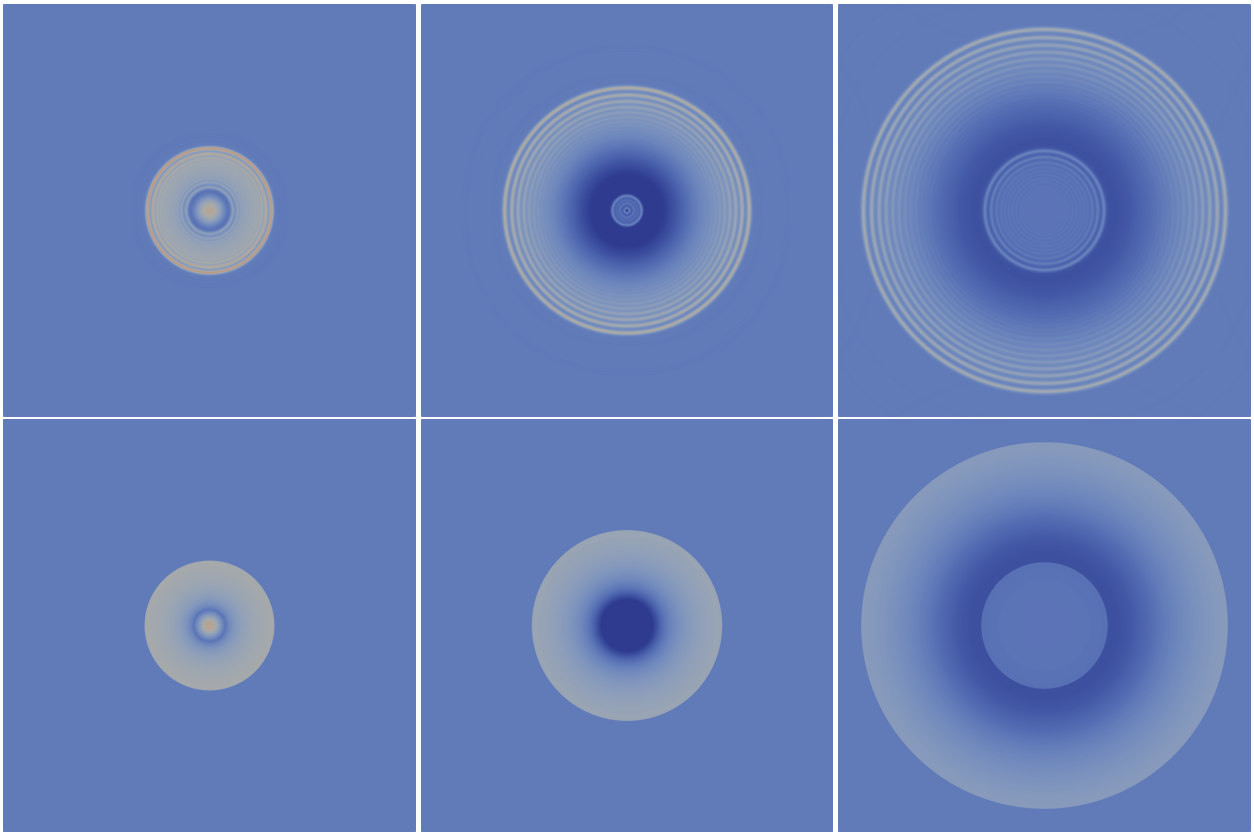


Figure 6.9: Circular dam break – Comparison with Hyperbolic Serre (top) and Saint-Venant (bottom) at $t = \{10, 30, 50\}$s.

$\mathbb{Q}_1$ nodes. The final time is set to $T = 50\,\mathrm{s}$ with CFL number of 0.075. In Figure 6.10, we show the water depth profiles for both the hyperbolic Serre model (top) and the Saint-Venant model (bottom) at the times $t = \{10, 30, 50\}$s.

6.6 Academic benchmarks

In this section, we reproduce several academic benchmarks seen in the literature. Since some of these benchmarks have not been reproduced using the Serre–Green–Naghdi equations, the goal is to highlight the computational results using the for such benchmarks. We also propose new synthetic benchmarks with interesting structures not seen in the literature.

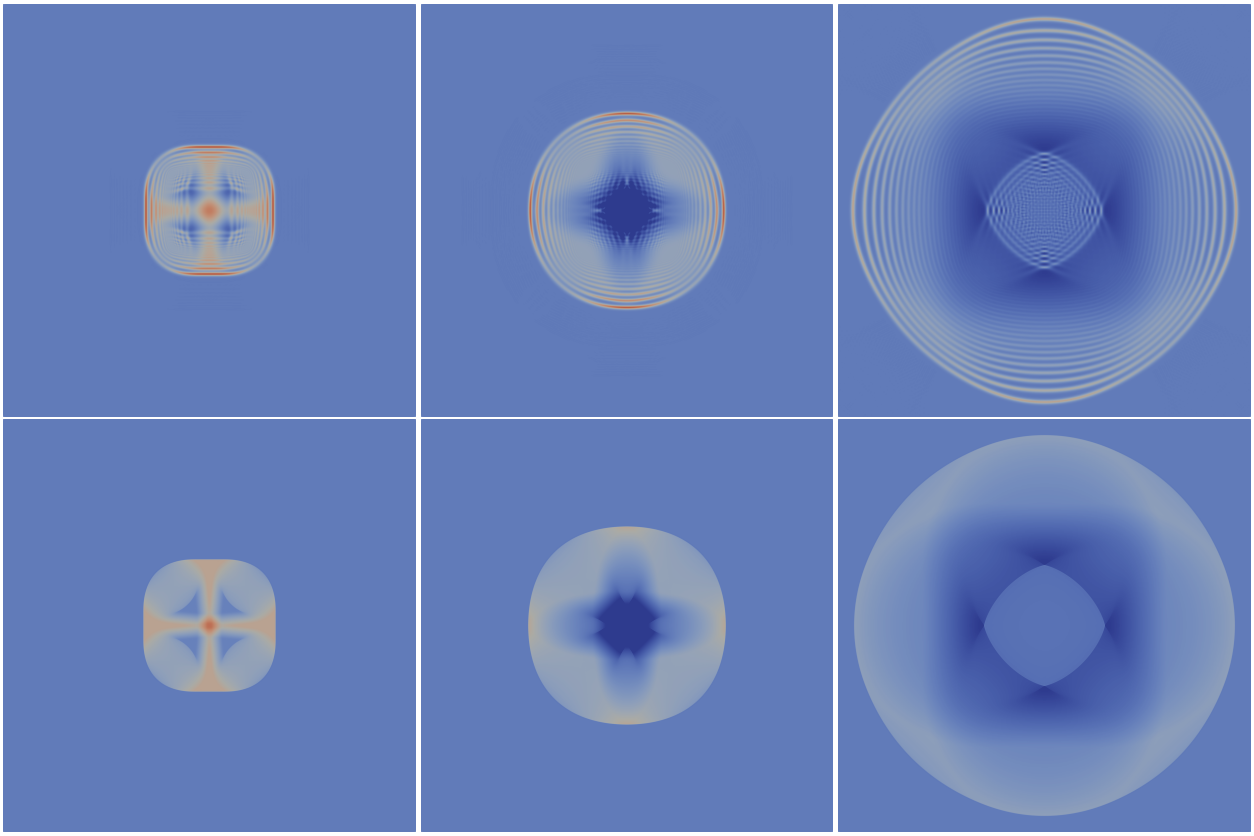


Figure 6.10: Square dam break – Comparison with Hyperbolic Serre (top) and Saint-Venant (bottom) at $t = \{10, 30, 50\}$s.

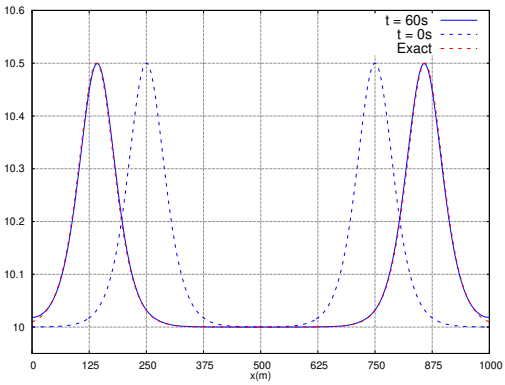
6.6.1 1D – Interaction of solitary waves

We now consider the the collision of multiple solitary waves in one spatial dimension. It is known that the Serre Equations admit quasi-elastic collisions of solitary waves [16] when the amplitude is small (unlike the Korteweg–De Vries (KdV) equation which admits elastic collisions of solitary waves). We say a collision is elastic if the solitary waves are unchanged after the collision. For the Serre Equations, the collision of two solitary waves of small amplitude will produce two waves propagating in the opposite directions with similar shape and height with potentially a small phase shift and loss of amplitude (or none at all). When the amplitudes of the colliding waves are larger, there is a stronger interaction between the waves which leads to a larger phase shift and loss of amplitude and the creation of a dispersive tail. We call this type of collision an

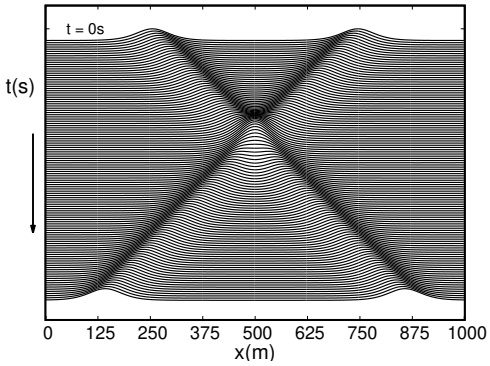
inelastic collision. For the study of such phenomena, we refer the reader to [16, 51] and references therein. Before discussing the different tests, note that for all computations in this section, the local mesh-size is set to $h = 0.5\,\mathrm{m}$ and the CFL number is set to 0.075.

The first test we consider is the collision of two solitary waves. In particular, we consider two different cases that show the phenomena of quasi-elastic collision (when the wave amplitude is small) and inelastic collision (when the wave amplitude is larger). The computational domain is set to $D = (0, 1000\,\mathrm{m})$. To initialize the solitary waves, we use the profiles defined by (6.3.1). We set the reference depth to $\mathsf{h}_0 = 10\,\mathrm{m}$. For the quasi-elastic case, we set the amplitude to be $\alpha = 0.05\mathsf{h}_0 = 0.5\,\mathrm{m}$. The first solitary wave (which travels to the right) is initiated at $x_1 = 250\,\mathrm{m}$ and the second solitary wave (which travels to the left) is initiated at $x_2 = 750\,\mathrm{m}$. For the second solitary wave, we switch the sign of the wave celerity in (6.3.1) (i.e., $c \to -c$) so that the wave travels to the left. The final time is set to $T = 60\,\mathrm{s}$. We show in Figure 6.11 the results for the quasi-elastic collision. In particular, we compare the final numerical solution with the exact solution (at the final time) in Figure 6.11a and see that the profiles are close to each other. In Figure 6.11b, we give a space-time plot which highlights the trajectory of the waves. For the inelastic collision, we set the amplitude of both waves to be $\alpha = 0.1\mathsf{h}_0 = 1\,\mathrm{m}$. We show in Figure 6.12 the results for the inelastic collision. Again, we compare the numerical solution with the exact solution and see that the numerical solution has produced small dispersive tails and a slight phase shift was introduced.

The second test we consider is the collision of four solitary waves. Again we consider two cases that show the phenomena of quasi-elastic collision and inelastic collision. The computational domain is set to $D = (0, 2000\,\mathrm{m})$. We set the reference depth to $\mathsf{h}_0 = 10\,\mathrm{m}$. For the quasi-elastic case, we set the amplitude of each wave to be $\alpha = 0.05\mathsf{h}_0 = 0.5\,\mathrm{m}$. The solitary waves are respectively initiated at $x_1 = 250\,\mathrm{m}\,, x_2 = 650\,\mathrm{m}\,, x_3 = 1350\,\mathrm{m}\,, x_4 = 1750\,\mathrm{m}$. The final time is set to $T = 100\,\mathrm{s}$. We show in Figure 6.13 the results for the quasi-elastic case. We see again that the waves are hardly unchanged and closely match the final exact profiles. For the inelastic collision, we set amplitude of each wave (left to right respectively) to be $\alpha_1 = 2\,\mathrm{m},\ \alpha_2 = 1\,\mathrm{m},\ \alpha_3 = 1\,\mathrm{m}, \alpha_4 = 1\,\mathrm{m}$. It is clear from Figure 6.14 that the larger amplitudes introduce stronger

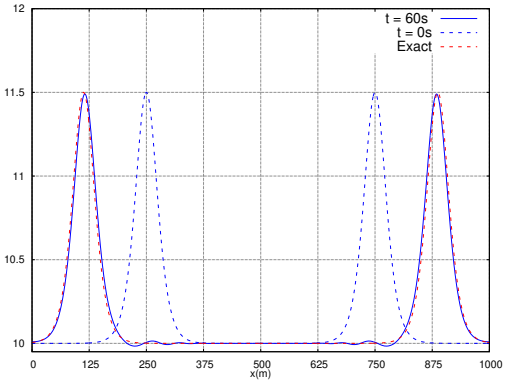


(a) Free surface elevation

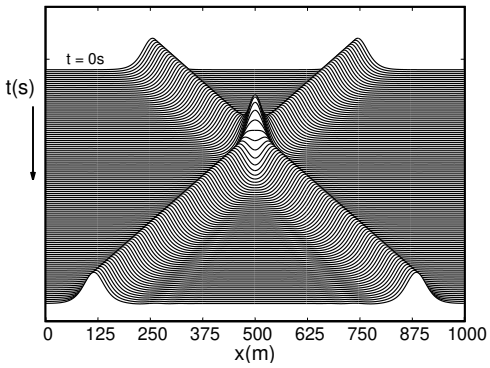


(b) Space-time plot

Figure 6.11: Two solitary waves – quasi-elastic collision

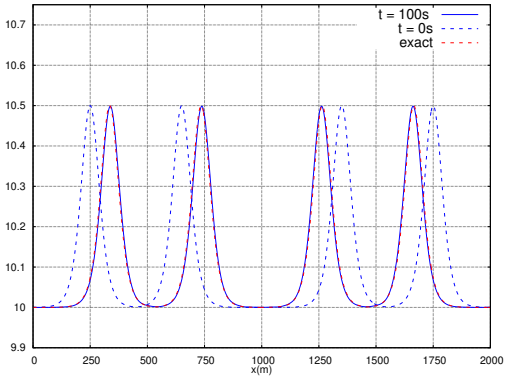


(a) Free surface elevation

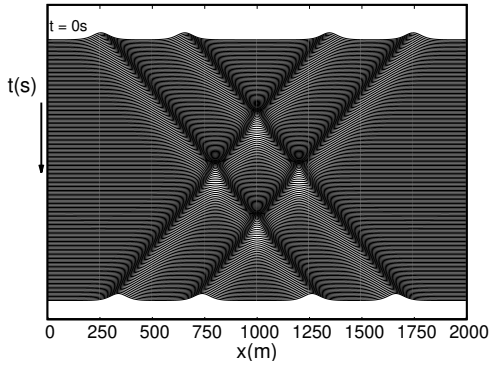


(b) Space-time plot

Figure 6.12: Two solitary waves – inelastic collision

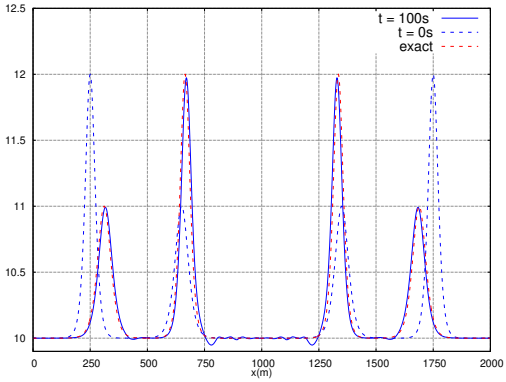


(a) Free surface elevation

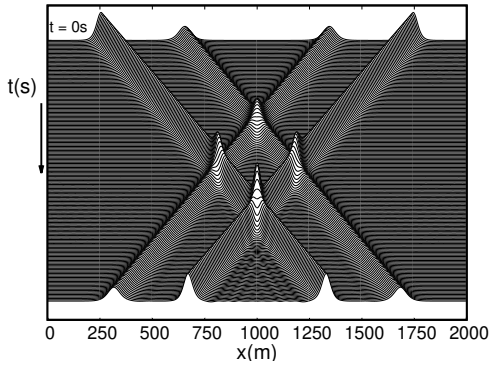


(b) Space-time plot

Figure 6.13: Four solitary waves – quasi-elastic collision



(a) Free surface elevation

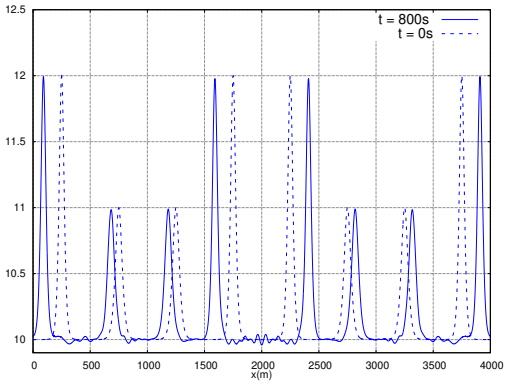


(b) Space-time view

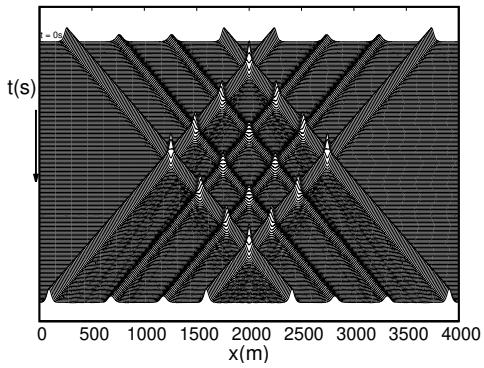
Figure 6.14: Four solitary waves – inelastic collision

interactions and affects the collision elasticity.

The next case we consider is the collision of 8 solitary waves. Since we have established the quasi-elasticity property of the Serre–Green–Naghdi equations, this test shown here is to highlight the robustness of the numerical method. The computational domain is set to $D = (0, 4000\,\mathrm{m})$. We set the reference depth to $\mathsf{h}_0 = 10\,\mathrm{m}$. Letting the integer "1" denote the left-most solitary wave, the amplitudes of the solitary wave are as follows: $\alpha_i = 2\,\mathrm{m}$ for $i = \{1, 4, 5, 8\}$ and $\alpha_j = 1\,\mathrm{m}$ for $i = \{2, 3, 6, 7\}$. The final time is set to $T = 200\,\mathrm{s}$. We show the computations for this test in



(a) Free surface elevation



(b) Space-time plot

Figure 6.15: Collision of Eight solitary waves

Figure 6.15.

6.6.2 2D – Dam Break over three obstacles with friction

We now consider the test case of a dam break over a dry bottom with three conical obstacles introduced by Kawahara and Umetsu [38] (and reproduced by others: Guermond et al. [33], Huang et al. [37], etc.). This benchmark tests the complex wetting/drying process and the method's ability to handle the Gauckler-Manning friction source. Here the Manning's coefficient is set to $n = 0.02\,\mathrm{m}^{-1/3}\mathrm{s}$.

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

$$z_1(\boldsymbol{x}) = 1 - \tfrac{1}{8}\sqrt{(x-30)^2 + (y-6)^2}, \tag{6.6.1a}$$

$$z_2(\boldsymbol{x}) = 1 - \tfrac{1}{8}\sqrt{(x-30)^2 + (y-24)^2}, \tag{6.6.1b}$$

$$z_3(\boldsymbol{x}) = 3 - \tfrac{3}{10}\sqrt{(x-47.5)^2 + (y-15)^2}. \tag{6.6.1c}$$

are thre three obstacles. The initial state is set to

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

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 $T = 20\,\mathrm{s}$ with CFL 0.075. In Figure 6.16, we show the computational free surface elevation $\mathsf{h} + z$ at several time snapshots. Then, in Figure 6.17, we show the comparison of the computations with the hyperbolic Serre model (4.4.1) and the Saint-Venant shallow water equations (2.2.1). We observe that more realistic structures are produced by the dispersive Serre model.

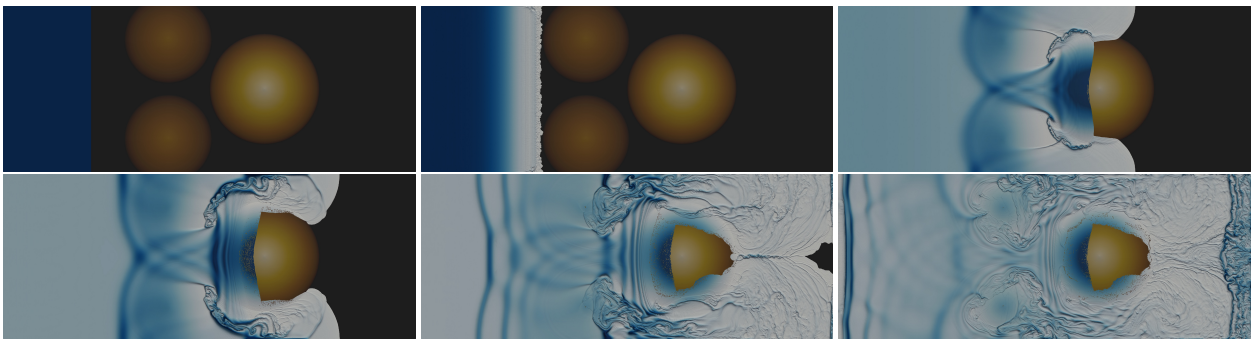


Figure 6.16: Dam break with bumps – Surface plot of the water elevation $\mathsf{h} + z$ at several time snapshots.

6.6.3 2D – Circular Dam Break with divergence-free velocity

We now reproduce a modified version of the circular dam break problem introduced in Section 6.5.3. Instead of initiating the dam at rest, we introduce a divergent-free velocity field so that circular dam is initially "spinning". More precisely, we initialize the circular dam with amplitude $0.8\,\mathrm{m}$ centered at $(x_c, y_c) = (250\,\mathrm{m}, 250\,\mathrm{m})$ with radius $R = 40\,\mathrm{m}$ and set the initial conditions to

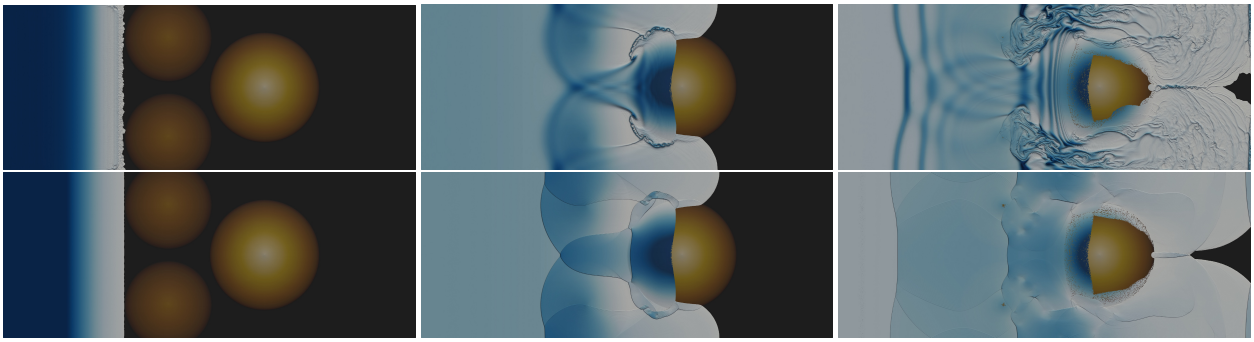


Figure 6.17: Dam break with bumps – Comparison with Hyperbolic Serre (top) and Saint-Venant (bottom) at $t = \{1, 7.8, 15\}$.

be:

$$\mathsf{h}(\boldsymbol{x}, 0) = \begin{cases} 1.8\,\mathrm{m}, & \sqrt{(x-250)^2 + (y-250)^2} < R, \\ 1\,\mathrm{m}, & \text{otherwise}, \end{cases}, \quad \boldsymbol{q}(\boldsymbol{x}, 0) = \mathsf{h}(\boldsymbol{x}, 0)\Big(y - y_c, -(x - x_x)\Big).$$

For this test case, we again use the `Ryujin` software. The computational domain is set to $D = (0, 500\,\mathrm{m}) \times (0, 500\,\mathrm{m})$ with 4,004,001 $\mathbb{Q}_1$ nodes. The final time is set to $T = 10\,\mathrm{s}$ with CFL number of 0.075. In Figure 6.18, we compare the water depth profiles for both the hyperbolic Serre model (top) and the Saint-Venant model (bottom) at the times $t = \{1, 5, 10\}\mathrm{s}$. Notice that at time $t = 1\,\mathrm{s}$, the Hyperbolic Serre model is producing structures reminiscent of a circular Kelvin–Helmholtz instability.

6.7 Laboratory experiments

We continue the numerical illustrations by reproducing several laboratory experiments that have been documented in the literature.

6.7.1 Shoaling of solitary waves over sloped beach

We now consider the 1994 experiments of Guibourg [35] conducted at LEGI (Laboratoire des Écoulements Géophysiques et Industriels) in Grenoble, France, to investigate the shoaling of solitary waves over a sloped beach. We consider 4 series of experiments proposed in [35] with a

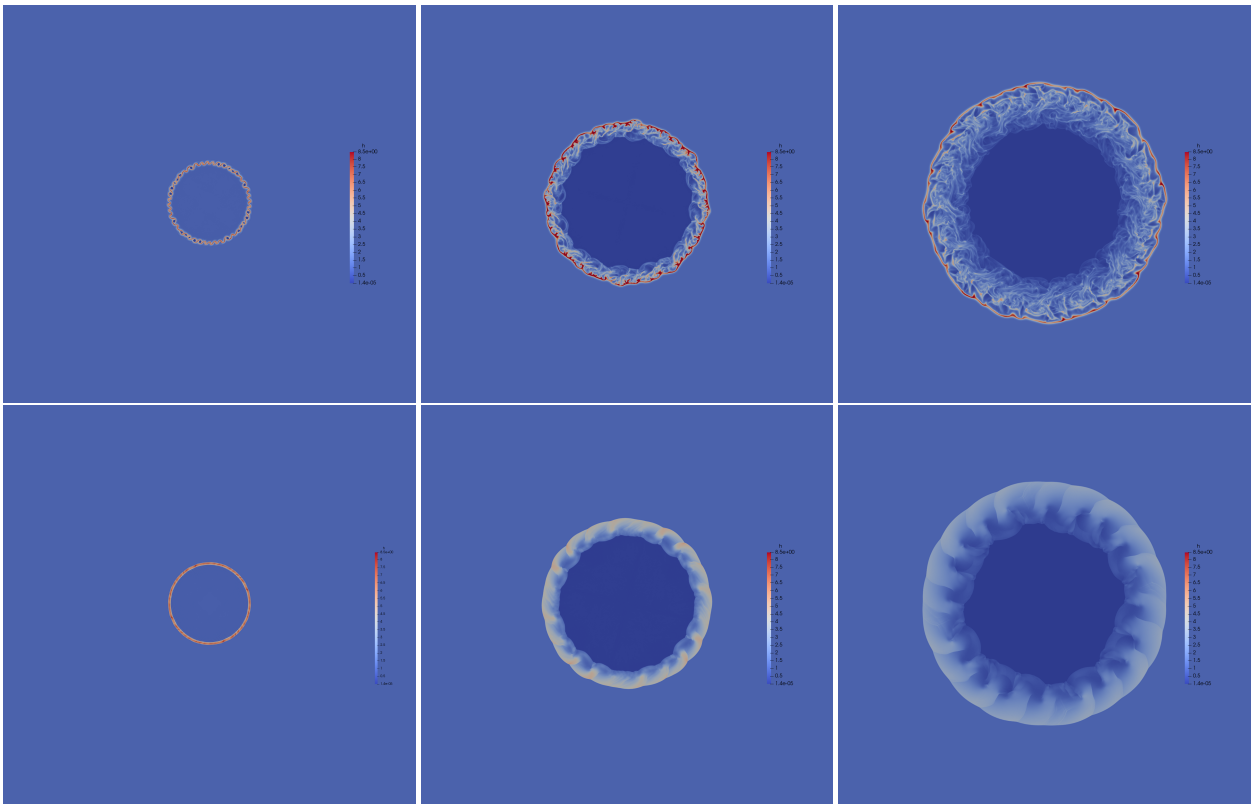


Figure 6.18: Circular dam break with divergent-free velocity field – Comparison with Hyperbolic Serre (top) and Saint-Venant (bottom) at $t = \{1, 5, 10\}s$.

reference water depth of $\mathsf{h}_0 = 0.25\,\mathrm{m}$ and different solitary wave amplitudes (see: Table 6.5). We simulate the experiments in one spatial dimension and reproduce the bathymetry as follows:

$$z(x) = \begin{cases} \frac{1}{30}(x - 2.5) - \mathsf{h}_0, & x \geq 25 \\ -\mathsf{h}_0, & \text{otherwise.} \end{cases}$$

The computational domain is set to $D = (-5\,\mathrm{m}, 35\,\mathrm{m})$. For each experiment, we initialize the solitary wave at $x_0 = 0\,\mathrm{m}$ with the profiles defined in (6.3.1) and the amplitudes shown in Table 6.5. We run the computations to the final time $T = 10\,\mathrm{s}$ on the three difference meshes with respective mesh-size: $h = \{0.05\,\mathrm{m}, 0.025\,\mathrm{m}, 0.0125\,\mathrm{m}\}$ (corresponding to $800, 1600, 3200\ \mathbb{P}_1$ elements). The CFL number is set to 0.1. We set wall boundary conditions at both ends of the domain.

In the experiments, the wave elevation was measured with three wave gauges (WGs) which

	Case 1	Case 2	Case 3	Case 4
α/h_0	0.096	0.2975	0.456	0.5343
WG1	7.75 m	5.75 m	4.25 m	4.25 m
WG2	8.25 m	6.25 m	5.0 m	5.0 m
WG3	8.75 m	6.75 m	5.75 m	5.75 m

Table 6.5: Solitary wave shoaling experiment [35] – configuaration values

were moved for each case. We report the location of the wave gauges in Table 6.5. In Figure 6.19, we show the comparisons with the numerical computations and the experimental data for each case. We observe that the numerical results match the experimental data very well. This set of experiments reinforces that the dispersive hyperbolic Serre model captures well the shoaling phenomenon induced by topography.

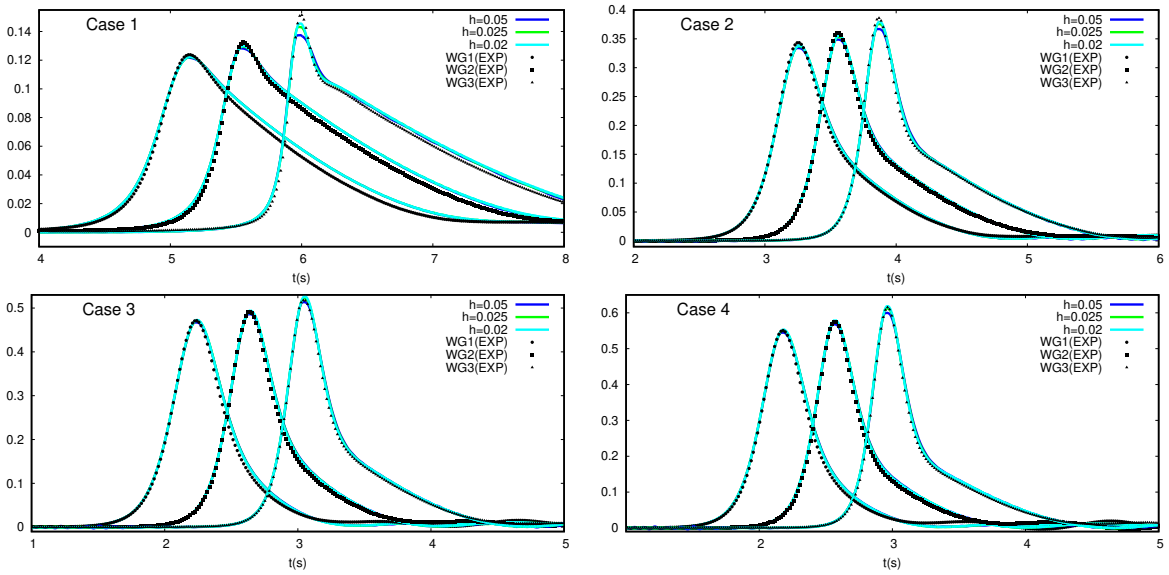


Figure 6.19: Comparison of numerical results with experimental data for solitary wave shoaling experiments of [35].

6.7.2 Periodic waves propagation over a submerged bar

We now consider the 1994 experiments conducted in Beji and Battjes [7] which investigate the propagation of periodic waves over a submerged trapezoidal bar. The goal of the experiments is to model the interaction of highly dispersive waves, and in particular, the release of higher-harmonics into a deeper region after the shoaling process.

We consider two of the experimental setups described in [7]: (i) sinusoidal long waves (SL) with target amplitude $a = 1\,\mathrm{cm}$ and period $T_p = 2\,\mathrm{s}$; (ii) sinusoidal high-frequency waves (SH) with target amplitude $a = 1\,\mathrm{cm}$ and period $T_p = 1.25\,\mathrm{s}$. We simulate these experiments in one spatial dimension and reproduce the bathymetry of the submerged bar as follows:

$$z(x) = \begin{cases} \frac{1}{20}(x-6), & 6 \le x \le 12 \\ 0.3, & 12 \le x \le 14 \\ 0.3 - \frac{1}{10}(x-14), & 14 \le x \le 17 \\ 0, & \text{otherwise.} \end{cases}$$

The computational domain is set to be $D = (-12.3\,\mathrm{m}, 37.7\,\mathrm{m})$. We impose two relaxation zones

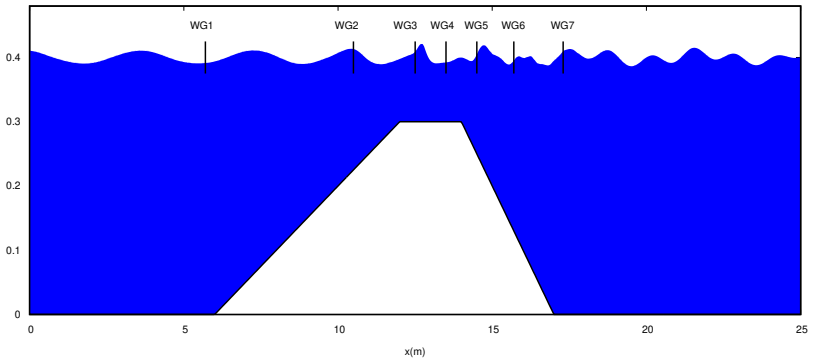


Figure 6.20: Submerged bar set up with gauge locations.

in the domain for the generation and absorption of waves (as described in Chapter 2). The length

of the generation zone for the SL case is $6\,\mathrm{m}$ (approximately 1.5 wavelengths) and $4\,\mathrm{m}$ for the SH case (approximately 2.0 wavelengths). The absorption zone for both cases is set to $D_{\text{abs}} = (25\,\mathrm{m}, 37.7\,\mathrm{m})$. We set the reference water depth to $\mathsf{H}_0 = 0.4\,\mathrm{m}$ and initialize the water height profile with $\mathsf{h}_0(x) = \mathsf{H}_0 - z(\boldsymbol{x})$ and discharge $q_0(x) = 0$. The periodic waves are introduced into the domain via the generation zone with the profiles given by:

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

where a is the amplitude, k the wave number and σ the wave frequency. Here, we define the wave frequency by $\sigma = \frac{2\pi}{T_p}$ and k is found by using the dispersion relation for the full Serre model: $k^2 = 3\sigma^2/(3g\mathsf{h}_0 - \mathsf{h}_0^2\sigma^2)$. We set the final time to be $t = 60\,\mathrm{s}$ and run with CFL=0.175. We run the computations on three different meshes with meshsizes $h = \{0.05\,\mathrm{m}, 0.025\,\mathrm{m}, 0.0125\,\mathrm{m}\}$ (corresponding to 1000, 2000, and 4000 $\mathbb{P}_1$ cells.).

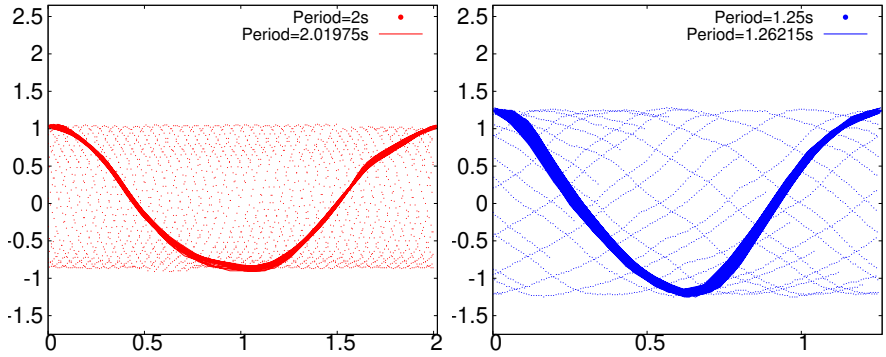


Figure 6.21: Illustration of *period-folding* with experimental wave gauge 1 with SL case (left) and SH case (right).

In the original experiments, seven wave gauges (WGs) were used to measure the water elevation: WG1($x = 5.7\,\mathrm{m}$), WG2($x = 10.5\,\mathrm{m}$), WG3($x = 12.5\,\mathrm{m}$), WG4($x = 13.5\,\mathrm{m}$), WG5($x = 14.5\,\mathrm{m}$), WG6($x = 15.7\,\mathrm{m}$), WG7($x = 17.3\,\mathrm{m}$). In Figure 6.20, we show the locations of these wave gauges with respect to the bathymetry. The experimental data used here was obtained from

the original author of the experiments, Serdar Beji. It was our experience that the experimental data did not quite match the targeted values of the period mentioned above. To illustrate this, we introduce a post-processing technique of the experimental data that we call *period-folding*. The idea is that given some experimental time series that is supposedly periodic with period T_p in the time interval $[t_0, T_{\text{final}}]$, the folding of the sequence obtained by the mapping $t \mapsto t - t_0 - \lfloor \frac{t-t_0}{T_p} \rfloor T_p$ should represent the evolution of the signal during one period (here $\lfloor \cdot \rfloor$ is the floor function). Doing this folding gives a better idea of the long time behavior of the experimental data than just looking at one specific window of length T_p as often done in the literature. In particular it reveals whether the signal is indeed periodic with period T_p. In Figure 6.21, we use period folding for the experimental data at WG1 with the targeted values of T_p. This process shows that the experimental data have not exactly the alleged period. We have been able to discover a good approximation of the actual period T_p^{adj} by doing the period folding with various values of T_p: (i) SL case, $T_p = 2\,\text{s}$, $T_p^{\text{adj}} = 2.019{,}75\,\text{s}$; (ii) SH case, $T_p = 1.25\,\text{s}$, $T_p^{\text{adj}} = 1.262{,}15\,\text{s}$. We also note that the wave amplitude value for the SH case is closer to $0.014\,\text{m}$ and this is what we use for our computations.

Using the adjusted values above and the period-folding technique, we compare in Figures 6.22 and 6.23 the experimental data and the results of the computations at the wave gauges 2–7 (Figure 6.22 for the SL experiment and and Figure 6.23 for the SH experiment). For the experimental data, we choose t_0 to be the time corresponding to the second maximum wave height of the signal at WG2. For the numerical simulations, we choose t_0 to be the time corresponding to the maximum wave height around $t \approx 40\,\text{s}$ at WG2. We observe that the numerical results converge as the mesh is refined. The SH experiment is relatively well reproduced at all the gauges. There are slight deviations at the last two gauges behind the bar for the SL experiment. It is possible that some wave breaking occurs between gauges 5 and 6 in this case.

6.7.3 Propagation of periodic waves over an elliptic shoal

We consider the 1982 experiments of Berkhoff et al. [8] conducted to study the propagation of monochromatic waves over an elliptic shoal. The goal of the experiments were to model the

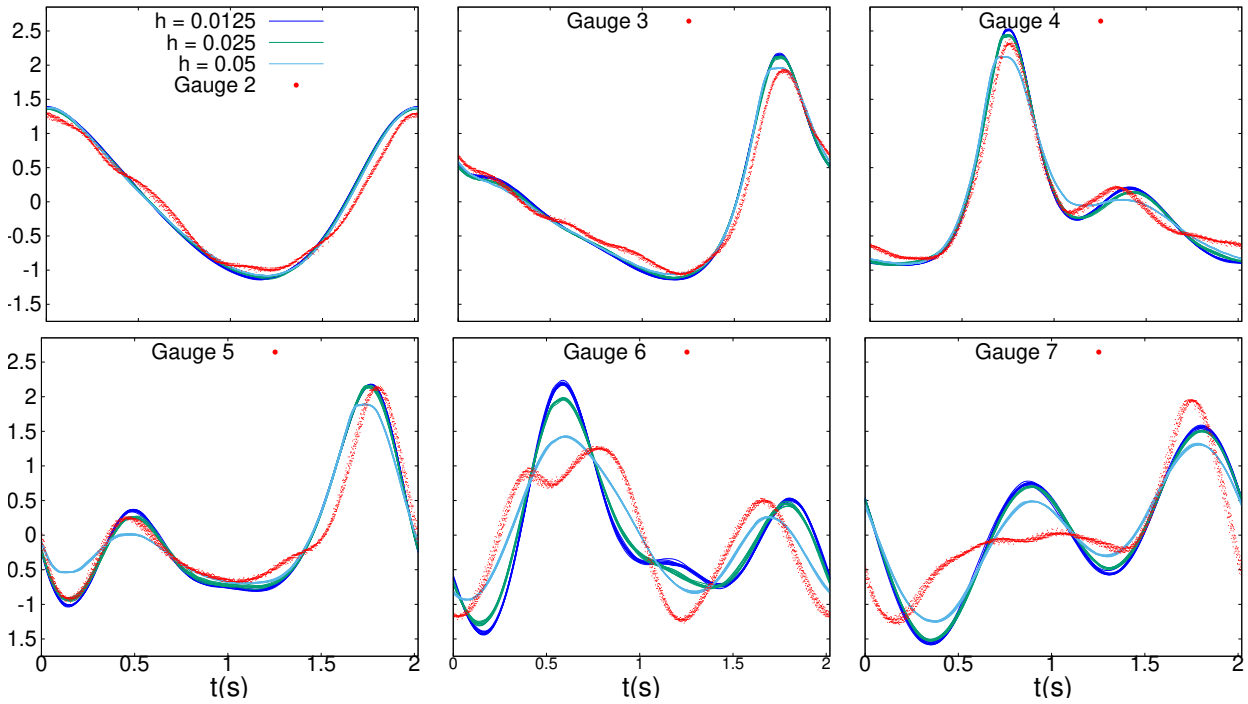


Figure 6.22: SL Case. Water elevation at seven gauges. Numerical results using three meshes, $h = \{0.05\,\mathrm{m}, 0.025\,\mathrm{m}, 0.0125\,\mathrm{m}\}$ (solid lines). Experimental data (red points).

refraction and diffraction of waves when propagating over a varying bottom. These experiment have become a benchmark for validating dispersive wave models (see: Duran and Marche [15], Ricchiuto and Filippini [52].)

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 $\boldsymbol{x} \mapsto \boldsymbol{x}_r(\boldsymbol{x})$:

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

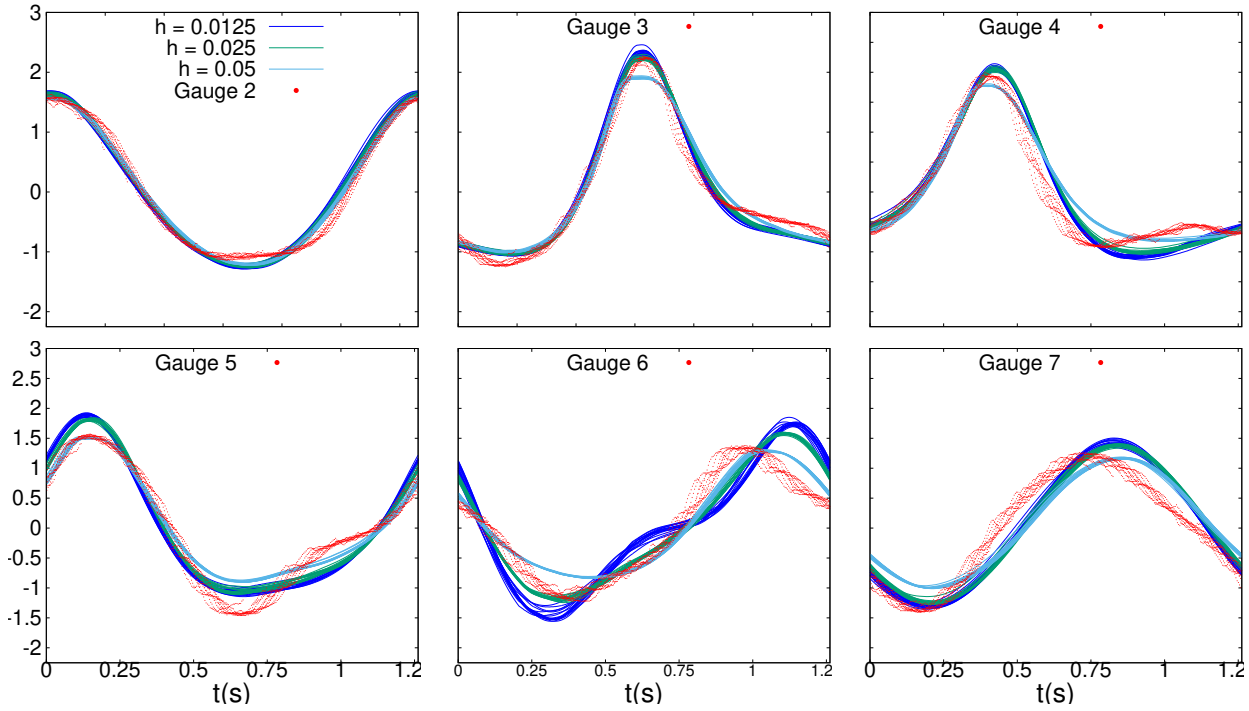


Figure 6.23: SH Case. Water elevation at seven gauges. Numerical results using three meshes, $h = \{0.05\,\mathrm{m}, 0.025\,\mathrm{m}, 0.0125\,\mathrm{m}\}$ (solid lines). experimental data (red points).

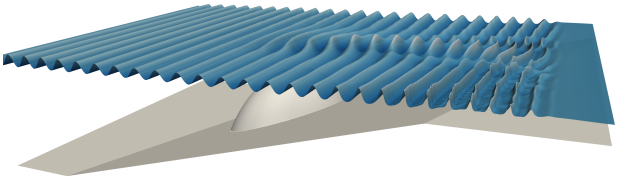
Then we define the sloping bottom and elliptic shoal profiles as:

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

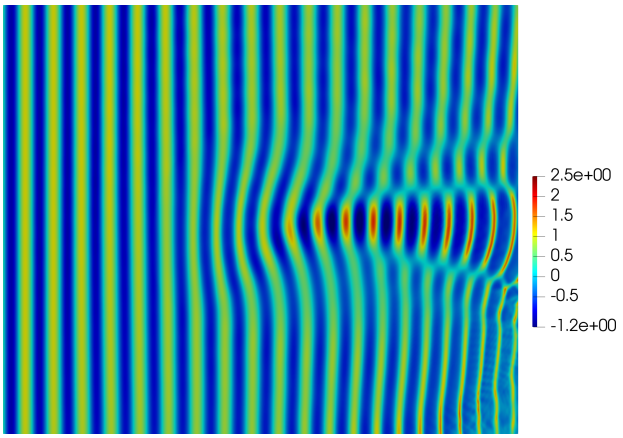
$$z_{\text{shoal}}(\boldsymbol{x}) := \begin{cases} -0.3 + \frac{1}{2}\sqrt{1 - (\frac{x_r(\boldsymbol{x})}{3.75})^2 - (\frac{y_r(\boldsymbol{x})}{5})^2}, & (\frac{x_r(\boldsymbol{x})}{3.75})^2 + (\frac{y_r(\boldsymbol{x})}{5})^2 \leq 1 \\ 0, & \text{otherwise}. \end{cases}$$

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 Berkhoff et al. [8] to include a flat portion at the right-end of the basin.

The reference water depth is set to $\mathsf{H}_0 = 0.45\,\mathrm{m}$. We initialize the water height with $\mathsf{h}_0(\boldsymbol{x}) =$



(a) Free surface elevation and topography; mesh refinement 1



(b) Normalized view of wave heights; mesh refinement 1

Figure 6.24: Elliptic shoal experiments

$\mathsf{H}_0 - z(\boldsymbol{x})$ and discharge $\boldsymbol{q}_0(\boldsymbol{x}) = \mathbf{0}$. For the simulation of the experiments, we use the `Ryujin` code. The computational domain is set to be $D = (-14, 18\mathrm{m}) \times (-10, 10\mathrm{m})$. We generate the periodic waves via the generation zone methodology described in §2.4.3 with the profiles:

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

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 by 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 $4\,\mathrm{m}$, i.e., $L_{\text{gen}} = L_{\text{abs}} := 4\,\mathrm{m}$, 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. In Figure 6.24, 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:

$$\text{section } 1 : \{x = 1\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, \qquad \text{section } 2 : \{x = 3\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\},$$

$$\text{section } 3 : \{x = 5\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, \qquad \text{section } 4 : \{x = 7\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\},$$

$$\text{section } 5 : \{x = 9\,\mathrm{m}, -5\,\mathrm{m} \le y \le 5\,\mathrm{m}\}, \qquad \text{section } 6 : \{y = -2\,\mathrm{m}, 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\},$$

$$\text{section } 7 : \{y = 0\,\mathrm{m},\ \ 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\}, \qquad \text{section } 8 : \{y = 2\,\mathrm{m},\ \ 0\,\mathrm{m} \le x \le 11\,\mathrm{m}\}.$$

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 Figure 6.25, 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.

6.7.4 Propagation of periodic waves over semi-circular shoal

We now consider the 1971 experiments of Whalin [63] performed at the U.S. Army Engineer Waterways Experiment Station (now the U.S. Army Engineer Research and Development 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 [63, Fig. 68]).

The experimental basin is designed to be $25.603\,\mathrm{m}$ in length and $6.096\,\mathrm{m}$ wide and the still water elevation is set to $0.4572\,\mathrm{m}$. Wse 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 $L_{\text{gen}} = L_{\text{abs}} = 8\,\mathrm{m}$ (which is roughly 2 wave lengths) with $x_{\text{min}} = -10$ and $x_{\text{max}} = 33$. The bathymetry is reproduced as follows: Defining $G(y) := \sqrt{y(6.096 - y)}$,

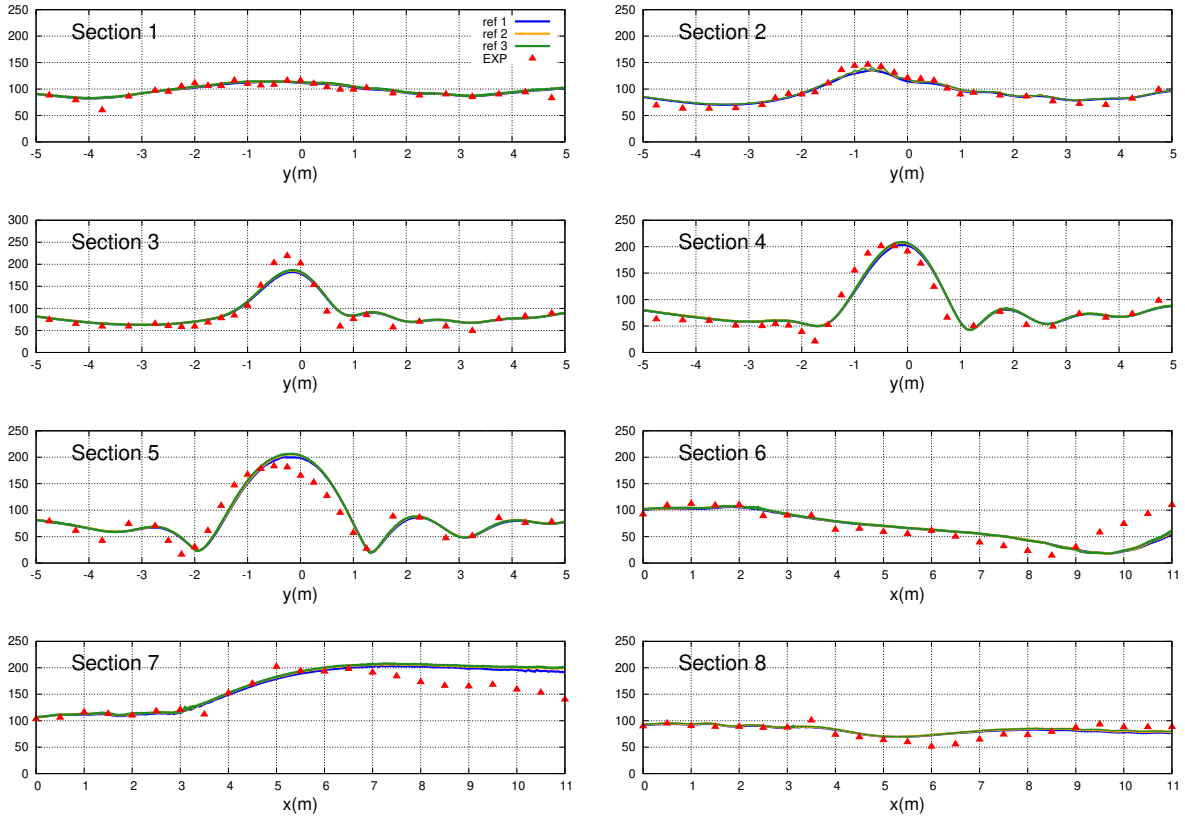


Figure 6.25: Elliptic shoal – Comparison of numerical results (3 mesh refinements) with the experimental data along the 8 sections. Experimental data: red triangles.

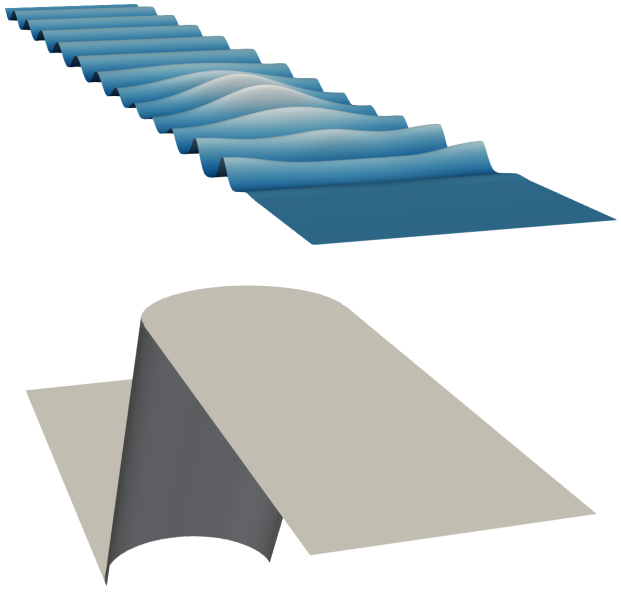
we set

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

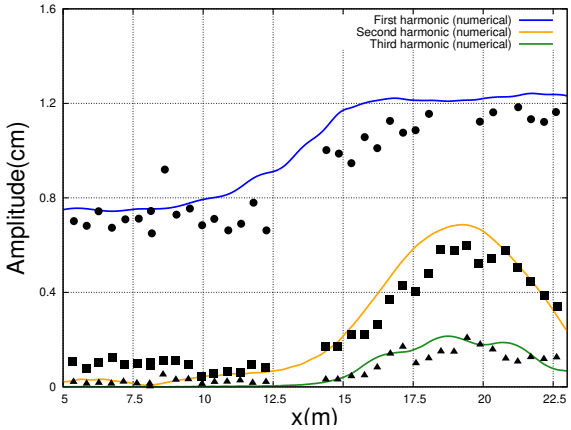
The computational domain is composed of a quadrilateral mesh with $265{,}761$ $\mathbb{Q}_1$ dofs. We run the numerical simulations until the time $T = 60\,\mathrm{s}$ to allow the waves to reach a steady state. The CFL number is set to 0.125.

In Whalin [63], 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 interpolate the centerline $y = 3.048\,\mathrm{m}$ with roughly 1400 points along the x-axis at every $0.001\,\mathrm{s}$ 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 Figure 6.26, 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.



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



(b) Comparison of numerical and experimental amplitude spectrum

Figure 6.26: Whalin semi-circular shoal results

6.7.5 2D Solitary wave run-up over a conical island

We consider the 1995 laboratory experiments conducted by Briggs et al. [11] at the US Army Waterways Experiment Station in Vicksburg, Mississippi (now the US Army Engineer Rsearch and Development Center). The laboratory experiments were motivated by several tsunami events

in the 1990s where large unexpected run-up heights were observed on the back (or lee) side of small islands. Several authors have used this experiment to study the run-up phenomena using the classical Shallow Water model and other dispersive models (see: Hou et al. [36], Lannes and Marche [45], Kazolea et al. [39]).

Let $r(\boldsymbol{x})$ by the radius from the center of the island located at $(12.96\,\mathrm{m}, 13.80\,\mathrm{m})$. Then the conical island bathymetry is defined by

$$z(\boldsymbol{x}) = \begin{cases} \min\left(\mathsf{h}_{\text{top}}, \mathsf{h}_{\text{cone}} - r(\boldsymbol{x})/s_{\text{cone}}\right), & r(\boldsymbol{x}) < r_{\text{cone}}, \\ 0, & \text{otherwise}, \end{cases} \tag{6.7.1}$$

where $\mathsf{h}_{\text{top}} = 0.625\mathrm{m}, \mathsf{h}_{\text{cone}} = 0.9\mathrm{m}$ and $r_{\text{cone}} = 3.6\mathrm{m}$. We reproduce two experiments, which we call Case B and Case C, with $\alpha/\mathsf{h}_0 = 0.091$ and $\alpha/\mathsf{h}_0 = 0.181$ where $\mathsf{h}_0 = 0.32\,\mathrm{m}$ and α is the amplitude of the solitary wave.

The computations are done in the domain $(0, 25\,\mathrm{m})\times(0, 30\,\mathrm{m})$ until the final time $T = 12\,\mathrm{s}$ with wall boundary conditions and CFL number 0.25. We initiate the solitary wave at $x_0 = 9.36 - \frac{L}{2}$ using (6.3.1) with $L = \frac{2\mathsf{h}_0}{k}\operatorname{arccosh}\sqrt{20}$ and $k = \sqrt{\frac{3\alpha}{4\mathsf{h}_0}}$. Here x_0 is the location of the experimental wave gauge 3 (WG3) which was used to measure the free surface elevation away from the island. In Figure 6.27, we show the surface plots of the free surface elevation $\mathsf{h} + z$ on a mesh composed of 52,129 $\mathbb{P}_1$ nodes at $t = \{0, 5.8, 8\,\mathrm{s}\}$.

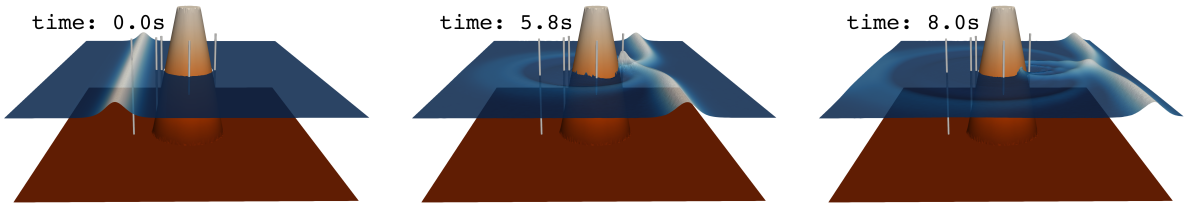


Figure 6.27: Experiment 4 – Surface plot of the water elevation $\mathsf{h} + z$ at several times for Case C. The thin grey cylinders represent the wave gauges WG3, WG6, WG9, WG16, WG22 (left to right).

In the experiment, several wave gauges were placed around the island to measure the free surface elevation and wave run-up. We compare the numerical results with the measurents at four of the experimental wave gauges: WG6$(9.36\,\mathrm{m}, 13.80\,\mathrm{m})$, WG9$(10.36\,\mathrm{m}, 13.80\,\mathrm{m})$, WG16$(12.96\,\mathrm{m}, 11.22\,\mathrm{m})$, WG22$(15.56\,\mathrm{m}, 13.80\,\mathrm{m})$. We show in Figure 6.28 the comparison with the experimental data and numerical simulations for both Case B (on the left) and Case C (on the right). For both cases, the numerical results show good agreement with the experimental data. We capture well the magnitudes of the run-up and draw-down at the front side of the island at WG9 with a slight overshoot in Case C. For both cases, we see very good comparison with WG16 which corresponds to the run-up and draw-down on the side of the island. We note that the experimental data shows subsequent free surface oscillations after impacting the island which is most notable in WG9, but our numerical simulation do not capture this effect. This phenomena is consistent with the literature and has been observed by others (see: Lannes and Marche [45], Kazolea et al. [39], Yamazaki et al. [66]), and is likely due to inconsistency in the original experiments.

6.7.6 Propagation over a solitary wave over a triangular shelf with conical island

We now reproduce the experiments of Swigler [59] and Lynett et al. [47] 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: Duran and Marche [15], Kazolea et al. [40], Roeber and Cheung [53]) have used these experiments for validation.

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

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

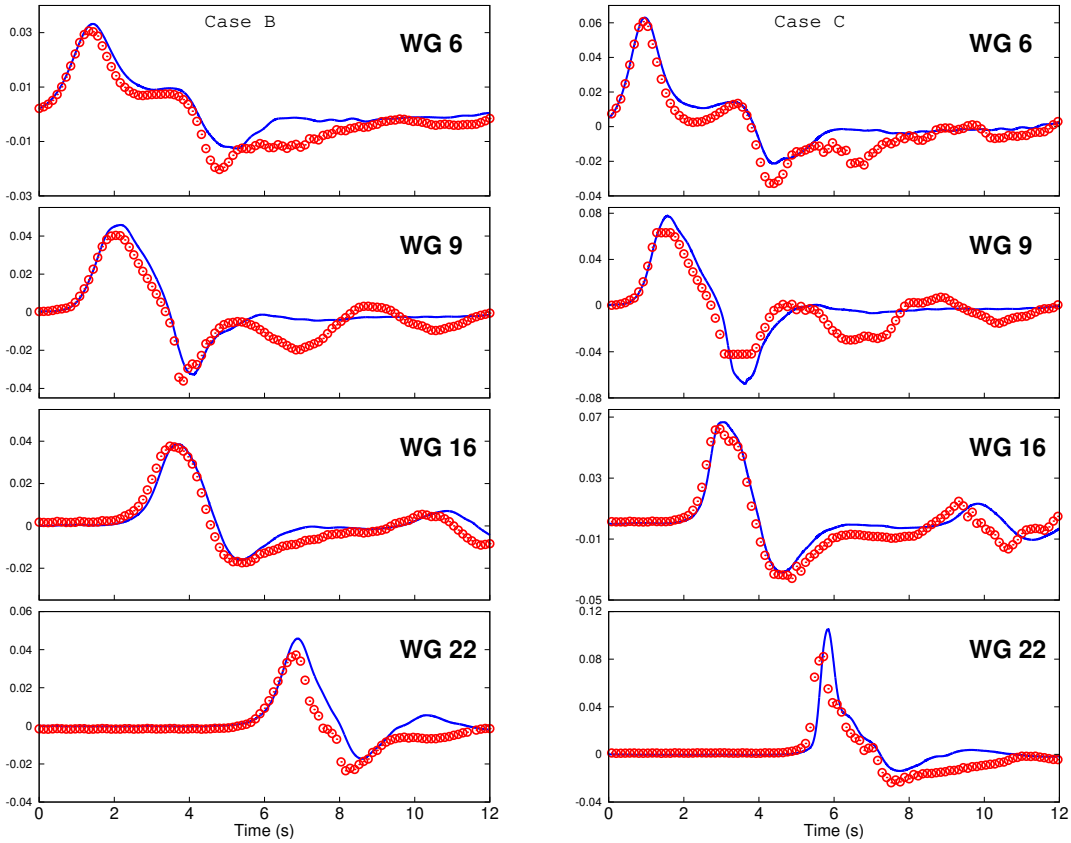


Figure 6.28: Experiment 4 – Temporal series over the period $t \in [0, 12\,\mathrm{s}]$ of the free surface elvation $\mathsf{h} + z$ in meters at the four WGs (blue solid) compared to the experimental data (red circles) for Case B (on the left) and Case C (on the right).

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

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

Then the full bathymetry is defined by

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

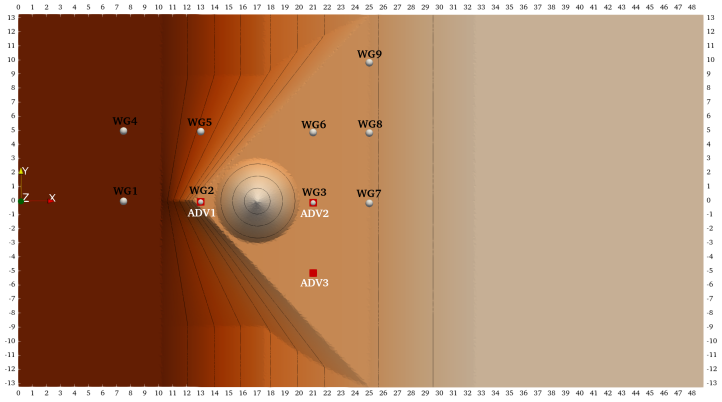
The setup of this complex bathymetry can be seen in Figure 6.29b.

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 (6.3.1). 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 Figure 6.31, 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 measure the velocity. In Figure 6.29, 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 Figure 6.30a, 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 Figure 6.30b, 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 $57{,}188$

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

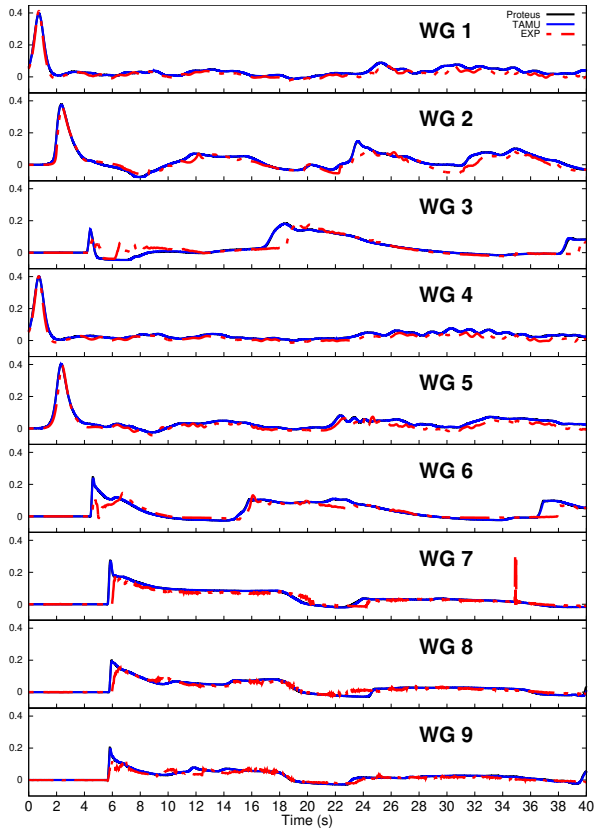
(a) WGs and ADVs coordinates.



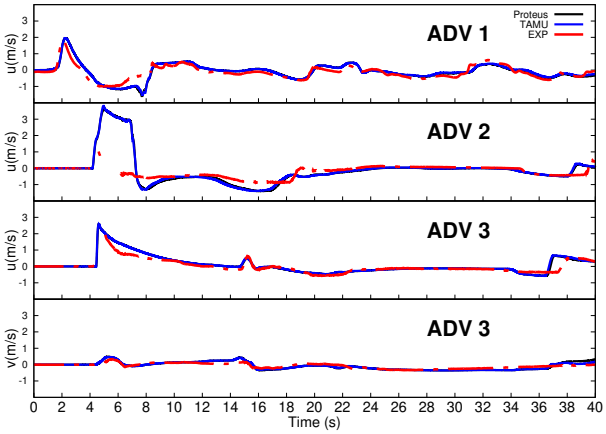
(b) Overview of bathymetry with WGs (white spheres) and ADVs locations (red boxes).

Figure 6.29: (a) Coordinates of the wave gauges and ADVs in meters; (b) Overview of their respective locations on the bathymetry.

$\mathbb{P}_1$ nodes and CFL number of 0.25.



(a) Wave gauges



(b) Acoustic Doppler Velocimeters

Figure 6.30: (a) Temporal series over the period $t \in [0, 40s]$ 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 $t \in [0, 40s]$ of velocity $\boldsymbol{v}$ (blue solid) and experimental ADVs (red dashed).

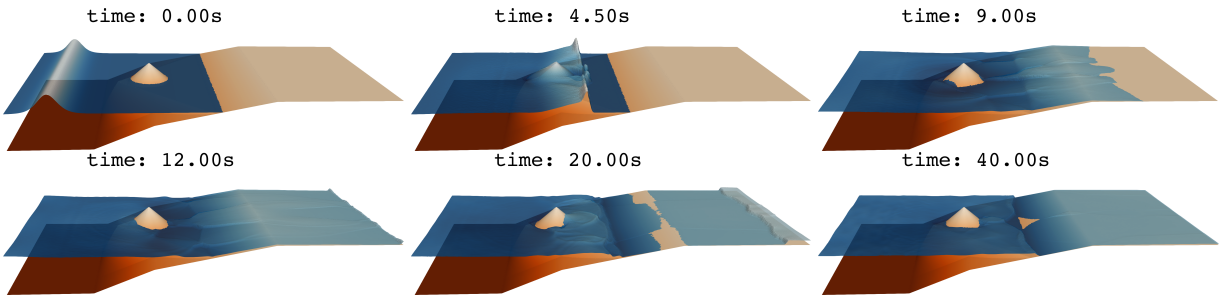


Figure 6.31: Surface plot of the water elevation $\mathsf{h} + z$ at several times.

REFERENCES

[1] Borys Alvarez-Samaniego and David Lannes. Large time existence for 3D water-waves and asymptotics. *Invent. Math.*, 171(3):485–541, 2008.

[2] George J. Arcement and Verne R. Schneider. Guide for selecting manning's roughness coefficients for natural channels and flood plains. Technical Report 2339, U.S. Geological Survey, 1989.

[3] Daniel Arndt, Wolfgang Bangerth, Bruno Blais, Marc Fehling, Rene Gassmöller, Timo Heister, Luca Heltai, Uwe Köcher, Martin Kronbichler, Matthias Maier, Peter Munch, Jean-Paul Pelteret, Sebastian Proell, Konrad Simon, Bruno Turcksin, David Wells, and Jiaqi Zhang. The `deal.II` library, version 9.3. *Journal of Numerical Mathematics*, 2021, accepted for publication. URL `https://dealii.org/deal93-preprint.pdf`.

[4] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, Rupert Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[5] Pascal Azerad, Jean-Luc Guermond, and Bojan Popov. Well-balanced second-order approximation of the shallow water equation with continuous finite elements. *SIAM J. Numer. Anal.*, 55(6):3203–3224, 2017.

[6] Eric Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25:315–337, 2004.

[7] S. Beji and J.A. Battjes. Numerical simulation of nonlinear wave propagation over a bar. *Coastal Engineering*, 23(1):1 – 16, 1994.

[8] J.C.W. Berkhoff, N. Booy, and A.C. Radder. Verification of numerical wave propagation models for simple harmonic linear water waves. *Coastal Engineering*, 6(3):255 – 279, 1982.

[9] Alfredo Bermúdez and Ma. Elena Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. Fluids*, 23(8):1049–1071, 1994.

[10] P. Bonneton, F. Chazel, D. Lannes, F. Marche, and M. Tissier. A splitting approach for

the fully nonlinear and weakly dispersive Green–Naghdi model. *J. Comput. Phys.*, 230(4): 1479–1498, 2011.

[11] Michael J. Briggs, Costas E. Synolakis, Gordon S. Harkins, and Debra R. Green. Laboratory experiments of tsunami runup on a circular island. *pure and applied geophysics*, 144(3): 569–593, 1995.

[12] Marie-Odile Bristeau, Anne Mangeney, Jacques Sainte-Marie, and Nicolas Seguin. An energy-consistent depth-averaged Euler system: derivation and properties. *Discrete Contin. Dyn. Syst. Ser. B*, 20(4):961–988, 2015.

[13] A. Chertock, S. Cui, A. Kurganov, and T. Wu. Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *Internat. J. Numer. Methods Fluids*, 78(6):355–383, 2015.

[14] Paul J. Dellar and Rick Salmon. Shallow water equations with a complete coriolis force and topography. *Physics of Fluids*, 17(10):106601, 2005. doi: 10.1063/1.2116747.

[15] A. Duran and F. Marche. A discontinuous Galerkin method for a new class of Green–Naghdi equations on simplicial unstructured meshes. *Appl. Math. Model.*, 45:840–864, 2017.

[16] Denys Dutykh, Didier Clamond, Paul Milewski, and Dimitrios Mitsotakis. Finite volume and pseudo-spectral schemes for the fully nonlinear 1d serre equations. *European Journal of Applied Mathematics*, 24(5):761–787, 2013. doi: 10.1017/S0956792513000168.

[17] G. A. El, R. H. J. Grimshaw, and N. F. Smyth. Unsteady undular bores in fully nonlinear shallow-water theory. *Physics of Fluids*, 18(2):027104, 2006.

[18] Alexandre Ern and Jean-Luc Guermond. *Finite Elements III*. Springer International Publishing, 2021.

[19] C. Escalante, M. Dumbser, and M. J. Castro. 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.

[20] N. Favrie and S. Gavrilyuk. A rapid numerical method for solving Serre–Green–Naghdi equations describing long free surface gravity waves. *Nonlinearity*, 30(7):2718–2736, 2017.

[21] Enrique D. Fernandez-Nieto, Martin Parisot, Yohan Penel, and Jacques Sainte-Marie. A hierarchy of dispersive layer-averaged approximations of euler equations for free surface flows. *Communications in Mathematical Sciences*, 16(5):1169–1202, 2019.

[22] S. L. Gavrilyuk and S. M. Shugrin. Media with equations of state that depend on derivatives. *Journal of Applied Mechanics and Technical Physics*, 37(2):177–189, 1996.

[23] Sergey L. Gavrilyuk, Boniface Nkonga, Keh-Ming Shyue, and Lev Truskinovsky. Generalized riemann problem for dispersive equations. 2018.

[24] Edwige Godlewski and Pierre-Arnaud Raviart. *Introduction*. Springer New York, New York, NvY, 1996. ISBN 978-1-4612-0713-9.

[25] A. E. Green and P. M. Naghdi. A derivation of equations for wave propagation in water of variable depth. *Journal of Fluid Mechanics*, 78(2):237–246, 1976. doi: 10.1017/S0022112076002425.

[26] A. E. Green, N. Laws, and P. M. Naghdi. On the theory of water waves. *Proc. Roy. Soc. (London) Ser. A*, 338:43–55, 1974.

[27] J. M. Greenberg and A.-Y. Le Roux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33(1):1–16, 1996.

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

[29] Jean-Luc Guermond and Richard Pasquetti. A correction technique for the dispersive effects of mass lumping for transport problems. *Computer Methods in Applied Mechanics and Engineering*, 253:186 – 198, 2013.

[30] Jean-Luc Guermond, Bojan Popov, and Yong Yang. The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations. *Journal of Scientific Computing*, 70(3):1358–1366, 2017.

[31] Jean-Luc Guermond, Murtazo Nazarov, Bojan Popov, and Ignacio Tomas. Second-order invariant domain preserving approximation of the euler equations using convex limiting. *SIAM*

J. Scientific Computing, 40(5):A3211–A3239, 2018.

[32] Jean-Luc Guermond, Bojan Popov, and Ignacio Tomas. Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Computer Methods in Applied Mechanics and Engineering*, 347:143 – 175, 2019. ISSN 0045-7825.

[33] Jean-Luc Guermond, Bojan Popov, Eric Tovar, and Chris Kees. Robust explicit relaxation technique for solving the Green–Naghdi equations. *J. Comput. Phys.*, 399:108917, 17, 2019.

[34] Jean-Luc Guermond, Matthias Maier, Bojan Popov, and Ignacio Tomas. Second-order invariant domain preserving approximation of the compressible navier–stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 375(1):113608, 2021.

[35] Sandrine Guibourg. *Modélisations numérique et expérimentale des houles bidimensionnelles en zone cotière*. PhD thesis, 1994. URL `http://www.theses.fr/1994GRE10160`.

[36] Jingming Hou, Qiuhua Liang, Franz Simons, and Reinhard Hinkelmann. A stable 2d unstructured shallow flow model for simulations of wetting and drying over rough terrains. *Computers & Fluids*, 82:132 – 147, 2013.

[37] Yuxin Huang, Ningchuan Zhang, and Yuguo Pei. Well-balanced finite volume scheme for shallow water flooding and drying over arbitrary topography. *Engineering Applications of Computational Fluid Mechanics*, 7(1):40–54, 2013.

[38] Mutsuto Kawahara and Tsuyoshi Umetsu. Finite element method for moving boundary problems in river flow. *International Journal for Numerical Methods in Fluids*, 6(6):365–386, 1986.

[39] M. Kazolea, A.I. Delis, I.K. Nikolos, and C.E. Synolakis. An unstructured finite volume numerical scheme for extended 2d boussinesq-type equations. *Coastal Engineering*, 69:42 – 66, 2012.

[40] M. Kazolea, A.I. Delis, and C.E. Synolakis. Numerical treatment of wave breaking on unstructured finite volume approximations for extended boussinesq-type equations. *Journal of Computational Physics*, 271:281 – 305, 2014. Frontiers in Computational Physics.

[41] C. E. Kees and M. W. Farthing. Parallel computational methods and simulation for coastal

and hydraulic applications using the proteus toolkit. *Supercomputing 11: Proceedings of the PyHPC11 Workshop.*, 2011.

[42] B. Khobalatte and B. Perthame. Maximum principle on the entropy and second-order kinetic schemes. *Math. Comput.*, 62(205):119–131, 1994.

[43] Alexander Kurganov and Guergana Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Commun. Math. Sci.*, 5(1): 133–160, 2007.

[44] D Lannes. Modeling shallow water waves. *Nonlinearity*, 33(5):R1–R57, 2020.

[45] D. Lannes and F. Marche. A new class of fully nonlinear and weakly dispersive Green–Naghdi models for efficient 2d simulations. *Journal of Computational Physics*, 282:238 – 268, 2015.

[46] David Lannes and Philippe Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21(1):016601, 2009. doi: 10.1063/1.3053183.

[47] Patrick Lynett, David Swigler, Hoda El Safty, Luis Motoya, Adam Keen, Sangoung Son, and Pablo Higuera. Study of the three-dimensional hydrodynamics associated with a solitary wave traveling over an alongshore-variable, shallow shelf. *Journal of Waterway, Port, Coastal, and Ocean Engineering (ASCE)*, 2019.

[48] P. A. Madsen, H. B. Bingham, and H. A. Schäffer. Boussinesq-type formulations for fully nonlinear and extremely dispersive water waves: derivation and analysis. *Proc. Roy. Soc. (London) Ser. A*, 459:1075–1104, 2003.

[49] Matthias Maier and Martin Kronbichler. Efficient parallel 3d computation of the compressible euler equations with an invariant-domain preserving second-order finite-element scheme. *ACM Transactions on Parallel Computing*, accepted, 2021. URL `https://arxiv.org/abs/2007.00094`.

[50] Fabien Marche. Combined hybridizable discontinuous galerkin (hdg) and runge-kutta discontinuous galerkin (rk-dg) formulations for green-naghdi equations on unstructured meshes.

Journal of Computational Physics, 418:109637, 2020. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2020.109637.

[51] Dimitrios Mitsotakis, Denys Dutykh, and John D. Carter. On the nonlinear dynamics of the traveling-wave solutions of the Serre system. *Wave Motion*, 70:166–182, April 2017. doi: 10.1016/j.wavemoti.2016.09.008.

[52] M. Ricchiuto and A. G. Filippini. Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. *J. Comput. Phys.*, 271:306–341, 2014.

[53] Volker Roeber and Kwok Fai Cheung. Boussinesq-type model for energetic breaking waves in fringing reef environments. *Coastal Engineering*, 70:1 – 20, 2012. ISSN 0378-3839.

[54] A.J.C. de Saint-Venant. Thèorie du mouvement non-permanent des eaux, avec application aux crues des riviéres et á l'introduction des marées dans leur lit. *C. R. Acad. Sc. Parios*, (73):147–154, 1871.

[55] Ali Samii and Clint Dawson. An explicit hybridized discontinuous galerkin method for Serre–Green–Naghdi wave model. *Computer Methods in Applied Mechanics and Engineering*, 330: 447 – 470, 2018.

[56] Fernando J. Seabra-Santos, Dominique P. Renouard, and André M. Temperville. Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacle. *Journal of Fluid Mechanics*, 176:117–134, 1987.

[57] François Serre. Contribution à l'étude des écoulements permanents et variables dans les canaux. *La Houille Blanche*, (6):830–872, 1953.

[58] C. H. Su and C. S. Gardner. Korteweg-de Vries equation and generalizations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. *J. Mathematical Phys.*, 10:536–539, 1969.

[59] David Townley Swigler. 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.

[60] Sergey Tkachenko. *Analytical and numerical study of a dispersive shallow water model*. PhD thesis, Aix-Marseille University, 2020.

[61] E.F. Toro. *Shock-capturing methods for free-surface shallow flows*. John Wiley, 2001.

[62] C. B. Vreugdenhil. *Shallow-water flows*. Springer Netherlands, 1994. ISBN 978-94-015-8354-1.

[63] R. Whalin. *The limit of applicability of linear wave refration theory in a convergence zone*. PhD thesis, Texas A&M University, 1971.

[64] Gerald B. Whitham. *Linear Dispersive Waves*. John Wiley & Sons, Ltd, 1999. ISBN 9781118032954.

[65] Inc. Wolfram Research. Mathematica, Version 12.3.1. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2021.

[66] Yoshiki Yamazaki, Zygmunt Kowalik, and Kwok Fai Cheung. Depth-integrated, non-hydrostatic model for wave breaking and run-up. *International Journal for Numerical Methods in Fluids*, 61(5):473–497, 2009.

[67] Steven T Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, 1979. ISSN 0021-9991.

[68] Yao Zhang, Andrew B. Kennedy, Nishant Panda, Clint Dawson, and Joannes J. Westerink. Generating–absorbing sponge layers for phase-resolving wave models. *Coastal Engineering*, 84:1–9, 2014. ISSN 0378-3839.